Logo

Description automatically generated ISO/IEC JTC 1/SC 29/AG 3 N150

**ISO/IEC JTC 1/SC 29/AG 03  
MPEG Liaison and Communication   
Convenorship: KR**

**Document type:** Output Document

**Title:** White Paper on MPEG-I Immersive Audio

**Status:** Approved

**Date of document:** 2024-04-26

**Source:** ISO/IEC JTC 1/SC 29/AG 03

**Expected action:** None

**Action due date:** None

**No. of pages:** 12 (including cover page)

**Email of Convenor:** kyuheonkim@khu.ac.kr

**Committee URL:** https://isotc.iso.org/livelink/livelink/open/jtc1sc29ag3

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG6, MPEG AUDIO CODING**

**ISO/IEC JTC1/SC29/AG3 N150**

**April 2024, Rennes, FR**

|  |  |
| --- | --- |
| **Title** | **White Paper on MPEG-I Immersive Audio** |
| **Source** | **AG 3, MPEG Liaison and Communications** |
| **Status** | **Approved** |
| **Serial Number** | **23975** |

# Introduction

The newly developed MPEG-I immersive audio standard is a comprehensive specification for compact and realistic representation and rendering of audio for virtual acoustic scenes with six degrees of freedom (6DoF) for the user, such as for Virtual, Augmented, and Mixed Reality (VR/AR/MR) collectively referred to as XR. In the proven MPEG tradition, great care has been taken to represent the associated information in a compact way to enable efficient distribution even over strongly bitrate-limited channels. There are a number of unique aspects that distinguish MPEG-I immersive audio over proprietary XR audio frameworks:

1. Interoperability  
   While there may be many different technical solutions for XR audio representation and rendering, it is the virtue of an open standard that the specification is open to anyone interested and can therefore be implemented by a broad range of different manufacturers on their specific software or hardware platforms. MPEG-I immersive audio is designed to be the interoperable common format between many XR platforms, thus allowing content to be exchanged freely between them. This saves both time and resources and secures the considerable investment necessary for creating high-quality XR content.
2. Long-term stability of format   
   Similar to the way MP3 files can be seamlessly played back 30 years after their creation and still sound identical, the MPEG-I immersive audio specification makes sure that produced XR audio content can be played for a long time into the future.
3. Streaming or download of XR content   
   MPEG-I immersive audio supports both streaming-based and file-based operation to serve a broad range of conceivable application scenarios.
4. MPEG-H 3D audio compatibility   
   In order to benefit from existing MPEG-H 3D audio content, production tools, content production pipelines, and decoders, MPEG-I immersive audio was designed to be compatible to MPEG-H 3D audio in that it can ingest all three content types: channels, objects and HOA.

# MPEG-I Immersive Audio

The MPEG-I immersive audio standard aims at providing a convincing solution for *compact representation* and for *high-quality real-time interactive rendering* of virtual audio content with six degrees of freedom (“6DoF”), i.e. the user can not only turn his/her head in all directions (pitch/yaw/roll) but also move around freely in 3D space.

By exploring the 6DoF virtual world, many acoustic effects of the real world must be modelled accurately to provide a realistic user experience, including properties of sound sources (e.g. level, size, radiation/directivity characteristics, Doppler processing) as well as effects of the acoustic environment (e.g. sound reflections and reverberation, diffraction, total- and partial occlusion). MPEG-I immersive audio features a plethora of technology components that support computationally efficient rendering of such aspects. Distinguishing from many existing technologies, it offers both scene description by *physics-inspired metadata* (for easier scene authoring from CAD scenes and material databases) as well as possibilities for *artistic tuning of the scene characteristics* to achieve the desired results.

During the standardization process[[1]](#footnote-2) extensive listening test comparisons and evaluation were conducted.

