Audio Engineering Society

Convention Paper 10234

Presented at the AES 159th Convention
2025 October 23–25, Long Beach, CA, USA

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

MPEG-I Immersive Audio – The Technology Of The New Standard For Virtual / Augmented Reality Audio

Sascha Disch[1], Leon Terentiv[2], Jeroen Koppens[3], Tommy Falk[4], Jussi Leppänen[5], and Isaac Munoz[6]

[1]*Fraunhofer IIS*
[2]*Dolby Germany GmbH*
[3]*Koninklijke Philips N.V.*
[4]*Ericsson AB*
[5]*Nokia Technologies*
[6]*Qualcomm Technologies, Inc*

Correspondence should be addressed to Sascha Disch (sascha.disch@iis.fraunhofer.de)

ABSTRACT

The MPEG-I immersive audio standard [1] for compressed audio representation and rendering for Virtual and Augmented Reality (VR/AR) applications with six degrees of freedom (6DoF) was technically completed in January 2025 by the MPEG Audio group (ISO/IEC JTC 1/SC 29/WG 6). This standard 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 description 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 paper presents its finalized technology.

1 Introduction

Virtual and Augmented Reality (VR/AR) has seen remarkable growth in recent years, paving the way for widespread consumer applications including educational entertainment, cultural performances and exhibitions, virtual concerts and conferences, and Social VR. For many applications, it is 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. Consequently, the ISO/MPEG standardization group initiated a new work item several years ago focused on high-quality compressed representation and efficient rendering of VR/AR media [2]. The audio component of this MPEG-I standard builds on the latest predecessor technology (i.e. the ISO/MPEG-H 3D Audio standard [3]) 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 three-dimensional space (x, y, z) and head orientation (yaw, pitch, roll). A previous paper provided an early overview of the ISO/MPEG-I immersive audio concept [4]. A considerable amount of technology has been added to the standard since then. Now that the stan-

dardization process is close to finalization, this paper presents a complete overview of the supported functionality.

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 [5]. 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 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 that needs preparation steps to enable the generation of suitable VR/AR test material, processing it by the technology under development, presention of it to subjects and evaluation of the technology's performance while experiencing full 6DoF interaction. A PC-based real-time Audio Evaluation Platform (AEP) [6] as well as a non-normative Encoder Input Format (EIF) [7] 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 [4].

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

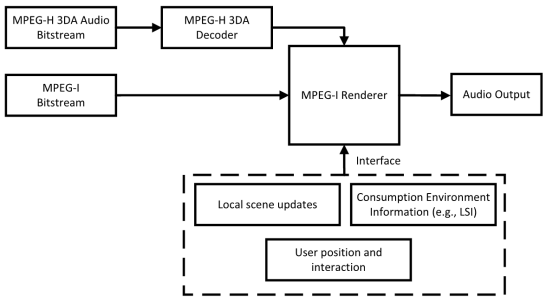


Fig. 1: Top-level MPEG-I immersive audio data flow.

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 incorporated into the Reference Model if they passed the evaluation successfully [4]. The technical work on the specification was finalized by issuing the Final Draft International Standard (FDIS) [1] in early 2025. The release of the International Standard (IS) is scheduled for the end of year 2025.

3 Renderer Overview

3.1 Bitstream Concept

MPEG-I immersive audio builds on the MPEG-H Audio Transport (MHAS) format that was introduced with MPEG-H 3D Audio [3], enabling it to additionally transmit the MPEG-I metadata. This is achieved through three newly defined MHAS packets: the Scene Config packet; the Scene Update packet; and the Scene Payload packet. This three-packet MHAS structure facilitates VR/AR applications across various scenarios, including client-server models and broadcasting. The allocation of necessary payloads across multiple Scene Payload packets can be adjusted to fit specific use cases, allowing for the interleaving of different MHAS packet types within a single stream. Existing MHAS mechanisms can be utilized for stream linking and synchronization.

3.2 Renderer Technology

Fig. 1 illustrates the high-level architecture of the normative renderer. The waveforms associated with a VR/AR scene are extracted from an MPEG-H 3D Audio bitstream and converted into PCM input for the

MPEG-I renderer. Another input to the renderer is the MPEG-I Audio metadata bitstream, carrying a compressed description of the VR/AR scene. It 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 is further detailed in Fig. 2. 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 the system. The renderer auralizes the Scene Controller's updated Scene State.

