OPEN ACCESS Freely available online

S. Disch, L. Terentiv, J. Koppens, T. Falk, J Leppänen, I. Munoz, J. Herre, A. Adami, A. Silzle, D. Fischer, P. Setiawan, C. Fersch, J.-H. Hanschke, S. Jelfs, W. de Bruijn, A. Eronen, S. Mate, S. Dick, C. Borss, P. M. Delgado, and M. M. Trinidad, "MPEG-I Immersive Audio–The ISO/MPEG Standard for Virtual/Augmented Reality Audio," *J. Audio Eng. Soc.*, vol. 74, no. 6, pp. 356–375 (2026 Jun.).
https://doi.org/10.17743/jaes.2022.0276.

MPEG-I Immersive Audio—The ISO/MPEG Standard for Virtual/Augmented Reality Audio

SASCHA DISCH,[1,*] (sascha.disch@iis.fraunhofer.de) **LEON TERENTIV,**[2] (leon.terentiv@dolby.com) **JEROEN KOPPENS,**[3] (jeroen.koppens@philips.com) **TOMMY FALK,**[4] *AES Member* (tommy.falk@ericsson.com),
JUSSI LEPPÄNEN,[5] (jussi.leppanen@nokia.com) **ISAAC MUNOZ,**[6] (isaacgm@qti.qualcomm.com) **JÜRGEN HERRE,**[7] *AES Fellow* (juergen.herre@audiolabs-erlangen.de) **ALEXANDER ADAMI,**[1] (alexander.adami@iis.fraunhofer.de)
ANDREAS SILZLE,[1] *AES Member* (andreas.silzle@iis.fraunhofer.de), **DANIEL FISCHER,**[2] (daniel.fischer@dolby.com) **PANJI SETIAWAN,**[2] (panji.setiawan@dolby.com)
CHRISTOF FERSCH,[2] *AES Member* (christof.fersch@dolby.com), **JAN-HENDRIK HANSCHKE,**[2] *AES Member* (jan-hendrik.hanschke@dolby.com), **SAM JELFS,**[3] (sam.jelfs@philips.com)
WERNER DE BRUIJN,[8] *AES Member* (werner.de.bruijn@ericsson.com), **ANTTI ERONEN,**[5] (antti.eronen@nokia.com) **SUJEET MATE,**[5] (sujeet.mate@nokia.com)
SASCHA DICK,[1] *AES Fellow* (sascha.dick@iis.fraunhofer.de), **CHRISTIAN BORSS,**[1] (christian.borss@iis.fraunhofer.de) **PABLO M. DELGADO,**[1] (pablo.delgado@iis.fraunhofer.de) **AND**
MIGUEL M. TRINIDAD[1] (miguel.trinidad.mendez@iis.fraunhofer.de)

[1] *Fraunhofer Institute for Integrated Circuits (IIS), Erlangen, Germany*
[2] *Dolby Germany GmbH, Nuremberg, Germany*
[3] *Koninklijke Philips N.V., Eindhoven, the Netherlands*
[4] *Ericsson AB, Stockholm, Sweden*
[5] *Nokia Technologies, Helsinki, Finland*
[6] *Qualcomm Technologies Inc., San Diego, CA, USA*
[7] *International Audio Laboratories Erlangen, Germany*
[8] *Ericsson Telecommunicatie B.V., Breda, the Netherlands*

The ISO/MPEG international standard on MPEG-I immersive audio was issued in November 2025 by the MPEG Audio group (ISO/IEC JTC 1/SC 29/WG 6). It provides technology for a compressed audio representation and real-time interactive rendering within virtual and augmented reality applications with six degrees of freedom. It enables efficient bitrate management, high-quality storage, and transmission of virtual environments, including audio sources with spatial dimensions and specific radiation traits (such as musical instruments), along with geometric descriptions of acoustically relevant scene details (e.g., walls, doors, sound reflecting, and occluding elements). The audio rendering process incorporates comprehensive modeling of room acoustics and intricate acoustic phenomena, including occlusion, reflection, and diffraction caused by sound obstacles, Doppler effect, and dynamic environment changes triggered by user interactivity. This article presents an overview of background, development, underlying technology, and selected application aspects of the new standard.

Keywords: interactive sound rendering, six degrees of freedom, spatial audio, virtual acoustics

1 INTRODUCTION

Virtual reality (VR) and augmented reality (AR) have seen remarkable growth in recent years, paving the way for widespread consumer applications including educational entertainment, cultural performances and exhibitions, virtual concerts and conferences, live sports, and Social VR in general. The term "Social VR" denotes applications where

*To whom correspondence should be addressed, email: sascha.disch@iis.fraunhofer.de

several people (often represented by their avatars) can gather in shared virtual scenes, talk to each other and interact with objects in the scene. Also, the necessary hardware including head-mounted displays (like, e.g., HTC VIVE [1], Microsoft HoloLens [2], Apple Vision Pro [3], Meta Quest [4], and PICO [5]) became available to mass market consumers and was adopted by several communities, for example, by VR gaming enthusiasts.

For many applications, it will be essential that the VR/AR content created can be represented in an open and enduring format, ensuring that the content remains functional and accessible for decades to come and compatible across different platforms. Consequently, the ISO/MPEG standardization group initiated a new work item several years ago focused on a high-quality compressed representation and efficient rendering of VR/AR media [6]. The audio component of this MPEG-I standard builds on the latest predecessor technology (i.e., the ISO/IEC MPEG-H 3D Audio standard [7]) and incorporates a wide range of VR/AR-specific technological elements, enabling users to navigate and interact within the virtual environment in six degrees of freedom (6DoF), which includes free movement in 3D space (x, y, z) and head orientation (yaw, pitch, roll).

Previous papers provided an early overview of the ISO/MPEG-I immersive audio concept [8, 9]. A considerable amount of technology has been added to the standard since then. Now that the standard is finalized, this article—based on a prior convention publication [10] that has been substantially extended for journal publication—presents background, development, underlying technology and selected application aspects of the new standard.

1.1 MPEG Spatial Audio History

The MPEG work on audio for VR and AR is based on extensive expertise in compressed representation and rendering of spatial audio, which is briefly outlined in the following. In fact, beginning around 1990, the standards established by the MPEG Audio group have effectively defined the cutting edge of perceptual audio coding for over 30 years. Gradually, the focus of these technologies has shifted from basic mono or stereo audio coding to the representation and interactive rendering of spatial audio and VR/AR audio. This progression represents a natural evolution toward greater immersion and user interactivity [11].

The key spatial audio technologies include the following developments: The first widely adopted multichannel audio codec standardized by MPEG in 1997 was MPEG-2 Advanced Audio Coding (AAC) [12, 13], which delivers EBU broadcast quality at a bitrate of 320 kbit/s for a 5.1-channel signal. A major advancement came with the introduction of MPEG-4 High-Efficiency AAC (HE-AAC) [14] in 2002/2004, which merges AAC technology with bandwidth extension and parametric stereo coding, enabling full audio bandwidth even at lower bitrates. For 5.1 content, HE-AAC achieves quality comparable to AAC at a reduced bitrate of 160 kbit/s [15]. Subsequent MPEG standards introduced generalized methods for the parametric coding of multichannel spatial audio: MPEG-D MPEG Surround (2006) [16, 17] and MPEG-D Spatial Audio Object Coding (2010) [18, 19], which facilitate the efficient transmission of multichannel audio and object signals, respectively. Both codecs can operate at lower bitrates (e.g., 48 kbit/s for a 5.1 signal).

Finally, MPEG-D Unified Speech and Audio Coding (USAC; 2012) [20, 21] integrated enhanced AAC coding with advanced full-band speech coding into a highly efficient system, enabling the transmission of good quality mono signals at bitrates as low as 8 kbit/s. By incorporating advancements in joint stereo coding, USAC offers improved performance compared to HE-AAC, even for multichannel audio. The MPEG-H 3D Audio codec [7, 22] delivers efficient and universal representation and rendering of all known production paradigms [channels, objects, and higher-order ambisonics (HOA)] across various loudspeaker configurations and binaural outputs. Although its waveform coding core is based on MPEG-D USAC, the innovative unified integration of these production paradigms, along with versatile rendering capabilities, establishes a comprehensive spatial audio system.

1.2 Immersive Audio Background

Interactive immersive 3D audio rendering is already being deployed in several popular interactive applications like in [23] and [24] and closed ecosystems (e.g., Apple Spatial Audio Format, Apple Positional Audio Codec, or ECLIPSA by Google/Samsung), using proprietary technology or audio middlewares like Wwise [25] and FMOD [26]. Wwise and FMOD, but also Google Resonance Audio [27], Meta XR Audio SDK [28], and Alliance for Open Media Immersive Audio Model and Format [29] occupy distinct layers within the immersive audio and media processing stack. Resonance Audio and Meta XR Audio SDK are spatial-audio software development kits (SDKs) designed to render 3D audio scenes for VR/AR. Resonance Audio offers cross-platform engine integrations (Unity/Unreal), while Meta XR Audio SDK provides optimization for Meta devices and ecosystems. Rather than being a dedicated spatial audio renderer, Alliance for Open Media Immersive Audio Model and Format focuses on broad media processing, codecs, and encoding/decoding within GPU-accelerated pipelines.

Middlewares like FMOD and Wwise are designed to be used together with extensive authoring tools that have direct game engine integration and thus allow for artistic auralization of interactive applications for various playback scenarios and end consumer devices. They are versatile sound composition tools that inevitably require artistic tuning to achieve the desired output [30, 31]. Often, the desired output is not a realistic simulation of natural acoustics but a sound composition that is designed to support narration or game-play and that usually sounds "larger than life." While this is a desired feature for some use cases, other applications might demand a realistic and plausible output without requiring elaborate expert authoring and dedicated sound design that is often required for authoring virtual acoustic scenes with middlewares.

Based on real-world acoustic parameters that can be controlled to the necessary level of detail, authoring a scene for the MPEG-I Immersive Audio renderer is simply a matter of transferring real-world acoustic parameters to the virtual environment, which in turn creates a plausible auralization of that environment. To some degree, tuning is possible, but MPEG-I Immersive Audio does not rely on it to produce plausible auralizations.

For many applications, it is also important that the VR/AR content to be created can be represented in an open and long-term stable format that will reliably play back for decades to come, just the way MP3 compressed audio files (or any of its standardized successor formats) can be, to preserve the investment to be made into the new content creation chain and to allow for long-term archiving of content. MPEG-I Immersive Audio is the first standardized format and renderer that potentially enables interoperability of devices of different manufacturers and is independent of a closed proprietary ecosystem.

2 MPEG-I WORK ITEM

Starting in 2017, initial conversations about creating a specification for audio in VR/AR applications at ISO/MPEG Audio were prompted by broader discussions on MPEG-I that encompassed both visual and audio representations, as well as the transmission of VR/AR content within a single bitstream or file. Requirements and general architecture of the audio specification were established and approved [32]. It was determined that the coding of source waveforms would adhere to the recently finalized MPEG-H Audio standard. Consequently, the MPEG-I Immersive Audio project concentrated on developing a compressed 6DoF scene representation and its rendering.

In contrast to the offline audio encoding and rendering of previous MPEG Audio standards, the development and evaluation of an audio standard for real-time interactive VR/AR content is a much more complex project. Such a project needs many preparation steps to enable the generation of suitable VR/AR test material and to process it by the technology under development. Also, content presentation to subjects and evaluation of the technology's performance, while experiencing full 6DoF interaction, needed novel tools.

A PC-based real-time audio evaluation platform (AEP) [33] and a non-normative encoder input format (EIF) [34] for uncompressed scene representation had to be prepared to aid the actual development work. EIF and AEP allowed for the creation of test content that includes both visual components (Unity) and audio elements (EIF), which are used to test the technology under development by simulating specific acoustic phenomena. The MPEG Audio group compiled a list of essential acoustic attributes related to anticipated functionalities, along with a collection of test scenes that address these attributes. A selection of these scenes were used as a test set for a thorough and critical evaluation of audio quality [8].