MPEG-I immersive audio supports many different use cases, including the following ones:

1. Virtual Reality (VR) rendering on both PC-based and mobile, computationally constrained, devices with binaural output to 6DoF head-tracked headphones and Head-Mounted Display (HMD) with possibilities for real-time interaction with the scene content (e.g.: playing a virtual basketball game).
2. Augmented Reality (AR) rendering on similar devices, but now allowing an augmentation of the real world by (possibly interactive) virtual acoustic components.
3. Loudspeaker Rendered XR: in contrast to other XR audio rendering technologies, MPEG-I immersive audio also allows audio content to be rendered on immersive loudspeaker setups from 5.1 to 22.2 channels. For audio-only applications, the ability to track the user position (x, y, z) allows the reproduction of existing spatial audio content (2.0 … 22.2, objects, HOA) with a large adaptive sweet-spot area and allows a sound image stability that is otherwise only known from Wavefield Synthesis type systems using hundreds of loudspeakers.
4. SocialVR: The user’s own voice can be rendered in the virtual environment enabling acoustic interaction with the scene supporting e.g. virtual meetings.

Figure 1 shows an overview of the MPEG-I immersive audio architecture.

Ein Bild, das Text, Screenshot, Diagramm, Design enthält.

Automatisch generierte Beschreibung

Figure 1 - MPEG-I immersive audio architecture

The MPEG-I encoder processes PCM audio objects, channels and HOA signals alongside with raw metadata that describes relevant acoustic properties of an XR scene and the associated audio signals. For authoring and storing the raw metadata that is input to the encoder, a dedicated scene description file format has been developed called Encoder Input Format (EIF) that holds definitions for e.g. sources and their positions, scene geometry (primitives, meshes or voxels), transforms, material acoustic coefficients, etc.

The output of the encoder is a compressed metadata bitstream that is packaged in MHAS (MPEG-H Audio Streaming) compatible data packets. It is, together with MPEG-H 3D audio encoded audio objects, channels and HOA, transmitted to the decoder.

The decoded PCM audio and the MPEG-I bitstream are input to the MPEG-I Renderer. Additionally, the renderer has an interface to input local information on the consumption environment, scene updates and on instantly tracked user position, orientation and user interaction with the scene. User interaction can e.g. be opening or closing doors or manipulating sources.

# Features

## Audio source types

The MPEG-I immersive audio standard can handle several types of commonly used audio representations, these can be fitted to a scene in any desired combination. As with MPEG-H 3D audio, the new MPEG-I standard can process channel-based audio (e.g. mono, stereo, 5.1, etc.), object-based audio, and Higher Order Ambisonics (HOA) audio (also referred to as “directional soundfield”).

An HOA source can be set either as a 3DoF “head-anchored” soundfield, or as a 6DoF source occupying a specific volume area of a scene. This latter case allows both a typical “internal” rendering, with the HOA soundfield surrounding the user when they are located within its source boundaries, and an “external” rendering that treats the HOA source as an object source with an extent when the user is outside of its designated boundary volume. The MPEG-I renderer ensures a smooth transition between these two rendering modes.

In addition to these common types, MPEG-I supports “multipoint” HOA audio files (MP-HOA), which consist of groups of concurrent and spatially separated HOA captures belonging to the same scene, that can be navigated in 6DoF within a rendered environment.

## Mesh- and voxel-based audio rendering capabilities

The MPEG-I immersive audio technology supports mesh- and voxel-based rendering approaches.

The mesh-based approach offers precise control over the location and high fidelity of shape representation for complex scene elements. It also allows for an efficient way to support dynamically changing environments (including smooth animations). The mesh-based rendering approach for simulating diffraction, high-order reflections and reverberation can create realistic and acoustically accurate results and provide immersive auditory experience.

The voxel-based approach for audio metadata representation is an alternative concept for audio processing. Voxel-based geometry representation are used in various application fields, including gaming, design and engineering. Voxels for audio rendering refer to discrete volumetric cubes assigned to specific material/media properties and audio rendering instructions. Voxels form elementary building blocks of a virtual world and can be considered as a simplified geometry representation of complex elements and structures for acoustics simulation.

