**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 N18127**

**January 2019, Marrakech, MA**

|  |  |
| --- | --- |
| **Source** | **Requirements** |
| **Status** | **Approved** |
| **Title** | **Requirements for MPEG-I Phase 2** |
| **Editor** | **Eric Yip, Rob Koenen** |

# Introduction

This document contains the draft requirements for MPEG-I Phase 2.

“The specification” shall mean any specification which may be targeted for MPEG-I Phase 2, as well as any supporting standards that may require updating, such as those specifying certain types of metadata.

# Definitions

This Section contains definitions for terms that are not defined in specifications related to MPEG-I Phase 1 and 1b.

|  |  |
| --- | --- |
| **Media Types** | Types of media (visual and audio) including at least: 2D video, spherical video, point clouds and various audio media data. |
| **6DoF scene** | A scene in which a user is able to move around freely with 6 degrees of freedom. |
| **Viewpoint** | The point from which the user observes (views and hears) the scene; it usually corresponds to a camera position. Slight head motion does not imply a different Viewpoint |
| **Audio Object** | Audio Signal and associated metadata as used in ISO/IEC 23008-3. |
| **Audio Channel** | Audio Signal and associated metadata as used in ISO/IEC 23008-3. |
| **HOA** | Audio Signal and associated metadata as used in ISO/IEC 23008-3. |
| **Earcon** | Audio Signal and associated metadata as used in ISO/IEC 23008-3:2019/AMD 1. |
| **Audio Element** | One or more audio signals and associated metadata. Audio Elements are audio objects, channels or HOA signals with associated MPEG-I 6DoF metadata and MPEG-H 3D Audio metadata if appropriate. |
| **Acoustic Environment** | Metadata describing the acoustic properties of the virtual scene to be rendered, e.g. room or locality. These might include reverberation times (RT60), zones of occlusion, etc. |
| **Acoustic Element** | Object in VR space that reflects, diffracts and/or absorbs sound. |
| **Audio Scene** | All audio elements, acoustic elements and acoustic environment which are needed to render the sound in the scene.  Note: Audio scene is not the same as scene-based audio (e.g., HOA). |
| **User** | The listener whose position and orientation are used for rendering. |
| **Audio Scene change** | Changes in the audio scene which impact the audio rendering. These changes might be triggered by the user. |
| **Audio Renderer** | Normative functionality that renders sound for the user. The output of the Audio Renderer are the audio signals to be reproduced over headphones or loudspeakers. |
| **Audio Rendering Parameters** | Parameters that control the audio renderer, such as enabling/disabling specific features (e.g., room acoustics, Doppler shifts) or min/max values, gain values, default values, etc. |
| **MPEG-H 3DA Decoder** | MPEG-H 3D Audio Low Complexity (LC) Profile decoder that receives as input an MPEG-H 3D Audio LC Profile MHAS stream and provides as output decoded PCM audio together with all metadata available in the MHAS packets. Decoded PCM audio contains channels, objects and reconstructed HOA as described in ISO/IEC 23008-3:2018 Section 17.10.  *Note: Not all metadata provided has to be used for MPEG-I 6DoF rendering.* |

# Requirements for MPEG-I Phase 2

## Content Representation

1. The specification shall support the efficient representation and compression of the defined Media Types.

## Audio with 6 Degrees of Freedom

Note: background information for the Audio requirements can be found in [1].

### General Requirements

1. The specification shall support user 6DoF (Degrees of Freedom) so that a single user perceives an experience consistent with user's movement in the environment (e.g., low and non-perceivable motion-to-sound latency).
2. The specification shall support efficient representation and compression of media and metadata.
   1. Media coding shall be done according to the MPEG-H 3D Audio Low Complexity Profile (including 3D Audio metadata). The specification shall support any combination of channel-based, object-based and HOA content types.

Note: Multiple HOA streams (e.g., sampled at different locations) may be supported simultaneously in one Audio Scene.

* 1. The specification shall support Loudness and Dynamic Range Control using MPEG-H 3D Audio tools.
  2. The specification shall support additional metadata as needed to support user 6DoF.
  3. The specification shall support delivery of the audio scene in multiple audio streams.

Note: This may already be fulfilled by MPEG-H 3D Audio.

