IMMERSIVE AUDIO WITH MPEG 3D AUDIO – STATUS AND OUTLOOK

Max Neuendorf, Jan Plogsties, Stefan Meltzer
Fraunhofer Institute for Integrated Circuits (IIS)
Erlangen, Germany

Robert Bleidt
Fraunhofer USA Digital Media Technologies
San Jose, California

ABSTRACT

Arguably, one of the hottest topics currently discussed in the movie and broadcast audio world is what is commonly referred to as "Immersive Audio" or "3D Audio". Immersive Audio literally raises audio reproduction to a new level with the introduction of channels for elevated (and lowered) sound sources. It is increasing the channel count to up to 22 channels plus 2 subwoofer channels. On top of that Immersive Audio provides means for the reproduction of individual audio objects, revolutionizing the way audio has been employed and perceived in a movie theatre, at home, or on portable devices.

This paper takes a close look at the upcoming MPEG 3D Audio open standard framework, which sets out to become the new reference for immersive audio production, transmission and rendering. Originating from the same international standardization body – ISO/IEC MPEG – the standard makes a perfect match for the MPEG HEVC / H.265 video codec and the MPEG-DASH adaptive streaming standard. MPEG 3D Audio boasts an unparalleled feature set including a holistic loudness solution, built-in dynamic range control, object based and parametric based dialogue enhancement, state-of-the-art high efficiency / high quality audio coding technology, flexible rendering to any speaker set-up, binaural rendering, and more. The paper discusses the most relevant features of MPEG 3D Audio, takes a look at the timeline, and evaluates how the standard relates to other ongoing standardization activities, e.g. in ATSC and DVB.

INTRODUCTION

The 5.1 audio systems of today's HDTV were a good match to the improved experience offered by HD video over earlier analog TV broadcasts when developed in the 20th Century and deployed in the last decade. The AC-3 codec used in ATSC and the MPEG AAC technology employed in ISDB and many DVB countries proved how high-quality digital surround audio can be delivered to the consumer. However, the viewers of the 21st Century are less homogenous and more difficult to retain than their 20th Century counterparts.

Viewers no longer watch their 20th Century rear-projection TVs with 5.1 Home Theater In-a-Box systems they purchased to experience DVDs. Today's consumer likely has a $200 two-channel soundbar below their $800 TV, and the focus on sound reproduction has shifted to wife (or husband) -acceptable unobtrusive and convenient products, perhaps at the expense of sound quality. While high-end viewers now have dedicated home theaters, younger consumers watch video primarily on tablets or mobile devices from web sources. All of these consumers are also used to the more interactive and personalized experiences offered by new media today.

Thus, future TV systems must evolve to offer a compelling audio experience on all the target devices for consumers use, while allowing them to tailor that experience to their liking. The MPEG-H standard offers an excellent base of technologies to enable this in future TV audio systems.

Personalization

MPEG-H's object coding will allow consumers to have personalized experiences ranging from simple adjustments, such as increasing or decreasing the level of announcer's commentary or actor's dialogue relative to the other audio elements, to perhaps future broadcasts where several audio elements may be adjusted in level or position to tailor the audio to his liking, as shown in Figure 1.

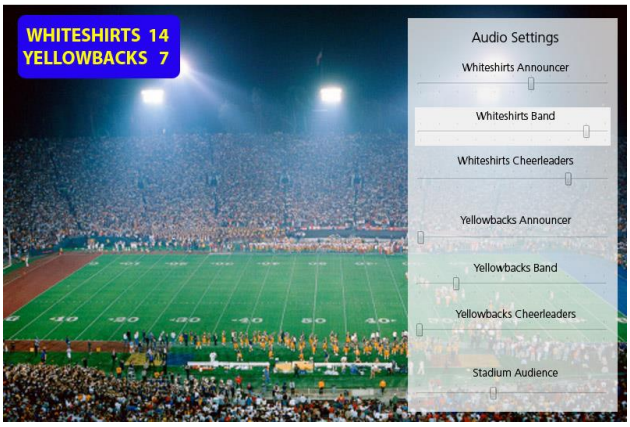

Figure 1. Hypothetical example of a future interactive football broadcast. A "reset button" to return to default settings would likely be needed as well.

Importantly, it is envisioned that this interactivity can be transmitted with very few changes to the existing production process for many content forms such as sports broadcasts.

Here the proposal is simply taking the existing sub-mixes or stems from the remote's audio console and transmitting them as objects, just as they would be prepared for the final mix to air today. This form of interactivity has been judged very desirable in consumer trials that were conducted with the BBC [1] and offers an inexpensive way for broadcasters to add a unique feature to their programming.