The competitive phase of development required technology proposals to be submitted by November 2021 for the real-time rendering of a set of pre-agreed test scenes on the AEP. The top proposals were chosen in January 2022 based on established testing and evaluation procedures, resulting in Reference Model 0, setting the technical foundation for all subsequent developments. Additional technical extensions were suggested, their effectiveness was evaluated using a method called the "Core Experiment Methodology," and they were incorporated into the reference model if they passed the evaluation successfully [8]. The technical work on the specification was finalized by issuing the final draft international standard in early 2025. The international standard was released in November 2025 [35], with a reference software implementation also standardized at the same time [36].

3 MPEG-I PERCEPTUAL QUALITY ASSESSMENT

The perceived audio rendering quality was extensively evaluated both during the development and, finally, at a verification test of the finalized standard [37]. The evaluation posed numerous challenges, for which several subjective evaluation techniques were explored as potential solutions. Unlike earlier MPEG audio standards, VR/AR quality assessment requires a new test paradigm that goes beyond subjective audio quality alone. The method has to consider three interacting senses—hearing, vision, and proprioception—and, unlike traditional MPEG Audio listening tests, there is no clear reference condition (such as uncompressed audio) for auralized VR/AR content [8].

Subjective evaluation of VR/AR applications often adapts established methods such as the Multi Stimulus test with Hidden Reference and Anchor, as defined in ITU-R BS.1534 [38], which has been used to evaluate 3D audio reproduction [40]. In the search for suitable references for VR/AR audio, some previous frameworks have incorporated task-based behavioral metrics (e.g., wayfinding in 6DoF [41]) and subjective comparisons to real acoustic scenes, including archetypal AR scenarios where computer-generated audio must be indistinguishable from physically present sound sources heard through transparent earphones [42].

However, evaluation based on external references can be problematic: reproducing identical acoustic environments (AEs) across laboratories is logistically challenging, and even small mismatches can bias ratings and scale usage. As scene complexity increases, faithful virtual reproduction of a real reference becomes infeasible. Conversely, subjective evaluations based on an internal reference have been reported as effective for 3D audio, where the perceived internal "ideal" reference can also vary with the application and is determined by the experience and expectations of the listener [43]. This reliance on internal references is formalized through the concept of "plausibility," which evaluates if a simulation is in agreement with a listener's prior real-world experiences and expectations. While "authenticity" requires a simulation to be perceptually indistinguishable from a physical sound field, "plausibility" is often a suffi-

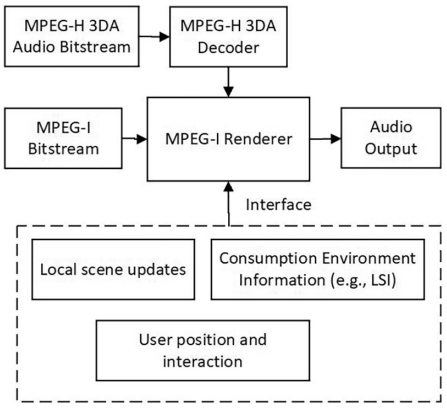


Fig. 1. Top-level MPEG-I Immersive Audio data flow.



Fig. 2. Geometry representation: visual geometry (a) versus acoustic geometry (b).

cient and more practical quality criterion for the vast majority of consumer VR/AR applications [44].

For MPEG-I, the motivation for the chosen subjective assessment was the need to balance scientific rigor with industry relevance while acknowledging the nonexistence of a physical reference. Consequently, MPEG-I evaluation intentionally moved away from traditional Multi Stimulus test with Hidden Reference and Anchor tests toward a multistimulus evaluation without an external reference according to ITU-R BS.2132 [39]. This measures both overall quality and plausibility [8, 45], using expert listeners involved in development of the standard but also experts stemming from other immersive audio domains.

Quality evaluation efforts progressed from initial A/B comparison tests during the call for proposals [8] to a final verification test utilizing a refined set of anchors [45]. Anchors are based on simplifications of the rendering pipeline for degrading spatial audio perception and additional deliberate timbre degradations. Extensive investigations were conducted to find stable anchor conditions; the group evaluated renderer modifications such as disabling early reflections (ERs) or portals but found the resulting degradation severity to be too scene-dependent.

Further tests compared simplified head-related transfer functions (HRTFs), specifically evaluating interaural level difference (IALD)–only models against stereo panning laws; panning was chosen because it provided a more stable spatial degradation by significantly reducing externalization. The low anchor was defined as a mono mixdown with a 3.5-kHz low-pass filter, while the mid anchor utilized stereo amplitude panning with a 7.5-kHz low-pass filter. This final large-scale verification test, involving ten diverse VR scenes across six international laboratories, confirmed that the renderer consistently achieved "excellent" ratings, with a median score of 84 points [45].

4 RENDERER OVERVIEW

This section provides an overview of the key aspects of the MPEG-I Immersive Audio technology. Fig. 1 illustrates the high-level architecture of the normative renderer. The waveforms associated with a VR/AR scene are extracted preferably from an MPEG-H 3D Audio bitstream and converted into pulse code modulation (PCM) input for the MPEG-I renderer. Additionally, PCM audio can also be obtained from local audio input or from the output of alternative codecs; for example, delay-optimized communication codecs may be leveraged for Social VR applications. Another input to the renderer is the MPEG-I Audio metadata bitstream, carrying a compressed description of the VR/AR scene. This is further detailed in SEC. 4.2.

4.1 Renderer Technology

The scene description is based on a discretized geometry representation of rooms and objects alongside their material properties and their positions and orientation within the scene. Fig. 2 shows two examples of visual geometry and a simplified, acoustically relevant geometry used for audio rendering. MPEG-I supports both common description paradigms: meshes (a set of connected vertices, edges, and faces) and voxels (volumetric elements on a regularly spaced, 3D grid). For more details of audio scene authoring, also refer to SEC. 7.6.

The renderer also receives real-time updates of the user's position and orientation, local scene updates, and, for AR applications, data of the acoustic and geometric characteristics of the consumption environment, the listener space information (LSI). Alignment of the virtual objects to the physical scene is implemented with AR anchors. The renderer accepts head-related impulse response (HRIR) datasets provided in SOFA format, specifically the SimpleFreeFieldHRIR convention, with HRIRs expected to match the renderer's sampling rate and preferably offer fine directional resolution. The HRIRs are used by the binaural spatializer as well as by rendering stages such as the multipoint HOA (MP-HOA), single-point HOA (SP-HOA), and homogeneous extent stages.

The renderer is further detailed in Fig. 3. The *scene state* is a central functional block which holds a description of the current state of all metadata including audio elements and geometry with associated materials. The *scene controller* handles all internal and external changes to the scene state data. The *stream manager* is an interface for the renderer to access audio streams that are associated with an audio element. A clock component is the time reference for

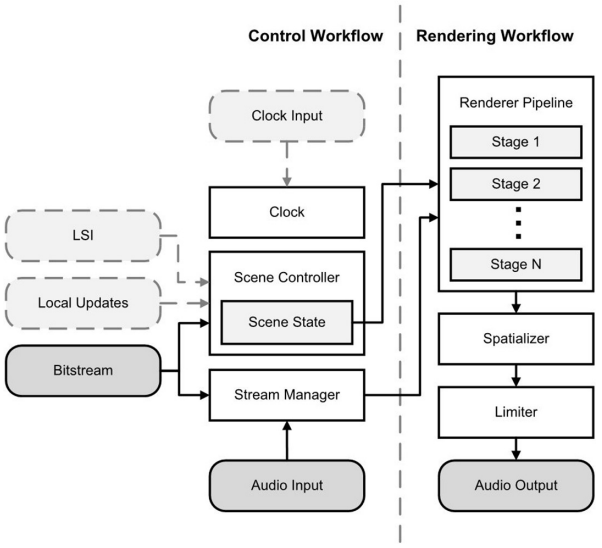


Fig. 3. MPEG-I audio renderer overview.

the system. The renderer auralizes the scene controller's updated scene state. The scene controller also determines whether there have been significant changes in the scene since the last update to the scene state and, if so, which sources have been updated. This allows for some rendering stages to determine if they can reuse previously calculated data, giving a reduction in computational complexity, or if new data must be generated.

Rendering is performed by sequential processing of audio elements in the *renderer pipeline* that comprises several *renderer stages*. Acoustic effects are modeled in dedicated stages [e.g., distance (simulating attenuation, propagation delay, and Doppler shifts), occlusion, diffraction, ERs by geometries carrying acoustic material properties, and room reverberation]. Other stages aim to reduce rendering complexity or aim at calculating supporting data for other stages. Render items (RIs) propagate through this pipeline and are processed or modified on their way. RIs originate from the scene state audio sources (primary RIs) or are created by the stages when secondary sources are needed (secondary RIs), like mirror image sources for ERs or substitute sources for diffraction. Most modifications, like gain or equalizer (EQ) changes, happen in the parametric metadata domain and alter metadata rather than directly processing audio samples. Audio processing is mostly carried out at the end of the pipeline and during spatialization.

Operations that directly depend on a snapshot of the time-variant scene state happen at a low control-step rate (e.g., 50 Hz), while operations on the actual audio streams run at an audio block rate of, for example, 187.5 Hz (48 kHz/256) using nonoverlapping blocks of typically 256 time-domain PCM samples and a sample rate of 48 kHz. Sample rates of f_s=32, 44.1, or 48 kHz and customizable block sizes are supported. After the audio effects processing takes place, the spatializer positions all RIs to their target locations, supporting both headphone binauralization or loudspeaker playback by tracked rendering. To prevent clipping in extreme situations, such as very loud nearby sources, a limiter is applied to the final output.

4.2 Bitstream Concept

MPEG-I Immersive Audio builds on the MPEG-H Audio Transport (MHAS) format that was introduced with MPEG-H 3D Audio [7], enabling it to additionally transmit the MPEG-I metadata. This is achieved through three newly defined MHAS packets: the Scene Config packet, Scene Update packet, and Scene Payload packet. This three-packet MHAS structure facilitates VR/AR applications across various scenarios, including client-server models and broadcasting. The Scene Payload packet carries all scene geometry and materials-related information required to render a scene. This information can be transmitted in one large packet, or it can be distributed to multiple smaller Scene Payload packets that can be sent sequentially, allowing for the interleaving of different MHAS packet types within a single stream, for instance, audio packets and metadata packets. Existing MHAS mechanisms can be utilized for stream linking and synchronization.

MPEG-I Immersive Audio uses specific encoding techniques for efficient transmission of bitstream metadata components (e.g., for acoustically relevant geometry, source directivity patterns, and precomputed ER and diffraction paths). While precomputation of reflection and diffraction paths reduces computational runtime complexity, it can significantly contribute to the overall size of the bitstream. The precomputed data is transmitted as "Voxel Data" (i.e., a key-value storage where voxel coordinates are used as keys for looking up data sets like path lists). Efficient encoding of this data is achieved by leveraging, for example, Voxel Coordinate Prediction, Inter-Voxel Redundancy Reduction, and Generic Codebooks that provide a flexible solution for entropy encoding of data series like ER and diffraction path lists. It allows for switching between fixed-length encoders and variable length encoders with individual codebooks that are transmitted together with the encoded series of symbols.

4.3 Listener Space Information

The MPEG-I Immersive Audio LSI file format [36] provides a structured approach for describing the geometry and acoustic properties of physical listening spaces to support immersive audio rendering in AR applications (note that LSI was termed "LSDF" in earlier publications, e.g., [8]). By specifying elements such as room boundaries, acoustic materials, reverberation parameters, loudspeaker configurations, and anchors, the LSI enables accurate alignment of virtual audio scenes with real-world environments. The format further facilitates integration of locally captured audio streams and object sources, ensuring adaptability to diverse AR and VR scenarios such as rendering of a device user's own voice. This direct provision of environmental data to the renderer enhances the realism and spatial fidelity of AR audio experiences, representing a significant advancement in the field of immersive audio technology.

