118th MPEG Hobart (TAS), Australia, 3 - 7 April 2017, Meeting Report

Panos Kudumakis and Krishna Chandramouli
qMedia, Queen Mary University of London

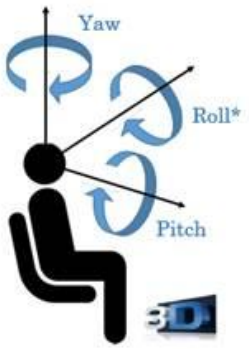
Contents

1 Technical Report on Immersive Media

A first set of specifications is required in time for a market launch of products and services in 2018. It is highly likely that MPEG can deliver solutions that are optimised in a longer time frame, which allows for more experiments and development. Since many believe that major market launch of VR 360 services will happen in 2020, a next set of specifications can be delivered in 2019. At the same time it is clear that there is a strong need for longer term work, notably in the video area, but possibly also in the audio space, on 6-degrees-of-freedom content. Thus, MPEG is planning standards in support of Immersive Media, including those for 360º Audiovisual Media, to be developed in the following phases:

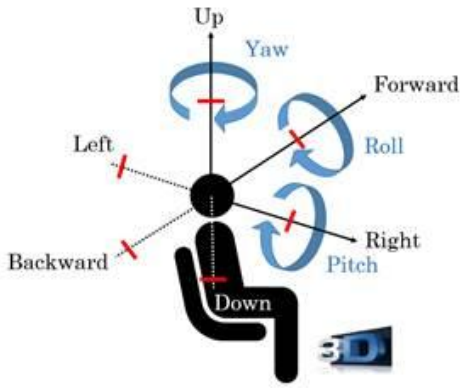
Phase 1a

- Timing is what guides this phase
- Goal: to deliver a Final Draft International Standard for up to 3 degrees of freedom 360 VR by end 2017.
- This phase aims to deliver a complete distribution system.
- Audio: a 3D Audio profile of MPEG-H geared to a 360 Audiovisual experience with 3 DoF,
- Transport: Basic 360 streaming, and if possible optimizations (e.g., Tiled Streaming)
- Video: Adequate tiling support in HEVC (may already exist) and projection, monoscopic and stereoscopic.



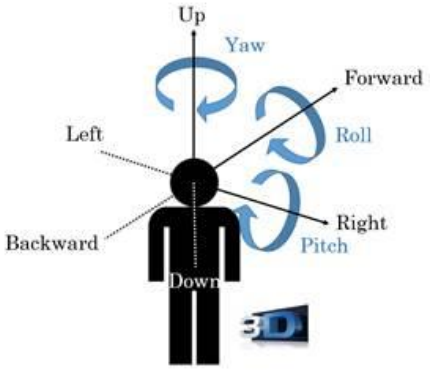
Phase 1b

- This phase is mainly motivated by desire by a significant part of the market to launch commercial services in 2020.
- It is intended for systems and services that deploy in 2020; the specification must be ready in 2019 (which may match 5G deployments).
- Phase 1b will be published as an extension of the Phase 1a specification; it will focus on VR 360 with 3 DoF, with some additional depth clues, that would, for instance, allow moving the viewpoint in a limited space. (Like in phase 1a, including monoscopic and stereoscopic).
- In addition, this phase is thought to comprise elements like:
    - Optimization in projection mapping
    - Further motion-to-photon delay reductions
    - Optimizations for person-to-person communications
- Unlike phase 1a, phase 1b should have some quality definition and verification.



Phase 2

- A specification that is ready in 2021 or maybe 2022
- Goal is support for 6 DoF
- Most important element probably new video codec with support for 6 DoF
- Audio support for 6 degrees of freedom
- Systems elements required in support of 6 DoF, as well as 3D graphics.
- Support for interaction with the virtual environment



Output Documents
N16918 - Working Draft 0.2 of TR: Technical Report on Architectures for Immersive Media

2 Thoughts on MPEG-I and 3D Audio

Providing Audio standards in support of Augmented Reality and Virtual Reality applications, services and systems will occur in stages. A first stage is to standardize technology to support 360-degree movies. This application permits 3 degrees of freedom by the user, and the required audio technology has already been standardized in MPEG-H 3D Audio. A next stage is to extend MPEG-H 3D Audio to 6 degrees of freedom. Technologies to explore may be point clouds that contain environmental meta-data (e.g. pertaining to acoustic characteristics) and compressed audio signals (e.g. ambient audio scenes or local audio sound sources). A more forward-looking stage is to investigate whether capture, compressed representation and reproduction of audio sound fields is a means to a more realistic user experience of virtual audio presentations.