In considering a future audio system for television, it is also important to consider the goal of maintaining parity with the consumer's audio experience in the cinema. Cinemas are starting to be equipped for 3D or immersive sound and feature films released with 3D mixes. Sound reproduction in a theater offers different challenges, but it is possible to ingest and transmit a feature film with 3D sound to the home, as might happen on a premium movie channel. More importantly, 3D sound offers broadcasters the opportunity to create a more realistic audio experience for their own content.

Immersion and Realism

The improvement offered by 3D sound over traditional 5.1 or 7.1 systems is substantial, since the realism is significantly enhanced by the reproduction of sound from above. Also, 3D formats offer the ability to localize on-screen sounds vertically, which will become more important as viewing angles increase with the transition to 4K and 8K video. Figure 2 shows the results of a double-blind subjective listening test comparing the overall sound quality obtained from 3D systems in comparison to today's stereo and 5.1 formats [2].

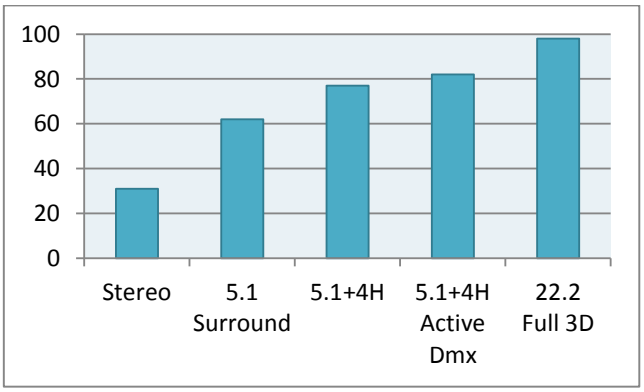

Figure 2. Overall sound quality improvement on a scale from 0 to 100 with expansion of reproduction system to surround and immersive / 3D formats.

Although broadcasters such as NHK have demonstrated full 22.2 channel audio systems, it is fair to say that a direct reproduction system using 24 speakers is impractical in any but the most elaborate home theaters, even though the MPEG-H 3D Audio system can support an even higher number of channels. A more practical recommendation for television broadcasts might be today's 5.1 channels plus four height channels and potentially 4 to 8 audio objects. Due to the improved coding efficiency of the MPEG-H system, broadcast quality audio with 14-18 channels or objects may be transmitted using the same ~400 kbps bitrates used for today's 5.1 broadcasts.

To render these channels in the home needs a way for the average consumer to hear them with a décor-friendly, no-installation, one-box solution. Although it is still an area of current research, the concept of "sound frames" around the viewing screen as opposed to sound bars offers promise as a means to accomplish this where the side, rear, and top channels are reproduced by psychoacoustic techniques [6].

Rendering on all play-back devices

On the other hand, rendering 3D or interactive audio on portable devices such as tablets and mobile phones imposes different challenges. Here there is a need to adapt the audio presentation and rendering to open-air earphones or half-inch loudspeakers in situations where the ambient noise in the listening environment may be 70 or 80 dBA SPL.

Thus, the system must be able to tailor the dynamic range (and usually equalization) for the best experience in these cases, but also offer a satisfying and realistic experience in cases where the viewer is listening to his tablet on good headphones in his bedroom or has his phone docked in his home theater system.

Finally, it is important to improve technical shortcomings of existing TV audio systems as well. In addition to tailoring dynamic range for each platform and environment, future systems must offer very flexible and precise loudness control for CALM act compliance and must adapt to the parallel network delivery strategies envisioned by upcoming transmission standards.

WHAT MPEG-H 3D AUDIO CAN DO

Next generation TV programs will have to serve several ways to deliver content to the user from highest-quality cable and satellite TV down to streaming to mobile devices. At the same time different types of audio content have to carried, from stereo to 5.1, 7.1 and a potentially high number of channels for immersive audio content.

Bit Rate Efficiency

The target bitrate range of MPEG-H 3D Audio specification is designed to meet both the quality and efficiency expectations. For today's 5.1 surround sound material bitrates of 96 to 256 kbps deliver good to excellent quality. For future audio content with a high number of channels (9.1 to 22.2) the same audio quality is achieved at bitrates from 256 kbps up to 1.2 Mbps. The efficiency of the MPEG-H Audio codec allows to carry more quality and/or more channels with the same bit budget, e.g. with commonly used broadcast audio data rates of 384 kbps, 11.1 audio channels for addressing height loudspeakers can be delivered including up to 4 additional object channels.