The voxels mode supports real-time, procedurally generated “seamless” and “endless” virtual environments (e.g., landscapes and buildings using AI-driven algorithms) where users can interact with the world in a natural way, such as digging, sculpting, creating and destroying elements. The main feature of voxel-based approach in MPEG-I immersive audio is that voxel geometry can be arbitrarily modified during the real-time audio rendering process (i.e., it does not require any pre-baking in the encoder). When the content creator allows it, all voxels for audio processing can be manipulated by users and handled by voxel interaction physics engine individually. The MPEG-I immersive audio has low computational complexity mode and voxel matrix compression tool for handing high-resolution voxel scenes.

The MPEG-I immersive audio technology can complement the video rendering by the corresponding sound rendering counterpart when both audio and video scene rendering technologies share the same underlying mesh- or voxel-based virtual world representation.

## Acoustic material properties

Acoustic materials can be assigned to mesh and voxel elements. These materials are characterized by parameters describing transmitted, reflected, coupled and absorbed audio energy parts. Materials are assumed to be isotropic but sound frequency dependent. The acoustic material coefficients can be obtained from material databases or created by virtual scene designers. The specification of material properties controls the modeling of acoustic occlusion, reflections and diffraction allowing for a nuanced simulation of how sound propagates and interacts with different elements in the virtual environment. The set of tunning parameters and settings provide means to precisely control these audio rendering algorithms to support content creator intent expression.

## Extended sources

With MPEG-I immersive audio, sound sources do not only have a position and orientation in space, but they may also have a specific size and shape, which can be described by a simple shape such as a box, or a more complex shape in the form of a mesh. Such “extended sound sources” can be used to represent object sources from a moderate size, such as a grand piano, to a huge size, such as a forest of rustling leaves.

The rendering of extended sound sources is adapted to the listening position so that the perceived width and height matches the shape of the object source as seen from the listening position. It is also possible for the user to enter inside the extent of a sound source (if the content creator has chosen to allow this), in which case the sound will completely envelop the user. The distance gain fall-off is adapted to the size of the extent using an acoustic model simulating physics. Extended sound sources can be partially occluded, e.g., by a wall, and/or have soft occlusion from, e.g., a curtain.

An extended sound source can be based on a single signal (homogeneous extended sources), in which case the renderer provides processing to generate a perception of the size of the extent, or it can be based on a multi-channel signal (heterogeneous extended sources) in which case the spatial information of the signals is used to establish the width and height perception. If an HOA signal is used as source signal for an extended sound source, an exterior rendering of the HOA signal is used for listening positions outside the extent and, if the user moves into the extent, a smooth transition to an interior rendering is provided.

## MP-HOA

MPEG-I immersive Audio allows for 6DoF rendering of scenes comprising one or more HOA sources. This is achieved via “Multi-Point Higher Order Ambisonics” (MP-HOA) rendering, which processes multiple HOA sources simultaneously and creates a 6 Degrees-of-Freedom (6DoF) listening experience for the user. The aim is to create a binaural output signal for the user based on the HOA source positions and orientations in the scene and accompanying HOA audio signals. For example, imagine navigating the sound of a concert recorded with a grid of HOA microphones). To achieve a smooth transition between the allocated zones of each HOA source in the group, the renderer may interpolate between two, or more, different capture points.

Two different MP-HOA rendering modes are provided, a parametric approach where the binaural output signal is achieved via interpolation of spatial metadata calculated for the HOA sources and a low-complexity non-parametric approach where HOA sources are rendered through virtual loudspeakers.

Binaural output is provided for listening positions within the area of the HOA source positions as well as outside the group Furthermore, if information about the sound source positions in the scene is available, “informed mode” processing, can be used to increase sound source localization and improve the distance attenuation effect.

The 6DoF audio scene can also be modified to fit in the real word listening space in case of AR scenes. For example an MP-HOA group, or multiple groups, may be representing real world captures obtained in large spaces than do not fit the user's reproduction space. In this case, a rescaling factor can be used to reshape the MP-HOA spatial distribution to the local geometry. This process can be automated such that the rescale factors are automatically determined in the renderer by comparing the scene and user space regions. When the content creator wants control over the degree to which each dimension is automatically rescaled, a tolerance can be set to limit the amount of rescaling.