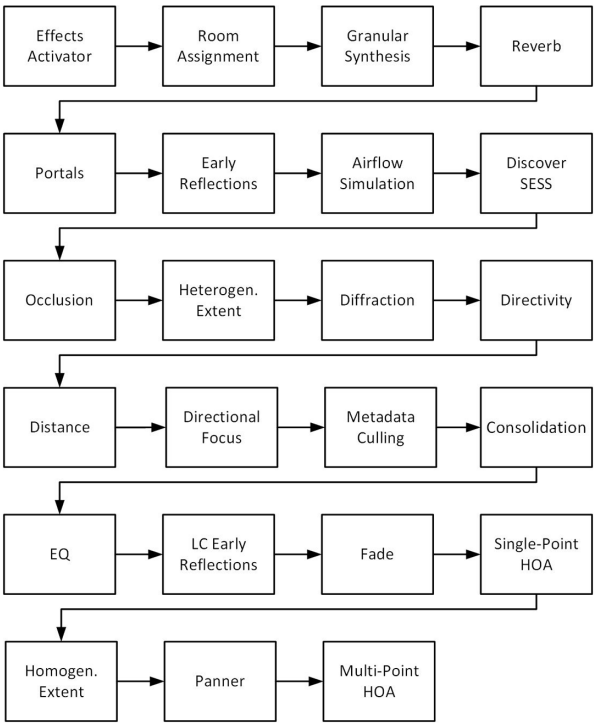


Fig. 4. MPEG-I audio renderer pipeline.

5 RENDERER PIPELINE TECHNOLOGY

5.1 General

The renderer pipeline auralizes all relevant acoustic effects. Its stages are shown in Fig. 4. Each stage can access and modify the list of RIs by adding, deleting, or deactivating them. The stages operate in a designated order, each fulfilling a distinct role that does not overlap with subsequent stages. Some stages prepare elements for further processing without modifying audio samples, while others, typically later in the pipeline, perform actual audio processing tasks.

5.2 Effects Activator

The standard generally relies on continuous audio signals, synchronized with animations in the visual content, but also supports sound effects that can be triggered by actions of the user or another local process. To optimize computational complexity, the sound source associated with the sound effect will not be rendered until the sound effect is triggered. With the duration of the sound effect being encoded as part of the bitstream, the stage can disable the sound source once the rendering of the sound effect has ended, freeing up computational resources, and without requiring knowledge of the content of the sound effect.

5.3 Acoustic Environment Assignment

The AE assignment stage updates metadata of RIs to reflect the current scene configuration regarding its AEs. This provides information on which signals are currently relevant for the late reverberation processors and which reverberation extent items are audible. A default AE can be applied in scene parts without an explicitly defined AE to create ambient, reverberation-like echoes whose parameters can be signaled in the bitstream.

5.4 Granular Synthesis

This stage generates procedural audio (i.e., sound that is generated in real-time in response to dynamic control input parameters) using a form of granular synthesis where several aspects of the sound can be controlled dynamically in real time [46]. This is useful for generating sounds that can react to user input or other changes in the virtual scene, such as vehicle sounds or the contact sound of two objects sliding against each other. The rendering is based on precomputed databases of sound segments, referred to as grains. To create such a database, original recordings are partitioned into segments, which are then stored in a multidimensional database where each dimension represents different aspects of the sound that can be controlled dynamically (e.g., the RPM of an engine). At rendering time, the sound is controlled by updating a target position in the multidimensional database and consecutive grains are selected for rendering using a weighted random selection where the weights depend on the target position, previously used grains, and other parameters of the granular database.

5.5 Portals

Portals designate spatial regions where sound can propagate between two environments. This is used for two purposes: multiroom reverberation rendering, as described in SEC. 5.6, and coupled source rendering. In the reverberation stage, the portal metadata is used to determine the energy contributions from external sources and the reverberation from connected AEs through the portal. The generation and management of the reverberation of extent sources is handled within the portal stage.

A coupling source simulates the effect that even without an airborne path between environments, occluded sources can still be audible from their energy inducing vibrations in an occluding material. If partial occlusion (as discussed in SEC. 5.10) were used to simulate this, a clearly localizable sound from the source's position behind the wall would be heard, whereas a coupling source localizes the sound across the coupling surface (typically a door or window). In the MPEG-I Immersive Audio standard portals are used to indicate where such coupling can occur. For portals with material having a nonzero coupling coefficient, the portal geometry is used to create a (homogeneous) extent source—a coupling source—that re-emits the energy of incident source(s). See Fig. 5. The user can hear the coupled energy emerge from the portal extent, rather than the original source itself. This effect can be combined with direct transmission, if desired, by setting the material properties correspondingly.

5.6 Reverberation

The reverberation stage simulates late reverberation for AEs using efficient feedback delay networks (FDNs). For virtual reverberating environments, the FDN is configured from the information in the bitstream, while for environ-

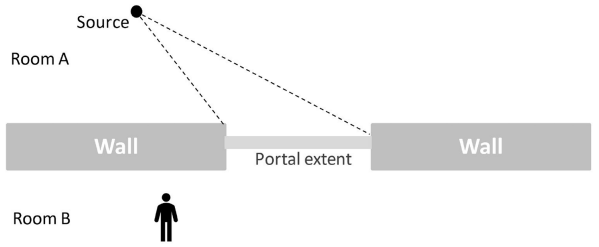


Fig. 5. Cross-section of portal with material properties, used as a coupling source.

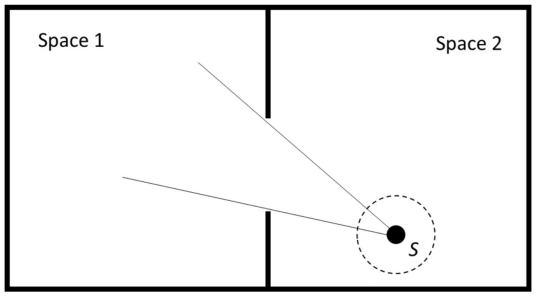


Fig. 6. A portion of the emitted source energy contributes to Space 1 and not to Space 2.

ments representing an AR listener environment, the FDN is configured slightly differently using the metadata and optionally a room impulse response (RIR) in the LSI. Reverberation time and level are controlled in a frequency-dependent manner. Furthermore, the feedback delays are adapted to the dimensions of the environment for a realistic reverberation effect. Each FDN is fed with a mono downmix of all audio sources.

The downmix process is controlled by how much energy each source contributes to the corresponding AE. When a source is inside the environment, this is the total source energy reduced by the energy lost through portals (see SEC. 5.5), while for external sources, this is the portion of the source energy reaching the environment through portals. With a source's energy emitted in all directions, part of that energy can go through a portal as illustrated in Fig. 6, a part of the source's energy is contributing to Space 1's reverberation and not to Space 2's.

Contributions from external sources are an important aspect for supporting multiroom reverberation in scenes with multiple connected environments. Two further concepts are present to realistically render multiroom reverberation. Firstly, reverberation for the environment containing the user is rendered immersively as a diffuse sound field reaching the listener from all directions. Reverberation from other environments is mapped to (homogeneous) extended sources (called reverberation extent sources) around the portals of the respective environment, such that the listener perceives those reverberations directionally from the location of the portals, and will be filtered by any occluding geometries such as doors or windows. Fig. 7 shows the cross-section of an AE with a reverberation extent geometry that is used to represent the reverberation of the room as a directional, extended source for user positions outside the AE.

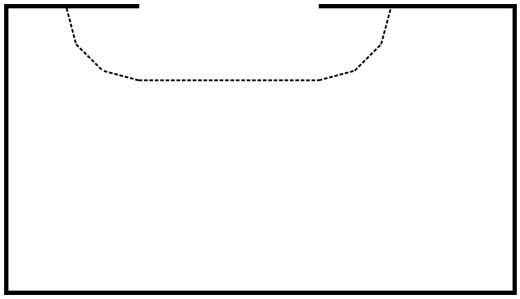


Fig. 7. Cross-section of AE with reverberation extent geometry at its portal.

Secondly, reverberation of one environment can reach another environment through the same portals and contribute to its reverberation, and vice versa. This effect is supported by feeding reverberation signals from each FDN into FDNs of all environments it is directly connected to via portals. These different multiroom reverberation aspects are integrated such that they create a smooth experience as a source or user moves through the environment. The source contribution of an external source naturally increases to half the energy as it approaches a portal, and the remaining energy in the source's environment decreases accordingly, causing a smooth transition of reverberations as the source crosses the boundary between two environments.

Similarly, with the reverberation extent sources placed at a distance from the portal (Fig. 7), the fully immersive playback of the reverberation is exchanged with the directional playback from the reverberation extent source as a user moves over this distance from the extent to the portal, resulting in a smooth transition. For large portals, this distance is larger to obtain an accordingly realistic, more gradual, effect. The multiroom aspects work similarly for connected VR environments as for virtual environments connected to a physical listener AE in AR scenes.

5.7 Early Reflections

The ER stage models low-order specular reflections on planar and curved surfaces of the given scene geometry. ERs are known to be important for auditory spaciousness, immersion, and speech and music clarity. Reflections on the faces of a triangle mesh are generated by the image source method [47]. This method results in a good approximation of sound reflection in indoor scenes, where most surfaces are large and perpendicular. If geometry data approximates curved surfaces by triangles or rectangles, the classic geometrical acoustics methods are no longer valid and artifacts become audible. The resulting "disco ball effect" is illustrated in Fig. 8. For a moving listener or a moving sound source, the visibility of the image source will alternate between visible and invisible, resulting in a permanently switching localization, timbre, and loudness. To mitigate this effect, a geometry preprocessor detects curved surfaces (i.e., "round edges" and "round corners") and an extended image source model is applied. If the listener is located in between the visible areas (white area in Fig. 8), then the round edge/corner is used as a reflection point.

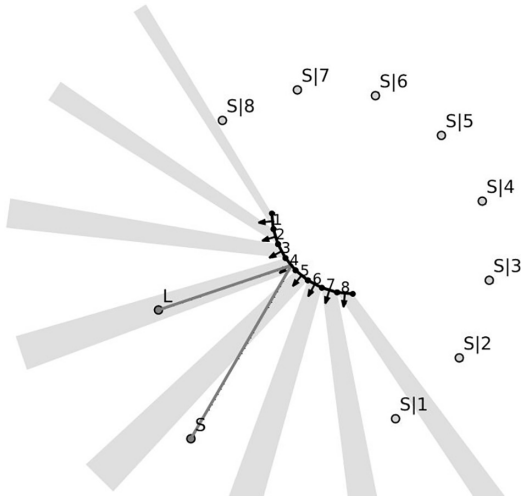


Fig. 8. Visualization of the *disco ball effect* (solid lines: reflecting surfaces; S|n: image sources; S: sound source; L: listener; gray area: area where the image sources are visible).

If geometry is represented by geometric primitives like spheres and cylinders, then also the changed distance attenuation caused by the scattering can be taken into account. The sound pressure of a point source and its reflection on planar surfaces follows a 1/r law where r denotes the propagation distance. This distance law no longer holds for reflections on spheres or cylinders, because of the lens effect (mirrored objects on spherical surfaces appear to be more far away, while mirrored objects on cylindrical surfaces appear to be compressed in one dimension). This property is taken into account by computing per dimension a magnification factor and using a derived "apparent distance" for the distance attenuation of the image source. In addition, the frequency dependent reflection filter is adjusted to not only take into account the material but also the radius of the geometric primitive [48, 49].

The calculation of ERs using the image source method can be demanding in terms of computational complexity. A low-complexity alternative for generating them is described in SEC. 5.19.

For dynamically updatable complex environments that are authored in voxel mode, the *voxel-based ER stage* offers a computationally efficient method to generate ERs. This method employs a uniform cuboid grid representation to compute first-order reflections without requiring precomputed encoder data. The algorithm utilizes segment grid traversing to verify potential reflection paths between a sound source and the listener. Origin points are established along the source–listener line segment, from which rays are cast in all directions covering the full 3D space to detect ray–voxel collisions. For valid reflection paths, the algorithm confirms that both segments (source-to-reflection and reflection-to-listener) remain unobstructed. Reflection paths that are too short or too long can be discarded based on configurable delay thresholds. The remaining valid paths are clustered into angular sectors, creating virtual audio sources that represent acoustically relevant ERs while meeting computational complexity constraints.