1. The specification shall support rendering of the audio scene, consistent with the acoustic elements and acoustic environments, resulting in a user experience consistent with the scene.
   1. [SYS] Audio elements shall be rendered consistently with their corresponding visual elements, if such visual elements exist.
   2. The specification shall support signalling of audio elements that have a fixed position relative to the user orientation and position (e.g., non-diegetic content).

Note: This may already be fulfilled using MPEG-H 3D Audio.

* 1. The specification shall support earcons.

Note: This may already be fulfilled using MPEG-H 3D Audio.

1. The specification shall support dynamic inclusion of audio elements in a sub-scene based on their relevance, e.g., audibility relative to the user location, orientation, direction and speed of movement or any other audio scene change.
   1. The specification shall support metadata to allow fetching of relevant sub-scenes, e.g., depending on the user location, orientation or direction and speed of movement.

Note: A complete audio scene may be divided into a number of audio sub-scenes, defined as a set of audio elements, acoustic elements and acoustic environments. Each audio sub-scene could be created statically or dynamically.

### Requirements on Audio Renderer

1. The specification shall support metadata describing the audio scene.
2. The specification shall support metadata for controlling and restricting the audio scene.

*Note: This may already be fulfilled by MPEG-H 3D Audio.*

1. The specification shall support control (e.g., via metadata or interface) of the audio rendering parameters (e.g., consumption space, player capabilities, etc.).
2. The specification shall support random-access in time (e.g. every 0.5 seconds) and space (e.g. jump within a sub-scene or to a new sub-scene).
3. The specification shall support metadata for enabling transition effects on audio rendering during user jumps between two different listener positions in the audio scene (e.g., fade-out fade-in).
4. The specification shall support metadata for enabling audio zooming (e.g., adjustment of prominence, dialog enhancement, simulation of depth-of-field effect, etc.).

*Note: This may already be fulfilled by MPEG-H 3D Audio.*

1. The specification shall support 3D spatial extent for audio objects, supported by metadata.
2. The specification shall support rendering of the radiation pattern of audio objects and channels, supported by metadata.
3. The specification shall support occlusion of audio elements, supported by metadata.
4. The specification shall support locally captured audio (e.g., user's own voice, side tone) in the audio scene.
5. The specification shall support accurate 3D spatial localization of audio objects (sound sources). Differences in location are with respect to what is perceivable.

### Interfaces and extensions

1. The specification shall support input interfaces for changing the audio scene.

Note: MPEG-H 3D Audio already provides interfaces enabling basic functionality that could be enhanced for MPEG-I Audio.

1. The specification shall enable extension of the rendering functionality (e.g., interfaces to external rendering tools, extension payloads, reserved bit fields etc.).
2. The specification shall support personal HRTFs in the audio renderer, including an interface for providing these filters.
3. The specification shall support personal headphone equalization filters including an interface for providing these filters.

### Presentation Modes

1. The specification shall support 6DoF head-tracked binaural rendering for headphone reproduction.
2. The specification shall support 6DoF head-tracked rendering to loudspeakers (e.g. to immersive configurations such as 7.1 + 4H).

Note: The user explores the scene by moving in the listening area and based on his position the audio is rendered.

1. The specification shall support 6DoF rendering to loudspeakers for the use case that the user's consumption position is fixed, while the virtual position changes.

Note: The user is located in the sweet-spot and navigates the scene for example using a joystick.

1. The specification shall support rendering to a combination of 6DoF head-tracked binaural headphones reproduction and loudspeaker reproduction.

### Social VR

1. The specification shall support rendering of speech and audio from other users in the virtual environment. The speech and audio may be immersive.
   1. The specification shall support low-latency conversation between users within a given virtual environment.
   2. The specification shall support low-latency conversation between a user within the given virtual environment and a user outside the given virtual environment.
   3. [SYS] The specification shall enable synchronization of audio and video of users and the scene.
   4. The specification shall support metadata specifying restrictions and recommendations for rendering of speech/audio from the other users (e.g. on placement and sound level).

### Interoperability between 3DoF and 6DoF audio platforms

1. The specification shall support decoding and presentation of MPEG-H 3D Audio Low Complexity Profile content on an MPEG-I 6DoF platform with an experience as with an MPEG-H 3D Audio Low Complexity Profile decoder.
2. The specification should support consumption of MPEG-I Audio content on MPEG-H 3D Audio Low Complexity Profile decoder (3DoF).

Note: By using the MHAS format this requirement can be fulfilled.