## Interaction

The MPEG-I immersive audio standard enables the user to interact with the content in various ways. The user's 6DoF movement within the scene is the most basic form of interaction. This can happen through motion tracking as well as through teleporting, allowing the user to move around occluding elements or into different rooms, which will change the way that sources, reflections and reverberation are rendered.

Furthermore, a scene can define specific interaction events that can be triggered by an external entity. Often this is caused directly by a user’s action but could also be prompted more indirectly, e.g., through a physics engine. These user interactions can change properties of the scene, such as the position or orientation of an element (e.g. opening and closing a door), or the properties of a sound source (e.g. playback level of a radio). In the voxel-based scene representation mode, the user interactions with the virtual scene support full control over the scene geometry, it is possible to create and modify scene geometry in real-time (similar to a *Minecraft* gameplay).

There are three types of scene updates, depending on how they are triggered:

* **Triggered scene updates** are fully defined changes that are executed when triggered. E.g. user opening a door or turning on a TV.
* **Dynamic scene updates** define which attributes of an element are changed, but the values are to be provided by the external entity that triggers the update. E.g. user picking up a radio and moving it, or creating a hole in a solid wall.
* **Conditional scene updates** define a specific condition that triggers the scene update. E.g. the user entering a certain region of the scene.

## Listening space description

For plausible 6DoF audio rendering in AR scenarios, characteristics of the content consumption space should be taken into account. MPEG-I immersive audio uses a listening space description format (LSDF) for providing the audio renderer with information about the space the user is in. The information describes the geometry of the listening space, its surface material properties and acoustic parameters such as reverberation time. For loudspeaker output, also the physical loudspeaker setup of the consumption space is stored in the LSDF. The information in the LSDF is used for rendering immersive reverberation matching user expectations for the listening space.

Furthermore, an anchor mechanism is used for placing scene content in appropriate positions in the listening space. For scenes with virtual scene information, the physical and virtual scene information is combined into a common scene state for rendering the scene.

## Procedurally generated audio

For sound sources that are synthesized in response to user interactions or dynamic changes within the virtual scene, e.g., the engine sound of a car that can be controlled by a user, or nature sounds that respond to weather parameters in the scene, procedurally generated audio is needed. MPEG-I immersive audio implements two tools for procedural audio generation.

### Granular Synthesis

A vast array of different procedural sounds can be generated using a multi-dimensional annotated database of *audio grains*. Audio grains are short excerpts of audio signals, each associated with a value of one or more control parameters. The different dimensions (up to three) of the database can represent different aspects of a sound, such as the RPM and the torque of a car engine. By controlling a target 'position' within the coordinate system of the database, the generated sound can be controlled in real-time following changes in the scene or user interaction.

The granular databases are typically based on real recordings of a sound source that are divided up into small portions a.k.a. grains. Each grain is given a position in the coordinate system of the database that corresponds to aspects that the dimensions represent, e.g., RPM of an engine, the flow of water from a water tap, or the intensity of rain in a recording from a forest. Using a controlled random selection of grains, a natural variation of the generated sound can be achieved without repeating the same grain or sequence of grains.

### Airflow Noise

The noise that airflow makes as it passes around a user's virtual head can be simulated. The feature can make scenes more realistic and immersive by having sound that matches visuals of wind, or when being around a fan or other airflow source.

Different types of airflow generators can be defined, determining where the airflow is present, with what strength, and from which direction it propagates. The noise characteristics are changed, depending on the velocity of the airflow and the user's orientation with respect to the airflow direction.

## Accessibility

To improve the experience of MPEG-I immersive audio for users that are hard of hearing, the audio rendering process can be adjusted according to the user-specific sensory conditions.

These adjustments can be provided via a user interface according to the user’s individual hearing abilities or sensory preferences. These user preference data are then converted to rendering parameters. For example, to increase intelligibility, a user could request to render the current audio scene with less late reverberation and attenuate the audio sources that are outside the viewing direction of the user.

In contrast to a dedicated postprocessing step, improving audio accessibility within the rendering process ensures low motion-to-sound latency and low computational needs for the accessibility processing.

# Technical description

## Renderer overview

A block diagram with all renderer components is shown in Figure 2.

Ein Bild, das Text, Screenshot, Diagramm, Zahl enthält.