5.8 Airflow Simulation

This stage simulates the sound heard by a listener when inside of an airflow, such as an atmospheric wind. As an example, when listeners are standing on a hillside with a wind blowing, they hear not only the sound of the wind in the trees, but also the sound of the wind passing around their head and over their ears. This sound depends on the speed of the wind and the angle relative to the listener's head, changing as the listener moves and rotates. This is simulated using efficient filtering of pink noise to increase immersion and plausibility, without requiring complex fluid dynamics simulations [50].

5.9 Discover Spatially Extended Sound Sources

This stage provides metadata necessary for rendering spatially extended sound sources (SESS). It uses ray tracing to sample the section of a source's extent geometry visible to the listener. Effectively, this provides the projection of the geometry onto the surface of a sphere centered in the listener. The result speeds up the rendering process by providing finite samples to check for occlusion and synthesis of extent binaural output. Both processes are handled by stages later in the pipeline [i.e., the occlusion stage (SEC. 5.10) and the homogeneous extent stage (SEC. 5.22)].

5.10 Occlusion

The occlusion stage uses ray tracing to determine whether the direct line of sight from a listener to a source is occluded. The occlusion status is defined by a set of acoustic materials retrieved from the occluding geometry. An occlusion EQ is generated by combining the transmission coefficients of the occluding materials and applied to the RI. This allows for arbitrary occlusion states (e.g., not, partially, or fully occluded). For extended sound sources, this process is applied for each of the samples produced by the Discover SESS stage (SEC. 5.9). The occlusion results are grouped by their set of occluding materials to be processed by the homogeneous extent stage (SEC. 5.22). This allows for arbitrary combinations of occluding materials on any number of sections of the visible surface of the extent.

The *voxel-based direct occlusion stage* offers a version of the stage that works on the 3D voxel grid. Using an efficient segment traversing technique, the algorithm analyzes all voxels on a line segment between listener and source and determines material type and position of occluders. An occlusion EQ is calculated based on material types encountered along the path. The activation of occlusion can be configured selectively per each audio source and occluding material type.

5.11 Heterogeneous Extent

This stage renders audio sources that have an associated spatial extent of defined size and shape and an input signal with more than one channel, representing a nonuniform spatial source character. Specifically, it handles the rendering of object sources with an extent and a multichannel source signal (supporting up to nine channels) as well as HOA sources with an associated extent. The stage provides

rendering of both a so-called exterior representation of the heterogeneous extended sound source when the listener is situated outside of the extent and a so-called interior representation of the extended sound source when the listener is inside the extent.

The exterior representation can be derived from a provided interior representation, or, conversely, the interior representation can be derived from a provided exterior representation. There is an interpolation between the rendering of the two representations whenever the listener moves between the inside and outside of the extent of the source to provide a smooth transition [51]. The rendering of heterogeneous extended sources is based on a set of virtual loudspeakers that are adaptively arranged so that they optimally represent the sound source at any given time, as the position of the listener changes relative to the sound source.

5.12 Diffraction

The diffraction stage models sound propagation around occluding objects (e.g., around corners) by deriving relevant diffraction paths from preprocessed geometrical data and line-of-sight occlusion checks. The resulting calculations are then used to generate secondary RIs at the diffraction edges closest to the listener. This approach creates perceptually plausible diffraction effects by using perceptually motivated, diffraction angle–dependent filters. Multipath diffraction modes are available to support complex geometries. In the case of the *voxel-based diffraction stage*, the diffraction path and corner detection can be determined at run time, using a jump-point-search path-finder algorithm. Higher-order diffraction modeling, the aggregating diffraction around multiple corners, is supported.

5.13 Directivity

The directivity stage models source directivity by applying direction-dependent radiation patterns typical of natural sound sources, such as human voices or musical instruments. It incorporates the corresponding directional equalization into the central EQ fields of the relevant RIs, which accumulate all EQ effects for subsequent application by the EQ stage (SEC. 5.18).

5.14 Distance

The distance stage simulates sound traveling from the source to the listener by addressing three key factors. First, it models a physically accurate *propagation delay* using a variable delay line with subsample spline interpolation and smoothing. The smoothly interpolated dynamic adjustment of propagation delay implicitly synthesizes a high quality Doppler effect. Second, it calculates signal attenuation caused by the geometrical spreading of the source energy. Third, it determines attenuation due to air absorption. Both attenuation effects are accumulated in the central EQ fields of the RIs. For sources with an extent, the attenuation due to the geometrical spreading is calculated from a physics-based model that takes the size of the extent into account [51]. This results in an attenuation-versus-distance behavior that accurately mimics the real-world behavior of differently sized sources. Optionally, a perceived location offset between visual position of an audio object and its auditory localization (due to different propagation delays of light and sound) can be modeled.

5.15 Directional Focus

The directional focus is an important part of the accessibility features (SEC. 7.1) of the renderer. It leverages the reduction of distracting sounds coming from areas outside a specified spatial region of interest to assist persons with hearing loss. The focus has a radial symmetry with a single main lobe. Its attenuation characteristics can be individually adjusted using parameters available in the accessibility interface. The directional focus attenuates the broadband gain of selected RIs. The gain attenuation is a function of the listener-relative azimuth and elevation angles of the RI. By default, the focus is directed toward the user's front viewing direction, but it can be adjusted to other orientations, allowing control through various services or input methods such as handheld controllers and eye trackers.

5.16 Metadata Culling

This stage deactivates RIs that become inaudible due to very low gain mandated by the current scene state (e.g., due to attenuation or occlusion) and thus reduces computational complexity. If loudness data for a source is available, loudness-based culling is also supported. Under certain circumstances, ERs might be perceptually masked by their source signal. A dedicated culling of such ERs is additionally supported.

5.17 Consolidation

The spatial resolution of human hearing and the ability to distinguish concurrent sound source positions is limited. The consolidation stage reduces the number of RIs and thus computational complexity while maintaining perceptual quality by combining RIs in positions that can perceptually not be distinguished. Groups of RIs that may be combined are identified based on a psychoacoustic model for perceptual localization thresholds via a hierarchical clustering algorithm. The respective audio signals of consolidated RIs are mixed, and the RIs are spatially combined into a representative RI at a consolidated position.

5.18 Equalizer

Global and frequency-dependent gains that were attributed to an RI in one of the preceding stages corresponding to acoustic effects, like occlusion, air absorption, etc., are accumulated in the central EQ fields of the RIs. In the EQ stage, minimum-phase FIR filters are designed based on gain parameters with third-octave resolution and applied to the corresponding RI signals to auralize the accumulated gains.

5.19 Low-Complexity ER

The geometric room acoustics model to generate ERs (SEC. 5.7) requires a high computational complexity that increases with the number of sound sources and desired

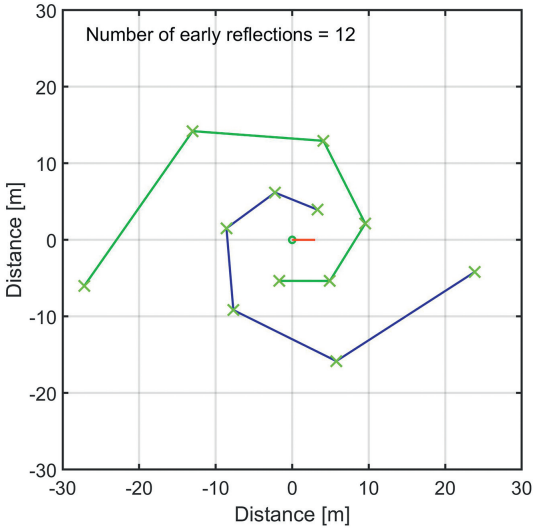


Fig. 9. Indoor LC-ER pattern plotted across location (top view).

reflection order. However, the exact simulation of RIRs is often less important for VR/AR applications. This alternative stage uses a simplified fixed ER pattern for a perceptually plausible and convincing experience at very low computational complexity. The number and positions of ERs are derived from simple room acoustic parameters as input, using different patterns for indoor and outdoor. The indoor low-complexity ER (LC-ER) pattern is presented in Fig. 9. In this way, only one common ER pattern is needed to process any number of sound sources in a scene.

The reflection model incorporates distance-dependent attenuation, room volume dependency, and automatic indoor/outdoor scene classification to appropriately adjust a static reflection pattern by number of reflections, spacing in time, positioning in space, and gains. It was proposed, discussed, and evaluated in [52]. Listening tests confirm that LC-ERs are equally plausible as a costly dynamically updated image source model. The bitrate reduction is shown to be 25% on average across all tested indoor scenes, and the overall relative complexity reduction is 39%, averaged across all nine tested scenes.

5.20 Fade

The fade stage performs fade-in and fade-out of the audio signals when their corresponding RIs are activated or deactivated [e.g., by the metadata culling (SEC. 5.16), by consolidation (SEC. 5.17), or during teleportation of the listener]. Activation/deactivation of RIs is performed in a coordinated manner across stages, such that stages continue processing of the RIs' audio signals until the fade-out has completed.

5.21 Single-Point Higher-Order Ambisonics

The SP-HOA stage provides 3DoF binaural output for HOA signals. A magnitude least-squares decoder is calculated during initialization based on the provided HRTFs [53]. During rendering, HOA sources are rotated according to the listener head orientation and decoded using the magnitude least-squares decoder to binaural output.

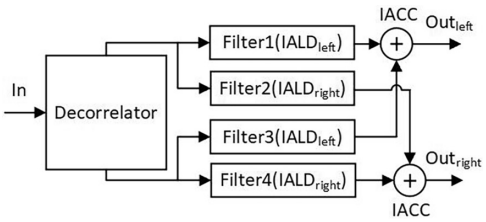


Fig. 10. Filter arrangement for the binaural synthesis of homogeneous extended sources.

5.22 Homogeneous Extent

The stage efficiently auralizes SESS that have a size and uniform sound characteristics directly to the binaural output using a set of adaptively designed filters. The target auditory cues are synthesized directly to binaural signals from just two decorrelated input signals using preprocessed tables derived from the HRTF data. Preprocessing is possible because no signal-dependent design is required and the employed processing filters solely depend on the desired spatial extent range. This approach is much more efficient than, for example, modeling extent by generating several decorrelated point sources spatially distributed over the desired extent range and rendering each of these point sources individually.

The process begins by identifying the binaural cues to be created by the SESS at the ear signals, namely interaural cross correlation (IACC), interaural phase differences, and IALDs. These are calculated in real time from the preprocessed HRTF data, the samples provided by the discover SESS stage, and their occlusion information. To synthesize IACC, a decorrelator based on all-pass filters and time-frequency variant envelope shaping is used [54]. The binaural output is produced by mixing the original and decorrelated signals using a set of four FIR filters that are adaptively designed dependent on the identified cues. The mixing as depicted in Fig. 10 is implemented by a set of four FIR filters where the EQ gains are determined by the IALDs and the mixing ratio between signals is determined by the IACCs [55].

5.23 Panner

The panner stage positions the immersive late reverb outputs of the reverberation stage (SEC. 5.6) into the output sound image using Vector Base Amplitude Panning [56] with additional controls, such as spatial spread. The stage implements high-quality triangulation of the immersive loudspeaker setup and ensures that the triangle edges do not cross the horizontal plane, which improves spatial localization.

5.24 Multipoint Higher-Order Ambisonics

The MP-HOA rendering approach in MPEG-I Immersive Audio supports 6DoF binaural output for scenes captured with one or more HOA sources. The system enables efficient capture, streaming, and 6DoF rendering of complete audio scenes using a sparse array of spatial microphones, without requiring close-up microphones at the sources. The

method employs interpolation of spatial parameters determined at microphone positions to estimate spatial parameters at the listener's location [57]. These interpolated spatial parameters are used to construct target binaural signal statistics for 6DoF rendering. The time-frequency domain recording from the microphone nearest to the listener is then adapted to match the desired binaural cues.