1. The specification shall enable consumption of MPEG-I Audio content on MPEG-I Audio platforms with reduced degrees of freedom e.g., 3DoF+, 3DoF, 0DoF platforms.

## MPEG-I Container for 6DoF Content Media

1. The specification shall support the storage of the defined Media Types.
2. The specification shall support the presentation and playback of different Media Types.
   1. The specification shall support the storage and presentation of multiple different Media Types such that they are synchronized both spatially and temporally.

*e.g. a 2D background video surrounding a point cloud object.*

1. The specification shall support the presentation of content media depending on the user-selected location and view within a 6DoF scene.

## Composition Description for 6DoF Content Media

Note: A separate requirements document exists for scene descriptions (N17537). The requirements in this section may contain overlaps with N17537; these overlaps will be removed in updated versions.

### *Spatial and Temporal Alignment of 6DoF Content Media*

*Note: This section is meant to target metadata which could be included into future OMAF amendments (not using scene graphs)*

1. The specification shall support metadata which allows natural and smooth switching between real camera captured views and virtual (interpolated) camera views within the 6DoF scene.
2. The specification shall support metadata for enabling the client to synthesize a Viewpoint that was not captured by a camera, using omnidirectional videos from multiple camera-captured Viewpoints as inputs.

### *Descriptions for Content Interactivity*

1. The specification shall enable realistic composition of a 6DoF scene depending on the user-selected location and orientation.

Note: Such composition may, e.g., include delivering proper lighting information and some form of geometry information of the scene so the view is rendered with realistic lighting and shadows.

1. The specification shall support a user interacting with objects within a given virtual environment.
   1. The specification shall support giving an object a place in the virtual environment, moving it and changing some attributes

Example: a rectangular screen that represents a TV set on which a different channel could be chosen; a point cloud object that is inserted into a representation of the real world

1. The specification shall support metadata describing objects and their features within a 6DoF scene.
   1. This metadata will be defined in updated versions.
2. The specification shall support replacement of 6DoF content media data.
   1. The specification shall support metadata for the processing of object features, namely preserving and replacing features for objects in a scene, either by the client, or in the network.

*Note: A typical scene may have multiple objects, with indices which describe their features.*

### *Descriptions for Multi-User Interactivity*

1. The specification shall enable realistic composition of user-embodiment within 6DoF content media.

Note: Such composition may, e.g., include delivering proper lighting information and some form of geometry information of the scene so user-embodiment is rendered with realistic lighting and shadows.

1. The specification shall support rendering of other users in 6DoF content media, including possible speech or audio from other users.
2. The specification shall enable multi-user VR applications in which several users are experiencing the same VR experience together.
   1. It shall be possible to detect & render interactions between users within the VR environment.
3. The specification shall support conveying metadata about the spatial alignment of a camera and the person that this camera is capturing, in the physical environment. This metadata shall include at least:
   1. the distance between the camera and the captured person
   2. the orientation of the camera
   3. the focal length and possibly other lens parameters
   4. the location of the captured person’s head in the video

## Delivery of 6DoF Content Media

1. The specification shall support low-delay delivery of different Media Types for 6DoF contents~~.~~
2. The specification shall support the partial delivery and presentation of content media depending on the user selected location and view within the scene.
3. The specification shall support full or partial delivery of media content depending on different network conditions, device capabilities and configurations.
4. The specification shall support differentiated service access by clients, through adaptation or partial delivery of the media content.

### *Delivery Requirements for Content Interactivity*

1. The specification shall support low delay processing and presentation of object features for a scene, in order to minimize motion-to-photon latency.
2. The specification shall support defining conditional switching between viewports
3. The specification shall support hotspots that trigger actions like switching viewports.
4. The specification shall support signalling how content needs to loop back or continue playing, where this behaviour may be triggered by certain interactive conditions

### *Delivery Requirements for Multi-User Interactivity*

1. The specification shall enable multi-user VR applications in which several users are experiencing the same VR experience together.
   1. The specification shall enable synchronous play-out between multiple users consuming the same scene, where the synchronization is accurate within 100 ms.

Example: multiple users viewing the same sports event and communicating about the event in real-time

Note: the end-to-end latency of the experience is much less important than the inter-user sync.