The actual rendering happens by serial 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, early reflections by geometries carrying acoustic material properties and room reverberation. Other stages aim at reducing rendering complexity or 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 early reflections or substitute sources for diffraction. Most modifications (like gain or 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 audio block rate (using non-overlapping blocks of typically 256 time domain PCM samples at sample rates f_s=32 kHz, 44.1 kHz or 48 kHz). 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.

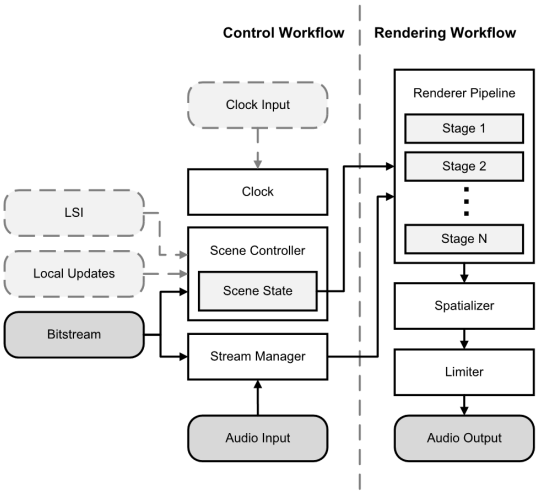


Fig. 2: MPEG-I audio renderer overview.

4 Renderer Pipeline Technology

The pipeline auralizes all relevant acoustic effects. Its stages are shown in Fig. 3. 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 unique roles that do not necessarily 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.

4.1 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 transmitted, the stage disables the sound source again when the rendering of the sound effect has ended.

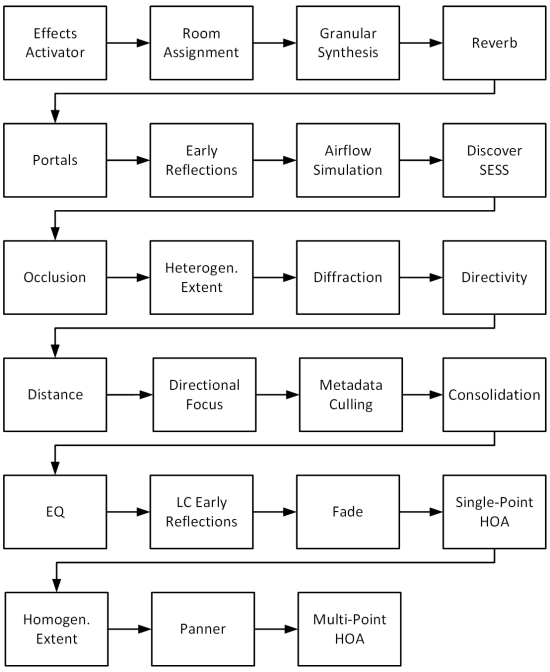


Fig. 3: MPEG-I audio renderer pipeline.

4.2 Room Assignment

The room assignment stage updates metadata of RIs to reflect the current scene configuration regarding its Acoustic Environments (AEs). This is necessary for a relevance check by the subsequent reverberation and portal stages, as well as others. A default AE can be applied in scene parts without an explicitly defined AE.

4.3 Granular Synthesis

This stage generates procedural audio using a form of granular synthesis where several aspects of the sound can be controlled dynamically in real-time [8]. 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. Rendering is based on pre-computed granular databases of sounds. To create a granular database, original recordings are partitioned into small parts, a.k.a. grains, which are stored in a multi-dimensional 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 multi-dimensional 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.

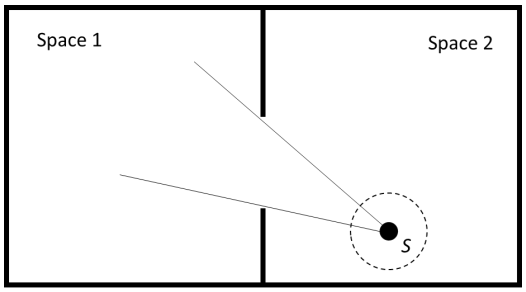


Fig. 4: A portion of the emitted source energy contributes to Space 1, and not to Space 2.