The method works best within the area defined by the microphone array positions but also supports plausible rendering at positions outside of the area, referred to as exterior rendering. In exterior rendering, the spatial parameters are first interpolated at the edge of the microphone area and then modified based on a projection radius and the listener distance from the microphone area edge [58]. A similar principle is used to handle the special case of 6DoF rendering using a single HOA source.

To enhance the clarity of relevant sound sources in complex immersive scenes, source locations may be provided to the renderer as metadata to enable so-called informed-mode rendering. The informed-mode rendering enables the renderer to construct more accurate target binaural signal statistics by estimating sound source characteristics via beamforming toward the known source positions. The informed mode rendering greatly improves the localization and distance attenuation performance for the sound sources of interest. A low-complexity MP-HOA stage provides binaural output from a set of HOA microphone signals using virtual loudspeaker objects derived from the HOA signals.

6 SPATIALIZER

6.1 Binauralization

The binauralizer positions the individual point source RIs into the binaural output signals. It selects the nearest HRIR in the provided HRIR dataset for a given RI position and convolves its associated audio waveform with the HRIR stereo filter through a fast partitioned convolution. A cross-fade is used to ensure smooth filter changes between positions. When the low-complexity mode is activated, in a first step, the signals of all RIs are panned to predefined representative virtual loudspeakers based on the RI positions and the virtual loudspeaker positions relative to the listener. For panning, precomputed time shifts and gains for left and right outputs are applied to the RIs. In a second step, the panned signals are summed at the representative positions and are convolved with HRIRs according to the listener-relative locations of the representative positions. The resulting binaural signals are then summed to form the auralized output signals.

6.2 Tracked Loudspeaker Rendering

While most VR/AR applications are expected to employ binaural head-tracked audio rendering, MPEG-I Immersive Audio also provides the option to render to spatial loudspeaker setups. Multichannel loudspeaker setups have a dedicated listening position, the so-called sweet spot, where optimal spatial perception is achieved. Moving with 6 degrees of freedom in a VR scene, while using a fixed loudspeaker setup in the physical reproduction space, will inevitably compromise the spatial perception severely and break any stable localization unless there is a compensation mechanism. To adapt the sweet spot of the loudspeaker setup to the current listener position (and avoid audio image collapse toward the nearest speaker), the position of the user is tracked in real-time.

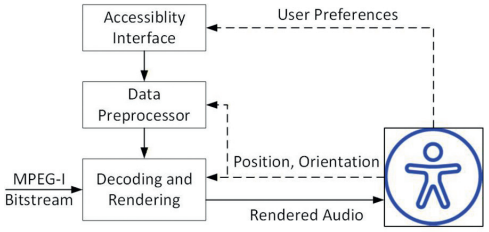


Fig. 11. Accessibility architecture.

Sweet spot adaptation is achieved by adaptive adjustments of level and delay [59] while maintaining appropriate direct sound levels [60]. Also directivity of high-frequency loudspeaker radiation can be corrected, utilizing a distance-dependent model of the loudspeaker in the listening room. This is referred to as the *physical compensation level* which can be utilized with any spatial loudspeaker arrangement, ranging from 2.0 to 22.2 channels. It allows the listener to move within the setup while maintaining a stable spatial sound image. Tracked rendering is thus an attractive feature for the playback of legacy audio content in that it both extends the sweet area dramatically and "detaches" the sound from the loudspeakers [61].

Based on these provisions, virtual scenes are rendered by the so-called *object rendering level*. A real-time adaptive 3D amplitude panning algorithm is employed that operates within loudspeaker layers, such as horizontal and height layers. Each layer utilizes a 2D panning algorithm to project the object onto the respective layer. The final 3D object is rendered by amplitude panning between the two virtual objects generated from the 2D panning in the two layers. Sound sources can be positioned both outside and inside the loudspeaker setup, with interior sources accessible to the listener from all directions. The target loudspeaker setup is input to the renderer via the LSI interface as described in SEC. 4.1.

7 APPLICATION ASPECTS

7.1 Accessibility

With the promise and aim of MPEG-I of delivering immersive experiences and connecting people via Social VR, the technology needs to be accessible to users of all ages and hearing abilities. Due to constantly increasing population aging, a growing number of consumers are affected by various degrees of hearing impairment, yet only about 14.2% of the people above age 50 years with a clear hearing impairment are actually using a hearing aid [62]. To account for this, MPEG-I rendering offers a range of accessibility enhancement options. Dynamic scene adjustments during rendering (as shown in Fig. 11) are possible via attenuation of ERs (SECS. 5.7 and 5.19) and reverberation

(SEC. 5.6), expansion or compression of distance-dependent gains (SEC. 5.14), directional focus (SEC. 5.15), and spectral compensation. Compared to hearing aids that can only process the rendered binaural mix under extremely constrained computational resources, accounting for accessibility directly within the MPEG-I rendering provides clear performance and quality improvements.

7.2 Social Virtual Reality Support

Social VR applications using the MPEG-I Immersive Audio standard promise to reshape virtual interactions by delivering immersive audio experiences that connect users in unprecedented ways. One of the key features of this standard is its support for external audio sources, thereby significantly enhancing the realism and engagement of virtual environments. This capability allows users to incorporate live audio feeds, such as music or voice from remote participants, into their virtual experiences. Additionally, users can have their own voices enabled to be heard in the virtual scene, further enhancing the sense of presence and personal connection. By blending real-world audio with the virtual environment, users can engage in social interactions that feel genuine and connected, regardless of their physical location.

Since the standard enables real-time auralization of room acoustics, it is ensured that all included sounds are spatially accurate and contextually relevant [63]. For instance, users can enjoy a virtual concert with live performances streamed directly into the VR space or gather in collaborative workspaces with their voice inputs rendered within the scene. Furthermore, including musical instruments as external audio sources enables the development of applications that facilitate jamming with other musicians in a virtual environment.

In Social VR applications, individual participants receive their personalized audio output from their individual renderers instances, while multiple renderers are interconnected via network, sharing and exchanging audio streams, avatar poses, and scene data. As the MPEG-I Immersive Audio renderer ingests decoded audio signals as PCM samples, it is not constrained per se by any specific audio coding technology. This enables, for instance, rendering from audio streams that are coded with 3rd Generation Partnership Project communication codecs such as Enhanced Voice Services [64] and Immersive Voice and Audio Services [65]. Specifically, this is important to facilitate low-latency communication with reasonable response times in conversations and user interaction.

7.3 Real-Time Acoustic Scene Manipulation

The MPEG-I Immersive Audio standard enables real-time dynamic scene modifications, either through user interactions or driven by a scene engine. In both available flavors of geometry descriptions, mesh and voxel mode, changes in the scene that affect the sound rendering can be triggered via scene updates (e.g., to model opening and closing of windows or doors). Such changes to a scene's geometry are instantaneously conveyed to the audio rendering, updating, for example, ERs, occlusion, and diffraction behavior. However, changing geometry freely by the user within the scene is uniquely possible through MPEG-I's voxel-based acoustic modeling approach, which allows prompt updates through voxel scene state modifications. For example, when walls are built, modified, or destroyed, the relevant voxel state transitions can be applied within a single update frame while avoiding complex geometric recalculations required by traditional ray-tracing methods. Voxel scenes can not only be manipulated but also constructed fully from scratch, adding reflective surfaces, changing materials, and parameters just as the visual scene is being generated.

7.4 Renderer Latency Considerations

An important application aspect is the latency introduced by the rendering process [66]. Especially in Social VR applications, the render's audio input-to-output latency is important: for local audio input, consisting of either the user's own voice or the input of a live musical instrument, a sufficiently low latency is one of the most important prerequisites for satisfactory user experience and a plausible immersion. Another fundamental latency value is the motion-to-sound latency (e.g., the delay of the renderer to adapt the output audio signal to a change of listener position or orientation in 6DoF space [67]). For example, to achieve a precise source localization and a stable spatial image perception, changes in listener's head rotations need to be handled without perceivable time lag.

In the MPEG-I Immersive Audio renderer, these delay values are mainly dependent on the sampling rate, the internal audio block rate and the control-step rate (SEC. 4.1). All three parameters can be configured and allow an optimization of the renderer latencies. In its default setting, with a sample rate 48 kHz and a block size of 256 samples, the renderer has 480 samples input-to-output latency (10 ms) and, assuming a control-step rate of 60 Hz, a maximum motion-to-sound latency of 832 samples (17.3 ms) if the update of incoming motion data is delayed by one full control step.

These values can be reduced by choosing a high sample rate, a small block size, and a fast control step rate. In practice, the choice will often be a trade-off between latencies, perceptual quality, and complexity. For example, more frequent control steps can improve the perceptual quality through a smooth and accurate adaptation of the rendering to movements of audio sources and listener within the scene and will reduce motion-to-sound latency. But, in turn, this will increase computational load due to more frequent parameter recalculations. As another example, a lower sampling rate will reduce computational complexity but will increase input-to-output latency at a given block size and provide a lower audio bandwidth that may have an effect on the perceived audio quality.

In applications, additional delays need to be considered, such as sound hardware latency, preprocessing, and software driver buffer delays [63].

7.5 Computational Complexity and Platform Compatibility

During development of the MPEG-I Immersive Audio standard, the specified technology was implemented as a plug-in for MaxMSP [68] and as a stand-alone offline renderer, both of which are part of the reference software specification [36]. The focus of the reference software implementation was on flexible testing and evaluation of different rendering technologies; hence, the final computational complexity of an optimized implementation can be difficult to gauge.

To assess the computational load of the reference software renderer, a dedicated complexity measurement tool was developed and integrated as an additional feature into the reference software. The tool quantifies the processing time consumed by each stage of the renderer pipeline, both for parameter update steps and real-time audio processing steps. Naturally, the computational workload of a 6DoF renderer substantially depends on the richness of a scene (e.g., number of active audio sources, detail level of scene geometry, etc.) and the actual position and movement of the user within the scene.

The complexity measurement tool can process recorded scene task trajectories for individual scenes and is able to evaluate a batch of these scenes automatically for a representative and reproducible measurement. This detailed measurement data provides valuable analysis information of the distribution of average and peak workload among all stages separated into update and audio steps and its dependency on scene content. Insights gained with the complexity measurement tool led to several code optimizations for specific stages already during the technology development of the standard. In the future, the tool will support further optimization of implementations on end-user devices.

As a proof of concept, the MPEG-I Immersive Audio reference renderer has already been successfully validated on multiple major extended reality (XR) platforms, showing the standard's readiness for broad adoption in commercial applications. The renderer software has been tested on Apple's visionOS running on head-mounted displays like Apple Vision Pro [3] and Android-based devices such as the Meta Quest 3 [4] and PICO 4 [5] headsets. Integration into XR applications has been facilitated both directly as a native library for Microsoft Windows and Android and as an AudioUnit plug-in utilizing the JUCE audio framework [69], allowing for integration into more advanced audio processing pipelines.

By setting up these exemplary systems and demonstrators, the standard has undergone extensive testing and validation in real-world usage, and no performance bottlenecks or stability problems were observed in principle during these testing periods. The standard can deliver the intended high-quality spatial audio experience it was designed to enable, even under the constraints and requirements of contemporary XR platforms.

The deployment process has also shown excellent compatibility of MPEG-I audio processing with the devices' native spatial computing capabilities: the MPEG-I renderer has been demonstrated to run on several hardware platforms (PC, Meta Quest 3, Apple Vision Pro) at MPEG meetings and, publicly, at a workshop of the 159th AES Convention in Long Beach, CA.