Scenarios

1. 360 Movie – 3DoF

- User is in "sweet spot" and can look in any orientation
  - Pitch, Yaw, Roll sensors detect orientation
  - Audio scene and Video scene render according to orientation
- 3D Audio Second Edition can do this now
  - LC Profile Sensor to ear lag is ~38 ms (algorithmic)
    - Literature suggests sensor to ear lag can be 60 to 100 ms
  - Need to publicize: "MPEG 3D Audio for VR"
- OMAF puts FF, Video, Audio together
- A potential extension of this scenario may be to allow the user to focus on one or more specific region of the video and audio scene (providing a kind of ROI) by applying spatial filtering and audio zoom. A user should be able to determine:
  - The position of the ROI (azimuth and elevation) and smoothly adjust it,
  - Select from a number of ROIs,
  - Adapt the size of a ROI and smoothly adjust it.

2. 360 Movie with head movement – 3DoF+

- User is in "sweet spot" and can look in any orientation, and in addition can move the head in (x,y,z) space. It is assumed that user cannot move from the seat, and so the (x.y,z) translation will be quite limited (hence the "+").

Note that MPEG-H 3D Audio CfP was evaluated with "off-sweet-spot" listening and found to perform with high quality, and so it should be a solution for this problem.

3. Windowed 6DoF

- Scenario: Your TV is a "window" into a virtual world
  - Experience like a movie, but not fully interactive
- User can move from center to either side of "window" to get limited alternate views
  - Renderer must place sound source objects to match view
  - Maybe channel bed doesn't need to change

This use-case represents a limited viewing and listening area in front of a screen.

In general it seems, that this scenario is also well covered by the MPEG-H 3D Audio LC profile. It should be studied to what extend the different listener locations in the virtual world can be covered by screen-related remapping and/or object remapping.

4. Virtual Reality – 3DoF

In this scenario, the user can take discrete positions in the virtual world, e.g. “Choose your seat in the virtual concert hall” or “VR street view” but with only the possibility of discrete positions along a street. The translation of the user is limited to certain predefined positions, while the orientation is defined by the user’s rotation (yaw, pitch and roll)

This scenario can be served already by MPEG-H 3D Audio if content authoring takes into account that the user can jump to different locations in the virtual world. E.g. if all sound sources in the audio scene are authored as objects the position and orientation of objects changes from virtual seat to virtual seat. The rendering takes into account only the relative direction and distance to the user, i.e. it is operating with 3-DOF. The MPEG-H 3D Audio decoder and renderer would need to be supplied with the relative positions of objects. The change of relative object locations could be signalled by using the mpegh3daElementInteraction().

For content that is not (fully) object-based, as may be the case for live-captured content (e.g. a music concert or sports event) or user-captured content, this scenario can be considered as a set of multiple instances of the "360 movie" scenario described above. So also in that case, 3D Audio is already well-equipped to serve this use case.

5. Virtual reality – 6 DoF

In this scenario, the user takes a free view point and orientation in the virtual world. The self-motion is induced by an input controller or sensors. From an audio perspective this seems very close to a gaming scenario. There, sound elements are typically stored as sound objects. With the user moving through the scene with 6 DOF a renderer takes care of appropriately processing the sounds dependent on the position and orientation.