4.4 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 environments representing an AR listener environment, the FDN is configured slightly differently using the metadata and optionally a room impulse response 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 acoustic environment. When a source is inside the environment, this is the total source energy reduced by the energy lost through portals (see 4.5), while for external sources this is the portion of the source energy reaching the environment through portals, as illustrated in Fig. 4. Contributions from external sources are an important aspect for supporting multi-room reverberation in scenes with multiple connected environments. Two further concepts are present to realistically render multi-room reverberation. Firstly, the environment containing the user renders the reverberation 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 filtered by any occluding geometries such as doors or windows. 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 multi-room reverberation aspects are integrated such that they create a smooth experience as a user or a source moves through the environment. The source contribution of an external source increases 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, the reverberation extent sources are placed at a distance from the portal. As a user moves over this distance, from the extent to the portal, the fully immersive playback of the reverberation is exchanged with the directional playback from the reverberation extent source, also resulting in a smooth transition. For large portals, this distance is larger to obtain a realistic, more gradual effect.

4.5 Portals

Portals designate spatial regions where sound can propagate between two environments. This is used for two purposes: multi-room reverberation rendering, as described in 4.4, and coupled source rendering. For reverberation, the portal metadata is used to determine contributions from external sources and the reverberation from connected AEs, and the playback of the reverberation extent sources are managed. When source energy reaches a portal with a material with a non-zero coupling coefficient, the portal geometry becomes a (homogeneous) extent source – a coupling source – that re-emits the energy of the incident source, see Fig. 5. These coupling sources simulate the effect that even without a path between environments, sources can be audible by the sound energy inducing vibrations in an occluding material. If partial occlusion (4.9) was used, a clearly localizable sound at the source's position behind the wall would be heard, whereas a coupling source localizes the sound across the coupling surface (e.g. door, window).

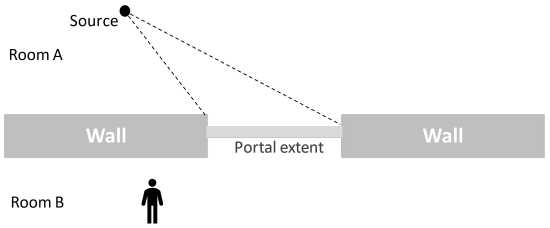


Fig. 5: Cross-section of portal with material properties, used as a coupling source.

4.6 Early Reflections (ER)

The ER stage models low-order specular reflections on planar and curved surfaces of the given scene geometry. Early reflections 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 [9]. 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. 6. 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. 6), then the round edge/corner is used as a reflection point. 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 fre-

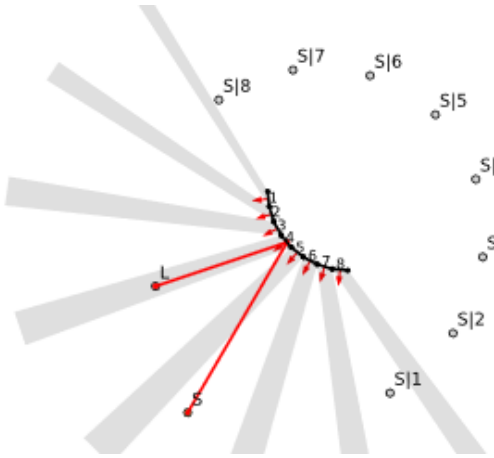


Fig. 6: Visualization of the "disco ball effect" (solid lines: reflecting surfaces; S|n: image sources; S: sound source; L: listener; grey area: area where the image sources are visible.

quency dependent reflection filter is adjusted to not only take into account the material but also the radius of the geometric primitive [10, 11].

The *voxel-based early reflection stage* offers a computationally efficient alternative to mesh-based reflections for dynamically updatable complex environments. 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 early reflections while meeting computational complexity constraints. A special ultra-low complexity mode exists that further combines calculated reflections from multiple sound sources. The reflection model incorporates both distance-dependent attenuation and material-specific reflection coefficients, with automatic indoor/outdoor scene classification to appropriately adjust reflection gains.

4.7 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 [12].

4.8 Discover SESS

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 (4.9) and the homogeneous extent stage (4.21).

4.9 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., non-, partially-, fully occluded). For extended sound sources, this process is applied for each of the samples produced by the Discover SESS stage (4.8). The occlusion results are grouped by their set of occluding materials to be processed by the homogeneous extent stage (4.21). 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.