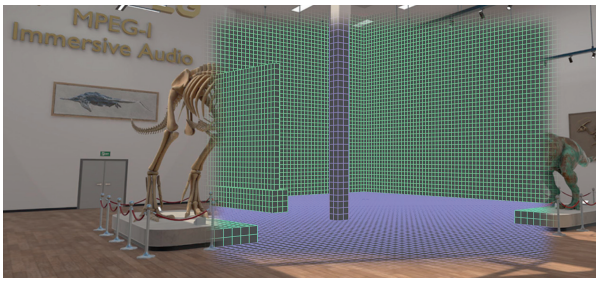


Fig. 12. Example of a VR scene visualization augmented with a voxel-based geometry representation used for MPEG-I Immersive Audio rendering technology.

Several MPEG-I Immersive Audio features exceed what is currently possible with default audio rendering engines, and it has shown to be straight-forward to make use of these features in combination with the devices' native visual and sensory capabilities.

7.6 Audio Scene Authoring

The MPEG-I Immersive Audio standard is engineered for direct integration into existing XR scene creation and authoring workflows. This integration requires the conversion of the original scene and scene asset formats, such as glTF, FBX, and USDZ, directly into the MPEG-I bitstream or through an intermediate format known as the MPEG-I Immersive Audio EIF [36].

The primary objective of conversion into the EIF is the elimination of acoustically irrelevant elements from the audio scene description, effectively transforming the EIF into an audio-centric scene description format that facilitates audio scene authoring. The EIF is an informative XML-based specification that references audio assets, scene geometry, sound environment descriptions, and other acoustic parameters. The format comprises static and dynamic entities, each identified by a unique ID. This scene description supports dynamic modification through atomic updates that alter entity properties based on time-based triggers, conditional logic (e.g., user proximity), or external scene events and user interactions. The conversion process to EIF encompasses scene geometry simplification and the incorporation of additional acoustically relevant data, including content creator intent settings. The conversion of complex visual scenes into EIF and the MPEG-I Immersive Audio bitstream can be performed automatically or semi-automatically using platforms such as Blender and Unity. Dedicated plugins facilitate the creation, visualization, and editing of MPEG-I Immersive Audio content.

Existing scene and 3D asset geometries can be adapted to the MPEG-I Immersive Audio format through remeshing or voxelization techniques. This process eliminates fine geometric details that produce a negligible acoustic impact on the rendered audio. Fig. 12 illustrates this approach, showing a visually detailed museum VR model and a simplified voxel representation consisting of only large elements such

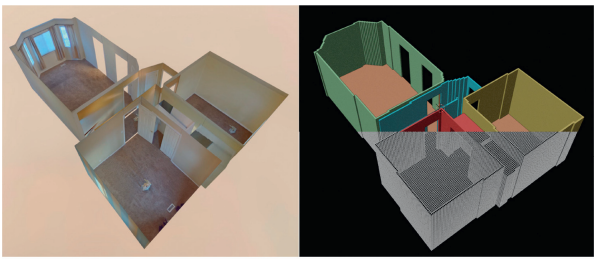


Fig. 13. Example of a 3D mesh data (a) and the corresponding voxel data (b) obtained from real environment images.

as walls, floors, and pillars. The resulting conversion and voxelization results can be further refined using specialized voxel editors such as MagicaVoxel or dedicated Blender and Unity plugins. EIF scene description can also be created directly from available low-polygon mesh models and voxel-based geometric 3D assets, bypassing the remeshing and conversion process entirely.

The MPEG-I Immersive Audio renderer can also leverage direct access to geometric data from 3D computer graphics engines like Unity or Unreal Engine. For instance, voxel-based platforms (e.g., Minecraft-like applications and games) can provide 3D voxel matrices (voxel chunks) directly for audio rendering purposes. Such an integration between scene engines and MPEG-I Immersive Audio rendering is essential for procedurally generated XR scenes. It also enables real-time assessment of the audio experience during scene creation and modification through user interactions.

The foundation for generating comprehensive visual and acoustic models lies in scanning of real environments. For example, Fig. 13(a) shows a multiroom 3D model, which enables user virtual navigation. This specific model was generated using combined information from five 360° image panoramas [70]. The resultant 3D mesh data is then automatically voxelized and converted in the format required by the MPEG-I Immersive Audio encoder and renderer [see Fig. 13(b)].

The selection of specific scene authoring tools and integration workflow approaches depends heavily on the intended application use-cases of the standard.

8 SUMMARY

The MPEG-I Immersive Audio standard was finalized in November 2025. This article sketches the background and development history of the specification work starting in 2017 and presents the concepts, supported functionality, and underlying technology in detail. It discusses approaches, methodologies, and scientific questions related to perceptual quality assessment of 6DoF audio. Specifically, it outlines how MPEG-I Immersive Audio was tested in a final verification test that needed to balance scientific rigor with industry relevance while acknowledging the nonexistence of a physical reference by formalizing the concept of "plausibility."

Additionally, insights into important practical implementation aspects are provided such as the built-in support for accessibility features, the processing of incoming real-time audio streams (e.g., for Social VR applications), and the on-the-fly acoustic scene manipulation capability. Finally, renderer latency considerations, computational complexity, and platform compatibility are discussed as well as concluding thoughts on audio scene authoring of MPEG-I Immersive Audio content.

As a nonproprietary standardized and open technology, it sets the foundation for the first long-term stable audio format specification in the VR/AR sector, allowing for interoperability of devices of different manufacturers. It is applicable to a wide range of consumer uses, including metaverse technologies and Social VR. Its metadata bitstream of efficiently compressed acoustically relevant geometry and associated side information, and the flexible transport concept based on MHAS packets enables the adaptation to VR/AR/XR broadcasting and streaming applications. In summary, MPEG-I Immersive Audio offers high-quality, bitrate-efficient representation and rendering for VR and AR audio in 6DoF applications.

9 ACKNOWLEDGMENT

Technical development and standardization of MPEG-I Immersive Audio was a long-term effort of many people. Acknowledgment goes to collaborators of the authors for their invaluable contributions (in alphabetical order): Carlotta Anemüller, Eric Creusen, Danilo Dara, Christof Faller, Alejandro Restrepo Garcia, Matthias Geier, Otto Harju, Jaakko Hyry, Kari Järvinen, Patrick Kechichian, Mikhail Korotiaev, Mikko-Ville Laitinen, Ernests Lavrinovics, Sangmoon Lee, Arto Lehtiniemi, Vensan Mazmanyan, Michael McCrea, Vitor Menezes, Chamran Moradi, Okke Ouweltjes, Lauros Pajunen, Nils Peters, Archontis Politis, Kahleel Hassan Porter, Raksha Ramakrishna, Thomas Robotham, Dennis Rosenberg, Olli Rummukainen, Simon Schwär, Michael Singer, Jan van Sweevelt, Luuk Tripp, Juha Vilkamo, Frank Wefers, Niklas Winter, Yun-Han (Nathalie) Wu, and Mengqiu Zhang. Special acknowledgment goes to Dr. Schuyler Quackenbush and Dr. Thomas Sporer for their leadership of the standardization process and to the entire ISO/MPEG Audio work group.

10 REFERENCES

[1] HTC, "VIVE Pro 2 Headset," https://www.vive.com/us/product/vive-pro/ (accessed 8 May 2026).

[2] Microsoft, "Microsoft HoloLens 2," https://www.microsoft.com/en-us/hololens (accessed 8 May 2026).

[3] Apple, "Apple Vision Pro," https://www.apple.com/apple-vision-pro/ (accessed 8 May 2026).

[4] Meta, "Meta Quest," https://www.meta.com/ie/quest/ (accessed 8 May 2026).

[5] PICO Immersive, "PICO Ultra," https://www.picoxr.com/global/products/pico4-ultra (accessed May 2025).

[6] ISO/IEC, "Definition of MPEG-I Phase 2 and Subphases," *ISO/IEC JTC 1/SC 29/WG 11, N18970* (2020 Jan.).

[7] ISO/IEC, "Information Technology - High Efficiency Coding and Media Delivery in Heterogeneous Environments - Part 3: 3D Audio," *ISO/IEC Standard 23008-3:2019* (2019 Feb.).

[8] J. Herre and S. Disch, "MPEG-I Immersive Audio – Reference Model for the Virtual/Augmented Reality Audio Standard," *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 229–240 (2023 May). https://doi.org/10.17743/jaes.2022.0074.

[9] J. Herre, S. Disch, C. Borss, et al., "MPEG-I Immersive Audio: A Versatile and Efficient Representation of VR/AR Audio Beyond Point Source Rendering," in Proc. *AES 6th Intl. Conf. Audio for Games* (2024 Apr.), paper 19.

[10] S. Disch, L. Terentiv, J. Koppens, et al., "MPEG-I Immersive Audio – The Technology Of The New Standard For Virtual / Augmented Reality Audio," presented at the *159th Conv. Audio Eng. Soc.* (2025 Oct.), paper 10234. https://aes.org/publications/elibrary-page/?id=23078.

[11] J. Herre, S. Quackenbush, M. Kim, and J. Skoglund, "Perceptual Audio Coding: A 40-Year Historical Perspective," in Proc. *IEEE Intl. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, p. 1–5 (Hyderabad, India) (2025 Apr.). https://doi.org/10.1109/icassp49660.2025.10887760.

[12] M. Bosi, K. Brandenburg, S. Quackenbush, et al., "ISO/IEC MPEG-2 Advanced Audio Coding," *J. Audio Eng. Soc.*, vol. 45, no. 10, pp. 789–814 (1997 Oct.).

[13] ISO/IEC, "Generic Coding of Moving Pictures and Associated Audio: Advanced Audio Coding," *ISO/IEC Standard 13818-7* (1997 Dec.).

[14] J. Herre and M. Dietz, "MPEG-4 High-Efficiency AAC Coding [Standards in a Nutshell]," *IEEE Signal Process. Mag.*, vol. 25, no. 3, pp. 137–142 (2008 May). https://doi.org/10.1109/MSP.2008.918684.

[15] EBU, "EBU Evaluations of Multichannel Audio Codecs," Tech. Rep. 3324 (2007 Sep.). https://tech.ebu.ch/docs/tech/tech3324.pdf.

[16] J. Hilpert and S. Disch, "The MPEG Surround Audio Coding Standard [Standards in a Nutshell]," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 148–152 (2009 Feb.). https://doi.org/10.1109/MSP.2008.930433.

[17] ISO/IEC, "MPEG-D (MPEG Audio Technologies), Part 1: MPEG Surround," *ISO/IEC Standard 23003-1:2007* (2007 Feb.).

[18] J. Herre, H. Purnhagen, J. Koppens, et al.., "MPEG Spatial Audio Object Coding – The ISO/MPEG Standard for Efficient Coding of Interactive Audio Scenes," *J. Audio Eng. Soc.*, vol. 60, no. 9, pp. 655–673 (2012 Sep.).

[19] ISO/IEC, "MPEG-D (MPEG Audio Technologies), Part 2: Spatial Audio Object Coding," *ISO/IEC Standard 23003-1:2010* (2010 Dec.).

[20] M. Neuendorf, M. Multrus, N. Rettelbach, et al., "The ISO/MPEG Unified Speech and Audio Coding Standard - Consistent High Quality for All Content Types and at All Bit Rates," *J. Audio Eng. Soc.*, vol. 61, no. 12, pp. 956–977 (2013 Dec.).

[21] ISO/IEC, "MPEG-D (MPEG Audio Technologies), Part 3: Unified Speech and Audio Coding," *ISO/IEC Standard 23003-1:2012* (2012 Apr.).

[22] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H Audio—The New Standard for Coding of Immersive Spatial Audio," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 770–779 (2015 Aug.). https://doi.org/10.1109/JSTSP.2015.2411578.

[23] Mojang AB, "Minecraft," https://www.minecraft.net (accessed 8 May 2026).

[24] "Roblox, "Roblox," https://www.roblox.com/ (accessed 8 May 2026).

[25] Audiokinetic, "Wwise," https://www.audiokinetic.com/en/wwise/overview/ (accessed 8 May 2026).

[26] Firelight Technologies, "FMOD," https://www.fmod.com/ (accessed 8 May 2026).

