MPEG-H Audio: Fitting Every Device

Jean Cruypenynck

May 2015

Abstract

While "linear" or traditionnal TV is getting stuck because of the expanding variety of media (such as TV, computers, tablets or phones) and new means of consumption, MPEG-H has been defined to be flexible; this is especially true when talking about the audio part of MPEG-H.

Introduction

Born in 2013, MPEG-H is the successor of MPEG-4. Still under development, it has yet got its main standards: "MPEG Media Transport", "High Efficiency Video Coding" and "3D Audio". These three first parts (formerly known as "Systems", "Video", and "Audio" in MPEG-2 and MPEG-4) constitute the core of MPEG-H. Widely known for its video part (H.265/HEVC), MPEG-H defines nevertheless a very interesting audio part. Three main audiovisual stakeholders, to wit *Fraunhofer IIS*, *Technicolor* and *Qualcomm* form the "MPEG-H Audio Alliance", responsible for this part.

1 Object-oriented audio

Let us consider a usual audio broadcasting framework, composed of three parts:

1. *Recording*, either capturing or creating audio information.
2. *Channel* transporting the recorded information with an efficient representation.
3. *Playback* re-creating the original information.

Traditionally, when working with multichannel audio, we try to match the number and position of microphones for the *recording* and channels for the *playback*.
It means that we have to provide as many different mixes as there are different configurations (see Figure 1). This way of thinking worked until now with mono, stereo and 5.1 signals but does not really suit the new needs of the audiovisual industry in this day and age: flexibility, mobility, immersion.

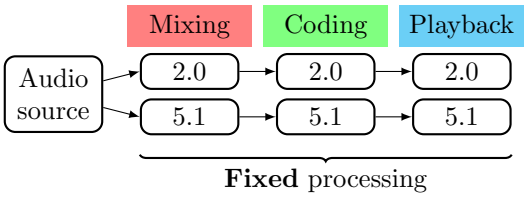

Figure 1: Traditionnal audiovisual framework

Here comes the *object-oriented audio* which, instead of working with channels, uses entities (as soundscape, commentary, sound FX, music, etc.).
The sound mixer creates only one mix by placing these entities in a 3D soundfield.
Then, these entities can be sent in the channel as data and position metadata.
Finally, the original audio can be re-created "easily" by remapping the objects in our given 3D audio soundfield.

There are some direct advantages coming from this method:

- Suitable to any configuration: knowing the number and the position of the loudspeakers, the decoder can decide where to map an object to render it at the best position.
- Allows easy re-mixing: as objects in the audio streams are intended to correspond to real-life objects (for instance, commentary and ambience in a sport event), viewer can adjust their own sound level with ease.

Nonetheless, one can quickly understand the issues tackled by this method:

- Sound scenes are often composed by an ambience and not only small and well-located objects.
  Consequently, this ambience should be represented in a channel-dependent format and creates what is called the "bed" on which different objects go.
- If object-oriented concept works well for a relatively low number of objects, it is impossible to apply and become counterproductive with highly complex signals.

An intermediate approach between channel-based solutions (with a high coupling between the *recording* and *playback* parts, but low complexity) and object-based solutions (with good decoupling, but high complexity) should be found.

2 High-Order Ambisonics

Invented in the UK in the early 1970s, ambisonics is a surround-sound technique which consists of giving a speaker-independent representation of a soundfield. This representation, also called *B-format* is then decoded to any specific speaker setup.

At a given point $M(r, \theta, \varphi)$ of a 3D space, ambisonics give us the following equation defining the sound pressure:

$$\left[\sum_{n=0}^{N} j_n(kr) \sum_{m=-n} n a_n^m(t)\, Y_n^m(\theta, \varphi)\right] e^{j\varphi} \xrightarrow[N \to +\infty]{} p(r, \theta, \varphi, t)$$

where:

- $j_n(kr)$ is a Bessel function of the first kind;
- $Y_n^m(\theta, \varphi)$ is the function giving the spherical harmonics;
- N is the "order" of the ambisonics, which determines the accuracy of restitution of the soundfield and the number of ambisonic coefficients (see Figure 2) which is $(N+1)^2$. In practice, $N \geq 3$ gives a good perception and is called in this case "high-order ambisonics" (HOA).

As technology did not enable real-time computation of HOA when it was created, we now have enough power to go until 4th-order.
Considering the 4th-order, it would give us 25 coefficients and by extension, 25 PCM 48kHz signals (if we consider sampling the soundfield at the usual 48kHz frequency).

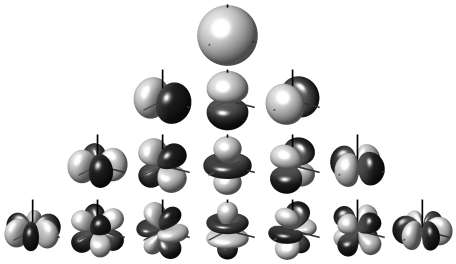

Figure 2: Spherical harmonics up to degree 3 (source *Wikipedia*)

However, these 25 signals are often highly correlated and allow a reduction to 6 PCM channels and some metadata, which can then optionally go through the MPEG-H compression engine.