- Full 6 degrees of freedom in world
  - pitch, yaw roll
  - x, y, z
- Example: Tour of Chengdu city streets
  - Virtual or Natural visual scenes
    - Ambient sounds
  - Virtual sound sources
    - Narrator giving information about current locale
- Virtual sound sources populate real environment
  - Sources may have rich metadata (e.g. sound directivity in addition to position)
  - Rendering of sound sources
  - “Dry” sound sources may need
    - distance, velocity treatment
    - environmental acoustic treatment, e.g. room reverberation

In a VR and non-VR gaming applications sounds are mostly stored locally in an uncompressed or weakly encoded form. Therefore, this has never been directly in scope of MPEG audio. However, MPEG-H 3D Audio could be applied in this context, e.g. if certain sounds are delivered from a far end or are streamed from a server. Rendering is critical in terms of latency and far end sounds and local sounds would have to be rendered simultaneously by the audio renderer of the game.
Therefore, it seems desirable to deliver sound elements from MPEG-H 3D Audio decoder by means of an output interface to an audio renderer of the game. Such interfaces exist in the MPEG-H 3D Audio standard and could be applied or adapted to support this use case.

For content that is not (fully) object-based, as may be the case for live-captured content (e.g. a music concert or sports event) or user-captured content, extensions to MPEG-H 3D Audio in the form of spatial conversion or -mapping methodologies may be needed to enable the 6DOF user movement within the virtual environment.

6. Augmented Reality

Augmented reality can be considered as presenting additional requirements on top of those for VR. The additional requirements come from the fact that the renderer has to be aware of the real environment. The additional local sensor data could be

Distance to walls, floor and ceiling
Location and size of nearby objects
Location of other users

Taking into account the information of the real environment would require the renderer to adapt to this information. This could be achieved by physical room modelling and modelling of other acoustic behaviour of objects such as distance, directivity and occlusion. Such behaviour is not part of the MPEG-H 3D Audio specification. It may be challenging to define such rendering behaviour given the missing test methodology and ground truth. However, it seems important to design interfaces of MPEG-H 3D Audio to feed audio renderers with the appropriate information about the virtual objects and the real environment.

Augmented reality could include:

- Audio and Video "Ad Hoc sampled space"
  - Dense sampling of 3D space with audio and video signals
    - Metadata can specify aspects of acoustic environment
    - AV signals can be "real" and augmented
    - Free navigation
  - Compression of audio signal at "sampling points"
  - Conditional transmission

Output Documents
N16753 - Thoughts on MPEG-I and 3D Audio

3 Common Media Application Format

Common Media Application Format is optimized for delivery of a single adaptive multimedia presentation to a large number and variety of devices; compatible with a variety of adaptive streaming, broadcast, download, and storage delivery methods. Several MPEG technologies have been adopted for much of the video delivered over the Internet and other IP networks (cellular, cable, broadcast, etc.). Various organizations have taken MPEG's core coding, file format and system standards, and combined them into their own specifications for their specific applications. While these specifications share major common parts, their differences result in both unnecessary duplication of engineering effort, and duplication of identical content in slightly different formats that increases storage and delivery costs. CMAF allows application consortia to reference a single MPEG specification (a "common media format") that allows a single media encoding to be used for many applications and devices.

The scope of Common Media Application Format (CMAF) is the encoding and packaging of segmented media objects for delivery and decoding on end user devices in adaptive multimedia presentations. Segmented Media Objects are derived from encoded tracks for storage, identification, and delivery. Delivery and presentation are abstracted by a hypothetical application model for segmented Media Objects described by a manifest that allows a wide range of implementations without specifying any.

CMAF constrains media encoding and packaging to allow interoperable adaptive delivery of alternative tracks of segmented media packaged as addressable media objects to different devices, over different networks. CMAF defines media objects using the ISO Base Media File format, file constraints, and a corresponding file compatibility brand. CMAF defines Media Profiles and compatibility brands that specify media track formats, codecs, and encoding constraints. CMAF defines Presentation Profiles that are conditionally required sets of CMAF Tracks that can be adaptively selected and seamlessly switched during playback. This enables most Internet devices to play an adaptive CMAF Presentation conforming to a specified CMAF Presentation Profile.

A manifest and player are assumed in the hypothetical application model. The manifest describes a CMAF presentation and its media Resources, which reference addressable CMAF Media Objects. A player can interpret a manifest, select, decode, synchronize, and present the CMAF Media Objects packaged as Resources in a continuous multimedia presentation consistent with the encoded CMAF Presentation. CMAF does not specify a manifest, player, or delivery protocol, with the intent that any that meet the functional requirements can be used.