1. The specification shall support synchronization of audio and video of users and the scene.
   * 1. Synchronization among users of all interactions (with VR objects and between users) shall be supported.
2. The specification shall support the low-latency delivery of interactions between users within a given virtual environment.
3. The specification shall support interactions where it will be rendered to a certain group of users.
   1. The specification shall support privacy protection features related to the delivery of client specific metadata (e.g. position, orientation etc), especially for interactions related to bi-directional delivery applications (e.g. social VR).

*NOTE: This requirement needs to be further clarified.*

## Processing and Adaptation of 6DoF Content Media

1. The specification shall support the network-based media processing framework.
2. The specification shall support the processing (for media distribution) of different Media Types.
3. The specification shall support the adaptation of multiple different Media Types to different types of devices in various network conditions.
   1. The specification shall support level of density adaptation of point cloud media depending on the user selected view location and end-to-end system capabilities.
4. The specification should support interoperable interfaces enabling the processing of different media data to create the scene.

### *Network (pre-)rendering of 6DoF Content Media*

NOTE: A separate requirements document exists for Network Based Media Processing (N17502). The requirements in this section may contain overlaps with N N17502; these overlaps will be removed in updated versions.

1. The specification shall support (pre-)rendering of immersive contents by a network, instead of the client which will consume the contents, where the output will be either 6DoF, 3DoF+, 3DoF, or 2D, and could be compatible with OMAF players.
   1. The specification shall support 6DoF content media metadata for (pre-)rendering by a network.
   2. The specification shall support metadata (such as different client network conditions, device capabilities and configurations) for (pre-)rendering by a network.
   3. The specification shall support remote encoding for converting immersive 6DoF content to a simpler representation, such as 2D image/video.

## Use Case Specific Requirements

Note: Requirements from this section may be updated to more generic requirements which match into the technology clustered sections above.

### Social VR

1. The specification shall support interactions in case of capability mismatching of user devices.
2. The specification shall support that a user can recognize objects to interact with.
3. The specification shall support metadata that provides the position/orientations of remote users in the VR environment of a local user
4. The specification shall support metadata that provides the direction of view of the local user in its VR environment
5. The specification shall support metadata that indicates which remote user is being looked at by the local user
6. The specification shall support metadata that indicates whether a video stream corresponds to user A looking straight into the camera (main camera) or not (side camera)
7. The specification shall support containers (e.g. ISOBMFF) to carry the above metadata to remote users.
8. The specification shall support metadata that provides x/y/z and yaw/pitch/roll coordinates that provide the position/orientation of a camera or audio source with respect to the perceived VR environment.
9. The specification shall support metadata that provides the type and other details of the camera or audio source.
10. The specification shall support metadata to identify and distinguish cameras and audio sources.
11. Metadata to signal details of the visual and/or auditive/haptic indication of the camera/audio source position/orientation in the VR environment
12. Containers (e.g. ISOBMFF) to carry the above metadata to the other user(s).
13. Metadata that provides the angle of view of the camera
14. Metadata that provides the distance between the camera and the video-captured person
15. Metadata about the position of the video-captured person in the video frame
16. Containers (e.g. ISOBMFF) to carry the above metadata to remote users.
17. The specification shall support positioning multiple objects in the same omnidirectional environment where these objects are represented by an image or a video

Note: the primary use case of this requirement is to place multiple users in, e.g., a video conferencing setting in an omnidirectional environment, together with other elements, e.g., a video screen.

* 1. The specification shall enable making the omnidirectional environment consistent:
     1. for all users that are embedded in their omnidirectional environment
     2. *between* the users (they can look at one another, and can see when another person is looking at them)
     3. for multiple users looking at / pointing at a common element in the omnidirectional environment (e.g., a video screen)
  2. The specification shall support synchronization of user viewpoints and orientations (i.e. where each user is looking at), as well as content playback status, between users.

Note: these individual environments need not necessarily be the same, as long as they are internally consistent for all participants individually

Note: “consistent” means the right visual perspective for all objects as well as audio/visual alignment

* 1. The specification shall enable bringing multiple users together in the same omnidirectional environment even when they are captured using their own individual 3DoF coordinate system
  2. The specification shall enable positioning a rectangular object (e.g., a video screen) inside the omnidirectional environment with correct perspective and parallax. The surface of this object may be at an arbitrary distance from the user, and may have an arbitrary orientation in 3D space.

# References

1. ISO/IEC JTC1/SC29/WG11 N18158, " MPEG-I Audio Architecture and Requirements", MPEG Audio Subgroup, January 2019, Marrakech, MA