In the design of MPEG-H, the latest generation of spectral and spatial coding was applied for the coding of

audio channels. Coding tools like bandwidth extension and noise filling, parametric stereo and object coding are combined to reach the highest quality level for the specific bandwidth and application. Although being based on established MPEG AAC, MPEG-H was not designed to be backwards compatible with legacy AAC.

Loudness Normalization

One of the essential features for a next generation audio delivery is proper loudness signaling and normalization. Within MPEG-H, comprehensive loudness related measures according to ITU-R BS.1770-3 or EBU R128 are embedded into the stream for loudness normalization. The decoder normalizes the audio signal to map the program loudness to the desired target loudness for playback. E.g. on a home AVR the target loudness is typically set to -31dB LKFS, while for mobile devices the target loudness is in the range of -12 to -15 dB LKFS. Downmixing and dynamic range control may change the loudness of the signal. Dedicated program loudness metadata can be included in the MPEG-H bitstream to ensure correct loudness normalization for these cases.

Dynamic Range Control

Looking at different target playback devices and listening environments, the control of the dynamic range is essential. In the framework of dynamic range control (DRC) in MPEG, different DRC gain sequences can be signaled that allow encoder-controlled dynamic range processing in the playback device. Multiple individual DRC gain sequences can be signaled with high resolution for a variety of playback devices and listening conditions, including home and mobile use cases. The MPEG DRC concept also provides better clipping prevention and peak limiting.

Audio Objects

Embedding of objects as additional audio tracks inside the audio broadcast opens a range of new applications as described in the previous section. Inside the MPEG-H 3D audio bitstream objects can be embedded that can be selected by the user during playback. Example applications range from different language tracks and visually impaired narration to additional commentary or close-up capture at sport events. The level of specific tracks can be adjusted for a personalized listening experience, e.g. raising the dialogue or commentary level over the background sound.

Objects such as dialogue can be controlled individually in terms of their dynamic range, which ensures audibility for all compression modes.

The notion of objects also allows accurate spatial reproduction of sounds in different playback scenarios. To do so object metadata can be embedded in the bitstream that can describe the geometric position. The MPEG-H decoder contains an object renderer that maps the object signal to loudspeaker feeds based on the metadata and the locations of the loudspeakers in the users home. As a result controlled positioning of sounds can be achieved for regular or unconventional loudspeaker setups, e.g. to align sounds with visual objects on the screen.

Flexible Rendering and Playback on Headphones

For audio production and monitoring the setup of loudspeaker is well defined and established in production practice for stereo and 5.1 [5]. However, in consumer homes loudspeaker setups are typically “unconventional” in terms of placement and diverse regarding the number of speakers. Within MPEG-H flexible rendering to different speaker layouts is implemented by a format converter that adapts the content format to the speaker setup available on the playback side. For well-defined formats specific downmix metadata can be set on the encoder to ensure downmix quality, e.g. when playing back 9.1 content on a 5.1 or stereo playback system.

It is foreseeable that media consumption is moving further towards mobile devices with headphones being the primary way to playback audio. Therefore, a binaural rendering component was included in the MPEG-H 3D audio decoder for dedicated rendering for headphones. The idea is to convey the spatial impression of immersive audio production on headphones.

Streaming

The use case of Internet streaming of audio or video to mobile devices poses very specific challenges to the delivery system as well as the contained media. In particular today's common mobile data connections via 3GPP or even LTE can never guarantee a constant bandwidth. Instead data throughput typically varies considerably as the user changes location. In order to guarantee continued play-back of content, recent devices make use of dynamic adaptive streaming techniques, such as MPEG-DASH, which allow seamless adaptation of the content's bit rate to the current connection quality.

While the idea of adaptive streaming is not new and has been employed in streaming of HE-AAC or MPEG-AVC (H.264), MPEG-H 3D Audio is inherently DASH-ready by design and will therefore allow much easier implementation of DASH-based services. The built-in concept of audio-I-Frames allows shorter tune-in time and much reduced implementation complexity to allow easy bit stream splicing in the production plant as well as ad-insertion even locally at the receiver's end.

MPEG-H may be even better known through its more prominent member HEVC (or H.265), the successor of the world-wide employed AVC or H.264 video codec. Both technologies, "HEVC" and "3D Audio", were developed in the same portfolio of standards and are thus perfectly matched to work in a combined system for A/V delivery.

PERFORMANCE

Inside MPEG several candidate technologies have undergone rigorous testing to select appropriate coding system for immersive audio. Test items were selected to represent typical and critical audio material. In the selection process more than 40000 answers from a total of 10 test labs were collected. A summary of the results and recommended bitrates are listed in Table 1.