Contents

- *CMAF Hypothetical Application Model, Media Object Model, and Profiles* describes the segmented media playback model and the associated objects defined by the CMAF.
- *The CMAF Track Format* describes the use of ISO file format for the Common Media Format brand.
- *Common Encryption of Tracks* details how digital rights management information and encryption is applied to the Common Media Format.
- *CMAF Video Tracks* describes the general video track format, specifics for NAL Structured Video tracks, and the AVC video format.
- *CMAF Audio Tracks* describes the general audio track format, and specifics for AAC Media Profiles.

- *Subtitles and Captions* describes the subtitle track format, and specifics Media Profiles for WebVTT and IMSC1 TTML subtitles, and signaling of CEA 608/708 captions embedded in video streams.
- *CMAF Media and Presentation Profiles* describes the general requirements for CMAF Media Profiles and CMAF Presentation Profiles.

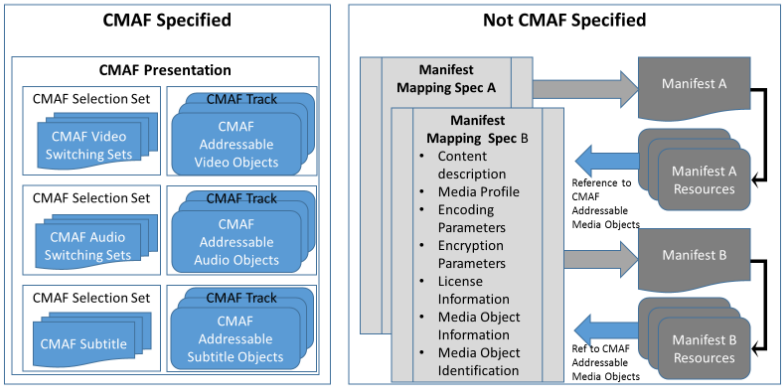


Figure 1 - CMAF Hypothetical Application Model using externally defined Manifests that describe the CMAF Presentation and Media Objects

Output Documents
N16818 - DoC on ISO/IEC DIS 23000-19 Common Media Application Format
N16819 - Text of ISO/IEC FDIS 23000-19 Common Media Application Format
N16820 - DoC on ISO/IEC 23000-19 PDAM 1 SHVC media profile and additional audio media profiles
N16821 - Text on ISO/IEC 23000-19 DAM 1 SHVC media profile and additional audio media profiles
N16822 - Workplan for CMAF conformance

4 Media Orchestration

The media orchestration standard provides specification for the orchestration of media and metadata capture, processing and presentation across multiple devices. The functional components of the specification are (i) orchestration of media capture; (ii) orchestration of media presentation; and, (iii) orchestration of processing.

i. Orchestration of media capture is about metadata and control in terms of which device captures what, when and how. What to capture is about device location, orientation and capture capabilities, e.g. zoom capabilities. When to capture is about synchronization with other devices, as well as start and stop of capture. How to capture is about frame rate, resolution, microphone gain, white balance settings as well as codecs used, metadata delivered, and possible processing to be applied.

ii. Orchestration of media presentation is about metadata and control in terms of which device presents what, when and how. What to present is about what media to retrieve and which parts of that media should be presented. When to present is about presentation synchronization with other devices. How to present is about where exactly to play out something (e.g. positioning of a media part in a screen, positioning of an audio object in a room, and possible processing to be applied).

iii. Orchestration of processing is about metadata and control for applying processing to combinations of captured media and/or metadata. This includes single-media processing (e.g. media synchronization in case of transcoding), as well as processing of multiple media and/or metadata together (e.g. performing video stitching, changing arrangements of media in space and time, or automated editing and selection processes).

Furthermore, the specification supports temporal orchestration at both source and sink, extending the DVD CSS specification. The messaging and control is achieved through the DVD-CSS-WC specification. The timed metadata, which cannot be rendered independently and may affect rendering, processing or orchestration of the associated media data is extended from Part 2 and Part 5 of the MPEG-V specification. The adhoc group carried out the analysis of DoC with comments from US, Germany and The Netherlands national bodies. In addition, joint group meetings were organised with MPEG-2 and MPEG-V.