3 MPEG-H

Hence, MPEG-H Audio "only" gathers the previous parts together. In fact, MPEG-H Audio allows to work with HOA, audio objects and traditional audio, together (see Figure 5).

We can imagine the following scenarios or use cases, as described in Figures 3 and 4:

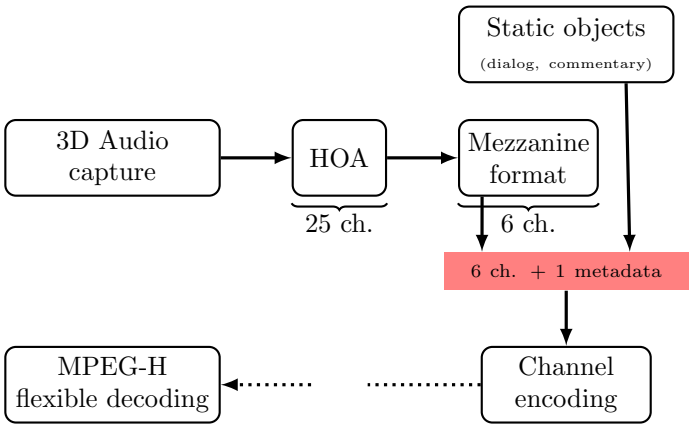

Figure 3: Broadcast worklflow

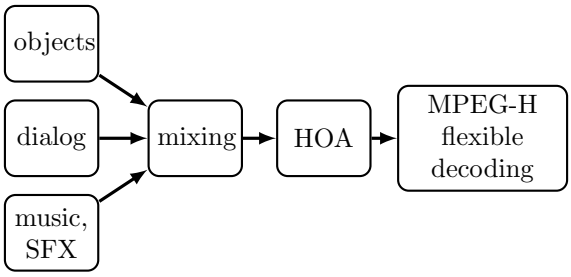

Figure 4: Movie workflow

Conclusion

We covered through this paper main aspects of the multichannel audio part of MPEG-H, which seems to answer most of the expectations of the current broadcast industry and to stay flexible enough to stay the standard during a long time. MPEG-H Audio is now ratified by DVB, ATSC 3.0 and Internet streaming, even though Youtube uses with its VP9 another multichannel audio codec: *Opus*.
Aside from the audio part, the two first parts of MPEG-H show that it stands by itself as the new top-notch codec.

References

[1] Fraunhofer. Mpeg-h audio alliance. `http://www.iis.fraunhofer.de/content/dam/iis/en/doc/ame/MPEG-H_Audio_Alliance.pdf`, 2014. [Online; accessed 10-May-2015].

[2] D. Mercier. *Le livre des techniques du son - Tome 1 - 4e édition - Notions fondamentales*. Audio-Photo-Vidéo. Dunod, 2010.

[3] D. Mercier. *Le livre des techniques du son - Tome 2 - 4e édition: La technologie*. Audio-Photo-Vidéo. Dunod, 2012.

[4] D. Mercier. *Le livre des techniques du son - 4e édition: Tome 3 - L'exploitation*. Audio-Photo-Vidéo. Dunod, 2013.

[5] The Future Trust. The case for mpeg-h audio: Laying the foundation for next generation immersive & interactive audio experiences. `http://thefuturetrust.technicolor.com/article/immersive-experiences-audio/mpeg-h-audio/`, March 2015. [Online; accessed 10-May-2015].

[6] Business Wire. Fraunhofer iis demonstrates real-time mpeg-h audio encoder system for broadcast applications at ibc. `businesswire.com/news/home/20140910005837/en/Fraunhofer-IIS-Demonstrates-Real-Time-MPEG-H-Audio-Encoder`, September 2014. [Online; accessed 10-May-2015].

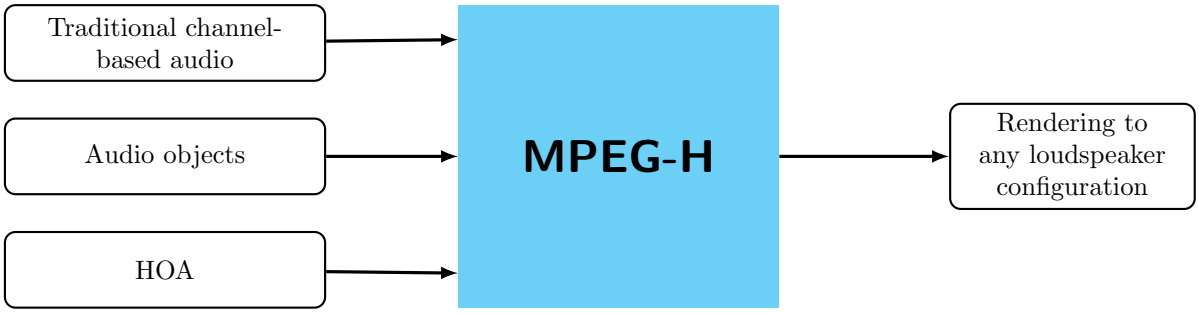

Figure 5: MPEG-H audiovisual framework