Bitrates in kbps for:	Good	Recomm'd	Trans-parent
22.2 Channels	256	512	1200
7.1 + 4 Height Ch. + 4 Obj.	200	**384**	800
5.1 Channels	96	160	256
2.0 Channels	32	56	160

Table 1 Subjective quality test results and estimates for different number of channels for various bitrates

In addition to quality aspects the system had to fulfill several requirements with regards to its capabilities of rendering to different loudspeaker layouts and headphones, random access and latency.

SCHEDULE AND TIMELINE

The work on immersive audio started in MPEG in 2012 with the definition of requirements for an open standard for the delivery of next generation audio content. With the track record of MPEGs successful audio standards, namely the AAC family of codecs, MPEG decided for a Call for Proposals of technology based on existing MPEG audio codecs [3].

After selection of the Fraunhofer IIS system as reference model mid-2013, additional requirements such as Loudness and DRC handling, bitrate adaptation and dialog enhancement were defined to design MPEG-H 3D audio codec towards broadcast and streaming application. Technologies to meet these requirements are integrated into the reference model until March 2014. At that time the MPEG Committee draft for MPEG-H 3D audio Version1 is issued. The final standard is scheduled for publication in February 2015.

APPLICATION STANDARDS

One of the main applications for MPEG-H audio is TV broadcasting, as TV broadcasters have to compete for viewers with cinema, Bluray and VOD delivery. With the introduction of new more immersive audio formats in cinemas, the next generation TV broadcast systems have to deliver a better audio experience than today's HDTV systems. Currently, all major TV broadcast systems are defining the next generation of their standards. While the Japanese ISDB and the European DVB standards are "only" changing the audio and video formats, ATSC is defining a completely new system under the working title "ATSC 3.0". On the video side the resolution will go up to 4k or even 8k with future improvements in bit depth and contrast levels.

The video codec for these new systems likely will be MPEG-H HEVC. On the audio side, diverging approaches are proposed. While the Japanese ISDB system already has selected a 22.2 configuration, the decision for DVB and ATSC 3.0 is still open although the basic requirements are already defined. MPEG-H Audio represents an excellent choice to fulfill these requirements. The 3D audio capabilities together with the flexible rendering enable an immersive audio experience for the user which matches the visual impression of high quality UHDTV pictures. The object based concept offers new opportunities for broadcasters to make their programs more attractive to the user. Elements like dialogue enhancement as well as loudness handling ensure that legal requirements are met by the broadcasters. Beside these new features, the MPEG-H core audio codec is the most efficient audio codec today, and can also be used to transmit stereo and 5.1 multichannel audio as known today.

Beside the use in TV broadcasting, streaming is another important use case for distributing A/V content in the future. We already see today that the borders between broadcasting and streaming delivery are vanishing. Especially with the use of the interactive features for audio as provided by the object based approach, the combined delivery of content over broadcast and streaming will become more commonplace. It is a likely scenario that the main audio content will be delivered as broadcast, while the optional additional audio objects are streamed to the user on an individual basis. With the introduction of audio-I-Frames in MPEG-H Audio the implementation of adaptive streaming is easier than before.

CONCLUSION

The new MPEG-H audio standard is the next generation MPEG audio codec with increased efficiency and more channels to support channel-based 3D audio. With the additional features like audio objects and flexible rendering, it provides the necessary tools for new interactive features and a truly immersive audio experience. Improved loudness and DRC control complete the concept of a more user focused audio standard. To conclude, the MPEG-H audio standard fulfills the requirements for a new audio system for the next generation TV standard as well as adaptive streaming, and is the natural companion of the MPEG-H HEVC video codec for UHDTV.

REFERENCES

[1] Fuchs, Meltzer et al, *Dialog Enhancement – Enabling User Interactivity With Audio*, NAB 2012

[2] Silzle et al, *Investigation on the Quality of 3D Sound Reproduction*, ICSA 2011

[3] ISO/IEC JTC1/SC29/WG11/N13411, *Call for Proposals for 3D Audio*, January 2013, Geneva, CH

[4] ISO/IEC JTC1/SC29/WG11/N13633, *Submission and Evaluation Procedures for 3D Audio*, April 2013, Incheon, KR
[5] ITU-R Recommendation BS.775-3 *Multi-channel Stereophonic Sound System with and without Accompanying Picture*, Geneva (2012)
[6] Sugimoto et al, *A Loudspeaker Array Frame Reproducing 22.2 Multi-channel Sound for Super Hi-Vision Flat Panel Display*, NAB 2012