[27] Google, "Developing with Resonance Audio," *Resonance Audio*, https://resonance-audio.github.io/resonance-audio/develop/overview (accessed 8 May 2026).

[28] Meta, "XR Audio SDK," *Audiokinetic*, https://www.audiokinetic.com/en/wwise/plugins/meta-xr-audio/ (accessed 8 May 2026).

[29] Alliance for Open Media, "Immersive Audio Model and Formats," AOM Working Group Draft (2025 Apr.). https://aomediacodec.github.io/iamf (accessed 8 May 2026).

[30] H. Fırat, L. Maffei, and M. Masullo, "3D Sound Spatialization With Game Engines: The Virtual Acoustics Performance of a Game Engine and a Middleware for Interactive Audio Design," *Virtual Real.*, vol. 26, pp. 539–558 (2022 Jun.). https://doi.org/10.1007/s10055-021-00589-0.

[31] A. Molkenthin, *Spatial Acoustics in 3D Games: A Comparative Study of Wwise Spatial Audio and Project Acoustics Workflow and Efficiency*, B.A. thesis, Darmstadt University of Applied Sciences, Darmstadt, Germany (2023 Feb.).

[32] ISO/IEC, "MPEG-I Audio Architecture and Requirements," *ISO/IEC JTC 1/SC 29/WG 6, N18158* (2021 Jan.).

[33] ISO/IEC, "MPEG-I Immersive Audio Documentation for the Audio Evaluation Platform (Max External Interface)," *ISO/IEC JTC 1/SC 29/WG 6, N0086* (2021 Jul.).

[34] ISO/IEC, "MPEG-I Immersive Audio Encoder Input Format," *ISO/IEC JTC 1/SC 29/WG 6, N0054* (2021 Apr.).

[35] ISO/IEC, "Information Technology — Coded Representation of Immersive Media, Part 4: Mpeg-I Immersive Audio," *ISO/IEC Standard 23090-4* (2025 Nov.).

[36] ISO/IEC, "Information Technology — Coded Representation of Immersive Media, Part 34: Immersive Audio Reference Software," *ISO/IEC Standard 23090-34* (2025 Nov.).

[37] ISO/IEC, "Summary of MPEG-I Immersive Audio Verification Test," *ISO/IEC JTC1/SC29/WG6, N0396* (2026 Jan.).

[38] ITU-R Recommendation BS.1534-3, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," *Intern. Telecom Union*, Geneva, Switzerland (2015 Oct.). https://www.itu.int/rec/R-REC-BS.1534-3-201510-I/en.

[39] ITU-R Recommendation BS.2132, "Method for the Subjective Quality Assessment of Audible Differences of

Sound Systems Using Multiple Stimuli Without a Given Reference," *Intern. Telecom Union*, Geneva, Switzerland (2019 Oct.). https://www.itu.int/rec/R-REC-BS.2132/en.

[40] A. Silzle, S. George, E. Habets, and T. Bachmann, "Investigation on the Quality of 3D Sound Reproduction," in Proc. *Intl. Conf. Spatial Audio*, pp. 334–341 (Detmold, Germany) (2011 Jan.).

[41] O. Rummukainen, S. Schlecht, A. Plinge, and E. A. Habets, "Evaluating Binaural Reproduction Systems From Behavioral Patterns in a Virtual Reality—A Case Study With Impaired Binaural Cues and Tracking Latency," presented at the *143rd Conv. Audio Eng. Soc.* (2017 Oct.), paper 9895. https://aes.org/publications/elibrary-page/?id=19292.

[42] J.-M. Jot and K. S. Lee, "Augmented Reality Headphone Environment Rendering," in Proc. *AES Intl. Conf. Audio for Virtual and Augmented Reality* (2016 Sep.), paper 8-2. https://aes.org/publications/elibrary-page/?id=18506.

[43] M. Schoeffler, A. Silzle, and J. Herre, "Evaluation of Spatial/3D Audio: Basic Audio Quality versus Quality of Experience," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 75–88 (2016 Feb.). https://doi.org/10.1109/JSTSP.2016.2639325.

[44] F. Brinkmann and S. Weinzierl, "Audio Quality Assessment for Virtual Reality," in M. Geronazzo and S. Serafin (Eds.), *Sonic Interactions in Virtual Environments*, Human–Computer Interaction Series, pp. 145–178 (Springer, Cham, Switzerland, 2023).

[45] A. Silzle, P. M. Delgado, A. R. Garcia, et al., "Perceptual Evaluation of the MPEG-I Immersive Audio Standard," in Proc. *160th Conv. Audio Eng. Soc.* (2026 May), paper 407.

[46] T. Falk, R. Ramakrishna, and W. de Bruijn, "Procedural Audio Using Generalized Granular Synthesis," presented at the *AES 5th Intl. Conf. Audio for Virtual and Augmented Reality* (2024 Aug.), paper 25. https://aes.org/publications/elibrary-page/?id=22674.

[47] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950 (1979 Apr.).

[48] W. M. Rabinowitz, J. Maxwell, Y. Shao, and M. Wei, "Sound Localization Cues for a Magnified Head: Implications From Sound Diffraction About a Rigid Sphere," *Presence*, vol. 2, no. 2, pp. 125–129 (1993 May).

[49] R. O. Duda, and W. L. Martens, "Range Dependence of the Response of a Spherical Head Model," *J. Acoust. Soc. Am.*, vol. 104, no. 5, pp. 3048–3058 (1998 Nov.).

[50] S. Jelfs and E. Creusen, "A Method to Improve Immersiveness in Virtual Acoustic Environment by Efficiently Simulating the Sound Experienced by a Person due to Air Moving Over the Ears," in Proc. *AES 5th Intl. Conf. Audio for Virtual and Augmented Reality* (2024 Aug.), paper 26. https://aes.org/publications/elibrary-page/?id=22675.

[51] W. de Bruijn and T. Falk, "Representation and Rendering of Heterogeneous Spatially Extended Sound Sources," in Proc. *AES 5th Intl. Conf. Audio for Virtual and Augmented Reality* (2024 Aug.), paper 24. https://aes.org/publications/elibrary-page/?id=22673.

[52] A. Silzle, S. Dick, S. Disch, and A. Adami, "Implementation and Evaluation of the Usage of Low-Complexity Early Reflection Patterns for Real-Time 6DOF Auditory Virtual Environments," in Proc. *AES 5th Intl. Conf. Audio for Virtual and Augmented Reality* (2024 Aug.), paper 19. https://aes.org/publications/elibrary-page/?id=22668.

[53] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality* (Springer, Cham, Switzerland, 2019). https://doi.org/10.1007/978-3-030-17207-7.

[54] S. Disch, "Decorrelation for Immersive Audio Applications and Sound Effects," in Proc. *26th Intl. Conf. Digital Audio Effects (DAFx)*, paper 3 (Copenhagen, Denmark) (2023 Sep.).

[55] C. Anemüller, A. Adami, and J. Herre, "Efficient Rendering of Spatially Extended Sound Sources," *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 281–292 (2022 May). https://doi.org/10.17743/jaes.2022.0069.

[56] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466 (1997 Jun.).

[57] A. Politis, L. Pajunen, J. Leppänen, and A. Eronen, "Wide-Area 6DOF Rendering of Multi-Point Ambisonic Recordings Based on Interpolation of Spatial Parameters," in Proc. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5 (New Paltz, NY) (2023 Oct.). https://doi.org/10.1109/WASPAA58266.2023.10248142.

[58] J. Leppänen, A. Politis, M.-V. Laitinen, L. Pajunen, and A. Eronen, "6DoF Rendering of Multi-Point Ambisonics Recordings at Listening Positions Exterior to the Area Defined by the Microphone Positions," in Proc. *AES 5th Intl. Conf. Audio for Virtual and Augmented Reality* (2024 Aug.), paper 13. https://aes.org/publications/elibrary-page/?id=22662.

[59] J. Herre and C. Faller, "Spatial Sound Stability Enhancement by Advanced User-Tracked Loudspeaker Rendering," in Proc. *155th Conv. Audio Eng. Soc.* (2023 Oct.), paper 178. https://aes.org/publications/elibrary-page/?id=22332.

[60] J.-H. Hanschke, D. Arteaga, G. Cengarle, et al., "Improved Panning on Non-Equidistant Loudspeakers With Direct Sound Level Compensation," in Proc. *155th Conv. Audio Eng. Soc.* (2023 Oct.), paper 10669. https://aes.org/publications/elibrary-page/?id=22250.

[61] J. Herre, C. Faller, S. Disch, and M. Schmidt, "Room Measurement Based Calibration of MPEG-I Tracked Loudspeaker Rendering," in Proc. *160th Conv. Audio Eng. Soc.* (2026 May), paper 10276.

[62] W. Chien and F. R. Lin, "Prevalence of Hearing Aid Use Among Older Adults in the United States," *Arch. Intern. Med.*, vol. 172, no. 3, pp. 292–293 (2012 Feb.). https://doi.org/10.1001/archinternmed.2011.1408.

[63] A. Silzle, S. Rose, F. Dennerlein, et al., "Perceptual Evaluation of Own Voice and Social Virtual Reality Using the MPEG-I Immersive Audio Standard," in

Proc. *159th Conv. Audio Eng. Soc.* (2025 Oct.), paper 376. https://aes.org/publications/elibrary-page/?id=23050.

[64] 3rd Generation Partnership Project, "Codec for Enhanced Voice Services (EVS); General Overview," 3GPP TS 26.441 (2016 Jan.).

[65] 3rd Generation Partnership Project, "Codec for Immersive Voice and Audio Services (IVAS); General Overview," 3GPP TS 26.250 (2025 Oct.).

[66] T. Hopkins, S. C.-C. Weng, R. Vanukuru, et al., "Studying the Effects of Network Latency on Audio-Visual Perception During an AR Musical Task," in Proc. *IEEE Intl. Symp. Mixed and Augmented Reality (ISMAR)*, pp. 26–34 (Singapore) (2022 Oct.). https://doi.org/10.1109/ISMAR55827.2022.00016.

[67] D. Brungart, A. J. Kordik, and B. D. Simpson, "Effects of Headtracker Latency in Virtual Audio Displays," *J. Audio Eng. Soc.*, vol. 54, no. 1, pp. 32–44 (2006 Jan.).

[68] Cycling '74, "MaxMSP," https://cycling74.com/products/max (accessed 8 May 2026).

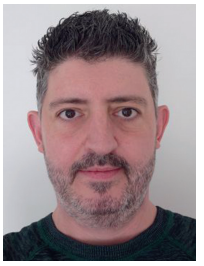
[69] Raw Material Software, "JUCE," https://www.juce.com (accessed 8 May 2026).

[70] walkaboutworlds, "Easy VR Creation: Multiple Rooms in Minutes," Sketchfab (2016 May). https://skfb.ly/O8Ay (accessed 8 May 2026).

THE AUTHORS


Sascha Disch


Leon Terentiv


Jeroen Koppens

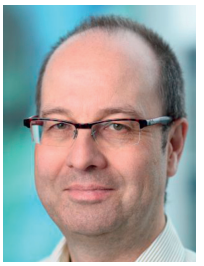

Tommy Falk


Jussi Leppänen


Isaac Munoz


Jürgen Herre


Alexander Adami


Andreas Silzle


Daniel Fischer


Panji Setiawan


Christof Fersch


Jan-Hendrik Hanschke


Sam Jelfs


Werner de Bruijn

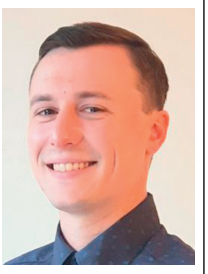

Antti Eronen


Sujeet Mate


Sascha Dick


Christian Borss


Pablo M. Delgado


Miguel M. Trinidad

Sascha Disch received his Dipl.-Ing. degree in electrical engineering from the Technical University Hamburg-Harburg in 1999 and joined the Fraunhofer IIS the same year. Ever since he has been working in research and development of perceptual audio coding and audio processing. From 2007 to 2010, he was a researcher at theLaboratory of Information Technology, Leibniz University Hannover, receiving his doctoral degree (Dr.-Ing.) in 2011. He contributed to the standardization of MPEG Surround, MPEG USAC, MPEG-H 3D Audio, and the 3rd Generation Partnership Project Enhanced Voice Services codec. Dr. Disch is also part of the development team of the MPEG-I Immersive Audio standard and one of its coleaders since 2022. His research interests as a Chief Scientist at Fraunhofer IIS include waveform and parametric audio coding, audio bandwidth extension, and digital audio effects.