4.10 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 non-uniform spatial source character. Specifically, it handles the rendering of object sources with an extent and a multi-channel 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). There is a seamless transition between the rendering of the two representations whenever the listener moves between the inside and outside of the extent of the source [13]. 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.

4.11 Diffraction

The diffraction stage models sound propagation around occluding objects (e.g., around corners) by deriving relevant diffraction paths from pre-processed 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. Multi-path diffraction modes are available to support complex geometries. In 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.

4.12 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 (see 4.17).

4.13 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. 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 [13]. 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.

4.14 Directional Focus

The directional focus is an important part of the accessibility features (6) 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 towards 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.

4.15 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, early reflections might be perceptually masked by their source signal. A dedicated culling of such ERs is additionally supported.

4.16 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 RI 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.

4.17 Equalizer (EQ)

Global and frequency-dependent gains that were attributed to a 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 and applied to the corresponding RI signals to auralize the accumulated gains.

4.18 Low Complexity ER (LC-ER)

The geometric room acoustics model to generate early reflections (4.6) requires a high computational complexity that increases with the number of sound sources and desired 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 in- and outdoor. In this way, only one common ER pattern is needed to process any number of sound sources in a scene. Results of subjective and objective tests confirm that LC-ERs are equally plausible as a costly dynamically updated image source model [14].

4.19 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 (4.15), consolidation (4.16) 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.

4.20 Single-Point HOA (SP-HOA)

The SP-HOA stage provides 3DoF binaural output for HOA signals. A Magnitude Least-Squares (MagLS) decoder is calculated during initialization based on the provided HRTFs [15]. During rendering, HOA sources are rotated according to the listener head orientation and decoded using the MagLS decoder to binaural output.

4.21 Homogeneous Extent

This stage efficiently auralizes Spatially Extended Sound Sources (SESS) that have a size and uniform sound characteristics directly to the binaural output. This 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 (IAPD) and Interaural Level Differences (IALD). These are calculated from the HRTF data in real time from the samples provided by the discover SESS stage and their occlusion information. To synthesize IACC, a decorrelator based on allpass filters and time-frequency variant envelope shaping is used [16]. 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 [17].

4.22 Panner

The panner stage positions the immersive late reverb outputs of the reverberation stage (4.4) into the output sound image using Vector Base Amplitude Panning (VBAP) [18] 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.

4.23 Multi-Point HOA (MP-HOA)

The MP-HOA stage provides 6DoF binaural output for a scene captured with several HOA sources by calculating spatial parameters in time-frequency domain at the HOA source positions and interpolating the spatial parameters to obtain an estimate of the spatial parameters at the listener position. A prototype binaural signal obtained from the HOA signal closest to the listener position is then adapted to match the binaural cues described by the interpolated spatial metadata. The MP-HOA stage can take the locations of sound sources as an input for improved spatial localization. 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.

5 Spatializer

5.1 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, the signal of each RI is distributed to pre-defined representative virtual loudspeaker positions based on the directions listener towards RIs and listener towards the representative positions. For panning, precomputed time shifts and gains for left and right outputs are applied to the RIs. Panned signals are summed at the representative positions and 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.

5.2 Tracked Loudspeaker Rendering

Tracking of user position data in real-time enables to mitigate the spatial audio image collapse towards the nearest speaker that happens outside the sweet spot. This is achieved by adaptive adjustments of level and delay [19] while maintaining appropriate direct sound levels [20]. Also directivity of high-frequency loudspeaker radiation can be corrected, utilizing a distance-dependent model of the loudspeaker in the listening room. This constitutes the so-called *physical compensation level*. It can be utilized with any spatial loudspeaker arrangement, ranging from 2.0 to 22.2 channels, allowing 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. Based on this, 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.

6 Accessibility

With the increase in 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 with a clear hearing impairment are actually using a hearing aid [21]. With the promise and aim of MPEG-I of delivering immersive experiences and connecting people via Social VR, the technology needs to be accessible by design to users of all ages and hearing abilities. 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 has obvious performance and quality improvements. Scene adjustments are possible via attenuation of ERs (4.6, 4.18) and reverberation (4.4), expansion or compression of distance-dependent gains (4.13), Directional Focus (4.14) and spectral compensation.