Automatisch generierte Beschreibung

Figure 2 - High-level block diagram of renderer.

The MPEG-I immersive audio renderer receives a metadata bitstream that describes the acoustic scene to be rendered, including identifiers to associated PCM audio streams, a dedicated representation of acoustically relevant geometry, material information and audio effect control data. The PCM audio data is handled by the Stream Manager.

The central part of the renderer is the Scene module. The Scene State always reflects the current state of all 6DoF metadata in the scene, including audio sources and geometry. It receives the input of a clock for progression of animations, gets modified by interaction events and controls the rendering process. Other components of the renderer can subscribe to changes in the scene state. Before rendering starts, all elements in the scene are created and their metadata is updated to the state that reflects the desired scene configuration when playback starts. Scene elements include audio sources, anchors/transforms and geometry.

Scene Control data and parameters are handled in an update thread in update steps at rather low rate (e.g., 50 Hz) while audio data is processed in an audio thread at 48 kHz sample rate.

Further renderer inputs are local updates: An API that allows changing certain scene properties. This can for example be used to update source or geometry positions. Information on the reproduction facility used for rendering can be input via LSDF.

At any given point in time, the renderer auralizes the current scene state. Auralization is organized based on so-called render items. A render item (RI) is the unified way of describing an audio element in the renderer pipeline.

The processing of the various perceptual effects on the RIs is done in renderer stages of the renderer pipeline. The spatializer is located after the renderer pipeline and is responsible for creating audio signals from the remaining RIs after the processing by the pipeline. The spatializer creates either a binaural stereo output or tracked rendering via loudspeaker.

## Renderer pipeline

The rendering is performed by several stages with dedicated responsibilities. A block diagram of the reference model software rendering stages is provided in Figure 3, but parallelization of some stages is possible.

A diagram of a system

Description automatically generated

Figure 3 - Default renderer stages order.

The **Effects Activator stage** activates RIs related to audio effects just for the duration of the effect.

The **Room Assignment stage** updates acoustic environment identifiers for the user and each RI at each update step to indicate in which acoustic environment they are located.

The **Granular Synthesis stage** creates interactive sounds from sound grain databases.

The **Reverb stage** renders the reverberation for each acoustic environment using feedback delay network (FDN) reverberators. Reverberators for acoustic environments having portals provide their output to the **Portal stage**. Portals are constructs that model the propagation of late reverberation and coupled transmission between acoustic environments.

For **Early Reflections** (ER), two alternatives are supported. The ER stage utilizes transmitted geometry data to compute reflections from reflecting scene geometry represented by mesh surfaces or voxels. As a low complexity alternative, different predefined ER patterns for indoor and outdoor environments are calculated in the **LC-ER stage**.

In the **Airflow Simulation stage** render items are introduced for direct playback on headphones that simulate the sound of airflow passing the user's head when airflow generators have been defined for a scene.

The **Discover SESS stage** is a helper stage for rendering Spatially Extended Sound Sources (SESS). Spatially and perceptually close RIs are combined in the **Consolidation stage** for computationally efficient rendering**.**

The **Occlusion stage** provides occlusion information with respect to a direct path from source to user.

The **Diffraction stage** provides the information required for generating diffracted sounds from hidden sources to a user around occluding elements.

The **Heterogeneous Extent stage** supports extended sources with up to nine channels where the channels represent signals for certain spatial segments of the extent as specified by the input layout parameter. The sources are rendered using an adaptive setup of virtual loudspeakers.

In the **Directivity stage**, directivity gains are calculated for RIs with an associated directivity pattern.

In the **Distance stage**, a propagation delay is applied to the RI signals to produce a physically accurate delay and Doppler effect using a variable delay line with subsample interpolation. Furthermore, distance attenuation gains are calculated from geometrical spreading. A model is used that takes the extent of the sound source into account, while air absorption filtering is modeled through a low-pass effect for audio sources that are far away from the user.

To improve the acoustic clarity for hearing impaired users, the **DirectionalFocus stage** acts like a configurable beamformer and attenuates audio sources that impinge from directions other than the (viewing) direction.