•

Leon Terentiv is a Staff Engineer at Dolby. He received his diploma in applied mathematics from Far Eastern State University, Vladivostok, Russia, in 2000, and both a master's degree in computational engineering and a Dr.-Ing. degree in the field of fluid dynamics modeling from the Friedrich Alexander University of Erlangen-Nuremberg, Germany, in 2005. In the same year, he joined the Fraunhofer IIS in Erlangen, where he began his work in the field of psychoacoustics and digital audio signal processing. His research focused on spatial audio coding and rendering, which served as the foundation for his long-term involvement in international standardization efforts. In 2017, he joined Dolby, where he continues to focus on development and optimization of audio codecs. He is a contributor to the MPEG standardization, specifically within the MPEG-D, MPEG-H, and MPEG-I frameworks, working to advance the next generation of immersive audio technologies.

•

Jeroen Koppens finished his master's degree electrical engineering at TU Eindhoven with a graduation project on a new psychoacoustic model for an AAC encoder at Philips Applied Technologies in 2005. After joining Philips, he continued contributing to several audio codecs being developed within the MPEG consortium: MPEG Surround, Spatial Audio Object Coding, USAC, and the start of MPEG-H. He also researched binaural and reverberation rendering until switching jobs to Dolby Sweden in 2013. There, he contributed to guided and machine learning–based unguided dialogue enhancement, loudness management, and dynamic range control for AC-4. When relocating back to the Netherlands in 2018, he returned to Philips as project lead and standardization delegate at MPEG, contributing to the development of MPEG-I Immersive Audio.

•

Tommy Falk received his master's degree in electrical engineering from Kungliga Tekniska Högskolan, Stockholm, in 1999. He joined the Audio Technology group in Ericsson Research directly after his master's thesis and has since gained over two decades of experience working as a researcher in the area of audio processing and spatial audio. In the past, he has worked on audio processing for mobile devices, spatial audio for teleoperated machinery, audio solutions for video conferencing, and spatial audio for in-game communications. More recently, as a Master Researcher, he has contributed concepts and technology to the 3rd Generation Partnership Project Immersive Voice and Audio Services and MPEG-I Immersive Audio standardizations.

•

Jussi Leppänen studied signal processing at Tampere University of Technology and received his M.Sc. degree in 2001. He joined Nokia in 2004 and has since worked on various topics, such as speech and context recognition and spatial audio rendering.

•

Isaac Muñoz, Ph.D., is a Senior Staff Engineer in San Diego with expertise in advanced audio systems and signal processing, focusing on spatial audio rendering and high dynamic range recording technologies. His work integrates hardware and software design to deliver cutting-edge audio solutions for next-generation devices. In addition to his industry role, Isaac is a lecturer at San Diego State University, where he teaches courses in recording, live sound, and audio programming.

•

Jürgen Herre has been involved in the development of perceptual coding algorithms for high-quality audio at the Fraunhofer IIS since 1989, including the well-known ISO/MPEG-Audio Layer III coder (aka "MP3"). In 1995, he joined Bell Laboratories for a postdoctoral term working on MPEG-2 AAC. Later, he returned to Fraunhofer IIS, developing more advanced multimedia technology including MPEG-4, MPEG-7, MPEG-D, MPEG-H, and MPEG-I, currently as Chief Executive Scientist for the Fraunhofer IIS Audio and Media Technologies division. In 2010, Dr. Herre was appointed professor at the University of Erlangen and the International Audio Laboratories Erlangen. Dr. Herre is a fellow of the AES and chair of the AES Technical Committee on Coding of Audio Signals. He has served as Chair/Vice Chair of the AES Technical Council for 25 years. He was a long-time active member of the MPEG Audio standardization group.

•

Alexander Adami studied electrical engineering at the Friedrich-Alexander University Erlangen-Nürnberg and received his Dipl.-Ing. degree in 2011. He then joined the International Audio Laboratories Erlangen, where he received his Dr.-Ing. degree in 2018. Since August 2017, he has been working in the Audio and Multimedia division at Fraunhofer IIS, Erlangen, where he contributed as one of the coleads to the development of the MPEG-I Immersive Audio standardization. Until 2025, he headed the "Audio for AR & VR" group at Fraunhofer IIS and was an active member of the MPEG Audio group. His research interests are in the fields of audio for VR applications, perception of transient sounds in general and applause in particular, and upmixing and downmixing techniques for audio signals.

•

Andreas Silzle studied electrical engineering with specialization in acoustics at the University of Karlsruhe (KIT), Germany. From 1991 onwards, he worked at the Institut für Rundfunktechnik (IRT) on multichannel audio coding. In 1998, he joined TC Electronic, Denmark. From 2002 onwards, he did his Ph.D. in quality evaluation of auditory virtual environments at the Institut für Kommunikationsakustik (IKA). In 2007, he joined Fraunhofer IIS, where he is now the Chief Scientist for Room Acoustics and Quality Assessment. He set up the soundlab laboratories and the soundlab group. He was actively involved in standardization work in ISO/IEC MPEG-H and ITU-R. In 2017 and 2018, he was the technical coordinator of the EU project ORPHEUS about object-based audio for broadcasting. In 2019, he coordinated a project with the Gewandhausorchester Leipzig. Since 2021, he has participated in the MPEG-I immersive audio standardization and taken the co-lead since 2022.

•

Daniel Fischer received a Dipl.-Ing. (FH) degree in electrical engineering from Berliner Hochschule fuer Technik, Germany. He joined Dolby Germany in 2017 and currently serves as Staff Engineer with a focus on MPEG Audio technologies.

•

Panji Setiawan received his M.Sc. degree in communications engineering from the Technical University of Munich, Germany, in 2002 and his Dr.-Ing. degree in electrical engineering and information technology from Munich University of the Federal Armed Forces (Bundeswehr), Germany, in 2009. Since the start of his professional career at Siemens, followed by Fraunhofer IIS and Huawei before joining Dolby Laboratories in 2020, he has been contributing to the development of various speech and audio codec standards in ITU-T, 3rd Generation Partnership Project (3GPP), and MPEG Audio. Recently he has been working on 3GPP Immersive Voice and Audio Services, MPEG-I Immersive Audio, and ITU-T biomedical and general waveform signal coding.

•

Christof Fersch is a Director at Dolby Laboratories with extensive experience leading complex technology projects. Holding a master's degree in software engineering and information technology and a Dipl.-Ing. in electrical engineering, he drives strategic planning and the implementation of innovative solutions across international teams. In addition to his corporate responsibilities, he holds leadership roles in ISO/IEC JTC1 SC29 WG6 (MPEG Audio) and serves in various other leadership positions within ISO, ITU-T, and Bluetooth standardization.

•

Jan-Hendrik Hanschke holds a master's degree in audio communication and technology from the Technical University of Berlin, Germany. He joined Dolby Laboratories in 2018, initially to complete his master's thesis on spatial audio capture. He currently serves as a Staff Research Engineer, contributing to multiple research and development initiatives in spatial and immersive audio processing.

•

Sam Jelfs joined Koninklijke Philips N.V. in 2011 having completed a Ph.D. at Cardiff University on the topic of modeling the Cocktail Party Effect and the role of room acoustics on binaural unmasking. As well as contributing to the development of audio standards, he also contributes to standardization efforts for the compression of biomedical signals. Outside of standardization, he has previously worked on a diverse range of topics, from efforts to reduce noise disturbance in intensive care units to novel signal processing techniques for the measurement of stress and blood pressure though on-body sensing.

•

Werner de Bruijn received M.Sc. and Ph.D. degrees in applied physics from Delft University of Technology in the Netherlands in 1998 and 2004, respectively. In 2003, he joined Philips Research in Eindhoven, the Netherlands, where he conducted and led research projects on audio reproduction technologies for a wide range of consumer applications such as soundbars, televisions, and home theater systems. In 2013, he became involved in MPEG standardization, first in MPEG-H 3D Audio and later in MPEG-I Immersive Audio. In 2018, he joined Ericsson Research in Stockholm, Sweden, where he has been working on audio technologies for VR and AR, primarily in the context of MPEG-I Immersive Audio standardization. In 2024, he relocated back to the Netherlands, where he continues to work with the Ericsson audio team in Stockholm.

•

Antti Eronen is a Bell Labs Fellow and Principal Researcher at Nokia. He holds an M.Sc. and D.Sc. (Tech.) in signal processing from Tampere University of Technology and has over 20 years of experience in immersive audio innovation. His work spans spatial audio rendering, sound recognition, and immersive audio standards. Widely published and a named inventor on numerous patents, Antti is passionate about turning advanced audio research into real-world impact.

•

Sujeet Mate received his M.S. in electrical engineering from The University of Texas at Dallas and Dr. Tech. from Tampere University of Technology, respectively. He works with Nokia as a Principal Researcher, Bell Labs Distinguished Member of Technical Staff. He has worked in the areas of end-to-end multimedia systems and immersive audiovisual multimedia. He has contributed to multiple standards (in MPEG and 3rd Generation Partnership Project SA4), including MPEG-I Immersive Audio, Immersive Voice and Audio Services real-time transport protocol payload format, multimedia telephony for integrated media systems, Omnidirectional Media Format second edition, and network-based media processing. His current research interests include immersive volumetric audio, immersive audiovisual systems in AR/VR (both streaming and conversational), and cloud-based processing of media.

•

Sascha Dick received his Dipl.-Ing. degree in information and communication technologies from the Friedrich Alexander University (FAU) of Erlangen-Nuremberg, Germany, in 2011, with a thesis on an improved psychoacoustic model for spatial audio coding, and joined the Fraunhofer IIS in Erlangen, Germany, as a Research Engineer in the same year. He has contributed to the development and standardization of audio codecs such as MPEG-H 3D Audio. In 2017, he joined the International Audio Laboratories at FAU to research psychoacoustic effects for 3D audio, and he received a Ph.D. degree for his thesis *Psychoacoustic Effects and Models for Processing and Coding of 3-Dimensional Audio* in 2023. He has since joined Fraunhofer IIS again to further pursue research and development of 3D audio applications. His research interests include psychoacoustics, multichannel signal processing, and audio coding.

•

Christian Borß received a Dipl.-Ing. degree in electrical engineering from Technische Universität Braunschweig, Germany, 2004. After his graduation in the field of communications technology, he joined the Institute of Communication Acoustics at the Ruhr-Universität Bochum, Germany, where he received a Dr.-Ing. degree in electrical engineering. Since 2011, he has been working as a researcher at Fraunhofer IIS, International Audio Laboratories Erlangen. He contributed to the development and standardization of the MPEG-H 3D Audio and MPEG-I Immersive Audio standard.

•

Pablo Delgado studied electrical engineering at the University of Buenos Aires and the École nationale supérieure

d'ingénieurs de Caen. He joined Fraunhofer IIS in 2014 to work in research and development of audio coding algorithms within numerous industry projects. From 2018 to 2022, he was a researcher at the International Audio Laboratories Erlangen, Germany, where he received his doctoral degree (Dr.-Ing.) in 2023. Dr. Delgado currently works as a Senior Scientist at the Advanced Audio Research Group in Fraunhofer IIS, where he focuses on the intersection of psychoacoustics, machine learning, spatial audio, and VR.

•

Miguel Mendez Trinidad studied image and sound engineering at the Polytechnic University of Madrid, where he graduated in 2021. After that, he joined Fraunhofer IIS to support the MPEG-I standard development effort, to which he contributed in testing and developing the reference software.