Output Documents
N16834 - Draft DoC on ISO/IEC CD 23001-13 Media Orchestration
N16835 - Study of ISO/IEC DIS 23001-13 Media Orchestration
N16836 - Technologies under Consideration for ISO/IEC 23001-13 Media Orchestration

5 MVCO Extensions on Time-Segments and Multi-Track Audio

MPEG-21 'Media value Chain Ontology (MVCO) Extensions on Time-Segments and Multi-Track Audio' both in terms of specification (21000-19:2010 FDAM 1) and the associated reference software (21000-8:2008 FDAM 4) promoted to Final Draft Amendment stage at the 118th MPEG meeting, Hobart, AU, 3 - 7 April 2017, that is the penultimate stage before issued as ISO/IEC International Standards.

MVCO facilitates rights tracking for fair and transparent royalties payment by capturing user roles and their permissible actions on a particular IP entity. However, widespread adoption of interactive music services (remixing, karaoke and collaborative music creation) - thanks to MPEG-A: Interactive Music Application Format (ISO/IEC 23000-12) - raises the issue of rights monitoring when reuse of audio IP entities is involved, such as, tracks or even segments of them in new derivative works. This amendment addresses this issue by extending MVCO functionality related to description of composite IP entities in the audio domain, whereby the components of a given IP entity can be located in time, and for the case of multi-track audio, associated with specific tracks. The introduction of an additional 'reuse' action enables querying and granting permissions for the reuse of existing IP entities in order to create new derivative composite IP entities.

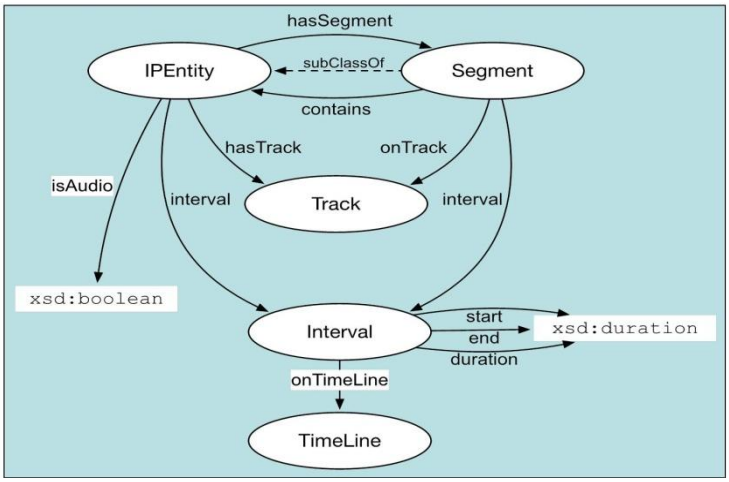


Figure 2 - Classes and relationships for the representation of IPEntity that contain other existing IP Entities. Segments may also be associated with individual Tracks of a Multi-track IP Entity.

With qMedia::C4DM proposed MVCO Extensions on Time-Segments and Multi-Track Audio, it is now possible to query for information about user collectives and the components of the composite IP Entities.

List members of a user collective:

```
$ java -jar rvdac.jar -r -lcu Performers
RVD Administration Console
Guitarist
Vocalist
```

List components of a composite IP Entity (including locations specified by segment and tracks where applicable):

```
$ java -jar rvdac.jar -r -lic MusicInstance
RVD Administration Console
LyricsInstance | segment: 30s to 150s | track: 2
GuitarInstance | track: 1
```

Further information on this particular use case can be found in the MVCO Extensions on Time-Segments and Multi-Track Audio Guidelines Document.

Output documents
N16798 - Request of ISO/IEC 21000-8:2008 AMD 4 MVCO Extensions
N16799 - Text of ISO/IEC 21000-8:2008 FDAM 4 MVCO Extensions
N16815 - DoC on ISO/IEC 21000-19:2010/DAM 1 Extensions on time-segments and multi-track audio
N16816 - Text of ISO/IEC 21000-19:2010/FDAM 1 Extensions on time-segments and multi-track audio