RIs that are inaudible due to low absolute or relative loudness, very low gain or EQ are deactivated in the **Metadata Culling stage** to save unnecessary computations in subsequent stages.

The **Equalizer (EQ) stage** applies frequency-dependent gain to RI signals after the frequency-dependent attenuation has been accumulated in other stages (e.g. occlusion, diffraction, reflection, directivity, medium attenuation) as well as an accumulated broadband gain. The **Fade stage** is responsible for applying fade-in and fade-out ramps to audio signals before RIs are activated or deactivated. It is also active during teleporting in order to briefly mute sound output.

The **Single Point Higher Order Ambisonics (SPHOA) stage** renders a single HOA source directly to binaural signals based on the user position and orientation with respect to the source position.

The **Homogeneous Extent stage** renders Spatially Extended Sound Sources (SESS) with uniform sound characteristic in the associated spatial range. It is designed for maximum efficiency by directly creating a binaurally rendered version of the SESS.

The **Panner stage** implements Vector-Base-Amplitude-Panning (VBAP) with additional controls such as configurable spatial spread to render some elements of the reverberation output.

The **Multi-Point Higher Order Ambisonics (MPHOA) stage** renders multiple HOA sources simultaneously.

## Metadata representation

The MPEG-I bitstream contains an efficiently compressed description of the scene and complementary data for supporting specific features achieving low complexity operation.

It carries, among others, metadata describing source locations and properties, reverberation properties of each acoustic environment and scene geometry. Also, portals are included for acoustic energy exchange between acoustic environments, and metadata indicating whether or not an audio source or geometry remains unchanged throughout the duration of the scene.

Complementary data is included to facilitate accelerated rendering of certain features, based on encoder-side analysis of the scene description. For example, diffraction edges on all the geometry and paths between sources and the user for diffraction rendering, and lists of relevant reflection surfaces and the possible sequences of reflection between them for early reflection rendering. The bitstream can also contain control parameters for alternative modes such as low complexity early reflections rendering that can be created in the renderer without complex geometry-based calculations.

Furthermore, the metadata includes definitions of what scene aspects can be modified by user interaction.

The MPEG-I immersive audio bitstream transport relies on MHAS (MPEG-H Audio Stream), by introducing three new packet types to carry the 6DoF metadata alongside the MPEG-H 3D audio signal assets.

# Conclusions

The MPEG-I immersive audio standard is a comprehensive specification that provides for compact and realistic representation and rendering of audio for Virtual, Augmented or Mixed Reality scenes. It enables high-quality real-time interactive rendering of the virtual audio content with six degrees of freedom (6DoF) for the user, e.g., the user can not only turn the head in all directions (pitch/yaw/roll) but also move around freely in 3D space (x/y/z). This permits a very high sense of user immersion in the scene.

Great care has been taken to represent the information required by the MPEG-I immersive audio renderer in a compact way to enable efficient distribution even over severely bitrate-limited networks. MPEG-I immersive audio provides for interoperability, is a long-term, stable format, is applicable for both streaming and download of content, and natively supports MPEG-H 3D audio for compression of audio content.

In the MPEG-I immersive audio standard, many acoustic effects of the real world are accurately modeled to provide a realistic user experience, including properties of sound sources (e.g. level, size, directivity characteristics and Doppler processing) as well as effects of the acoustic environment (e.g. sound reflections and reverberation, diffraction, total- and partial occlusion). The standard provides computationally efficient rendering of these aspects. Distinguishing it from many existing technologies, it offers both scene description by physics-inspired metadata (for easier scene authoring from CAD scenes and material databases) as well as possibilities for artistic tuning of the scene characteristics to achieve desired results.

The MPEG-I immersive audio specification will be published as International Standard “ISO/IEC 23090-4, Immersive audio” in early in 2025.

1. Herre, J. and S. Disch, MPEG-I Immersive Audio – Reference Model For The New Virtual / Augmented Reality Audio Standard. J. Audio Eng. Soc., 2023. 71(5) DOI: https://doi.org/10.17743/jaes.2022.0074. [↑](#footnote-ref-2)