7 Summary

The ISO/MPEG MPEG-I immersive audio standard is expected to be finalized at the end of 2025. In this paper, we sketched the background and development history of the specification work starting in the year 2017, and we presented the concept, supported functionality and underlying technology in detail. MPEG-I immersive audio allows for high-quality, bitrate-efficient representation and rendering for Virtual and Augmented Reality

(VR/AR) audio in six degrees of freedom (6DoF) applications. It sets the foundation for the first long-term stable audio format specification in the VR/AR sector, applicable to a wide range of consumer uses, including metaverse technologies, social VR, broadcasting and streaming.

8 Acknowledgements

The authors would like to express their thanks for contributing material to this paper (in alphabetical order) Alexander Adami, Christian Borss, Werner de Bruijn, Pablo Delgado, Antti Eronen, Christof Fersch, Daniel Fischer, Alejandro Restrepo Garcia, Jan-Hendrik Hanschke, Jürgen Herre, Sam Jelfs, Panji Setiawan, Andreas Silzle and Miguel Mendez Trinidad. 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.

References

[1] ISO/IEC N0301 - Text of ISO/IEC FDIS 23090-4 MPEG-I immersive audio, 2025.

[2] ISO/IEC JTC 1/SC 29/WG 11, "N18970, Definition of MPEG-I Phase 2 and subphases," 2020.

[3] ISO/IEC 23008-3:2019, Information technology - High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D Audio, 2019.

[4] Herre, J. and Disch, S., "MPEG-I Immersive Audio Reference Model For The Virtual/Augmented Reality Audio Standard," *J. Audio Eng. Soc.*, 71(5), pp. 229–240, 2023.

[5] ISO/IEC JTC 1/SC 29/WG 6, "N18158, MPEG-I Audio Architecture and Requirements," 2021.

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

[7] ISO/IEC JTC 1/SC 29/WG 6, "N0054, MPEG-I Immersive Audio Encoder Input format," 2021.

[8] Falk, T., Ramakrishna, R., and de Bruijn, W., "Procedural Audio Using Generalized Granular Synthesis," in *AES 5th Int. Conf. on AVAR*, 2024.

[9] Allen, J. B. and Berkley, D. A., "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, 65(4), pp. 943–950, 1979.

[10] Rabinowitz, W. M. et al., "Sound Localization Cues for a Magnified Head: Implications from Sound Diffraction About a Rigid Sphere," *Presence*, 2(2), pp. 125–129, 1993.

[11] Duda, R. O. and Martens, W. L., "Range Dependence of the Response of a Spherical Head Model," *J. Acoust. Soc. Am.*, 104(5), pp. 3048–3058, 1998.

[12] Jelfs, S. and Creusen, E., "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." *J. Audio Eng. Soc.*, (26), 2024.

[13] de Bruijn, W. and Falk, T., "Representation and Rendering of Heterogeneous Spatially Extended Sound Sources," in *AES 5th Int. Conf. on AVAR*, 2024.

[14] Silzle, A., Dick, S., Disch, S., and Adami, A., "Implementation and Evaluation of the Usage of Low-Complexity Early Reflection Patterns for Real-time 6DOF Auditory Virtual Environments," in *AES 5th Int. Conf. on AVAR*, 2024.

[15] Zotter, F. and Frank, M., *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, Springer Nature, 2019.

[16] Disch, S., "Decorrelation for immersive audio applications and sound effects," in *Proc. DAFx-23*, 2023.

[17] Anemüller, C., Adami, A., and Herre, J., "Efficient Rendering of Spatially Extended Sound Sources," *J. AES Special Issue on Audio for Virtual and Augmented Reality*, 2022.

[18] Pulkki, V., "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, 45(6), pp. 456–466, 1997.

[19] Herre, J. and Faller, C., "Spatial Sound Stability Enhancement By Advanced User-Tracked Loudspeaker Rendering," in *155th AES Conv., Express Paper 178*, 2023.

[20] Hanschke, J.-H., Arteaga, D., Cengarle, G., Lando, J., Thomas, M. R. P., and Seefeldt, A., "Improved Panning on Non-equidistant Loudspeakers with Direct Sound Level Compensation," *J. Audio Eng. Soc.*, 2023.

[21] Chien, W. and Lin, F. R., "Prevalence of Hearing Aid Use Among Older Adults in the United States," *Archives of Internal Medicine*, 172(3), pp. 292–293, 2012.