An Introduction to MPEG-H 3D Audio

Simone Füg[1], Achim Kuntz[2]

[1] Fraunhofer Institut für Integrierte Schaltungen IIS, 91058 Erlangen, E-Mail: simone.fueg@iis.fraunhofer.de
[2] International Audio Laboratories Erlangen, 91058 Erlangen, E-Mail: achim.kuntz@iis.fraunhofer.de

Introduction

With consumers desiring to view and hear content on more devices and in more places than before, producers and broadcasters are increasingly put under pressure to meet demands for new and better sound experiences by delivering enhanced services to a broad range of end-user services and devices reaching from high-end home theaters, to tablets and smart phones.

Recently, a new generation of spatial audio formats were introduced that include elevated loudspeakers. These systems deliver higher spatial fidelity than the established 5.1 setup [1] and surpass traditional surround sound in terms of a more realistic spatial sound experience and a feeling of immersion. Typical examples of such 3D loudspeaker setups include 7.1 with two height channels [2], 9.1 [3] and 22.2 [4].

With these current developments of next generation audio, there is the need for a 3D audio system that offers bitrate-efficient distribution, interoperability and optimal rendering of 3D audio content. Based on such consideration, MPEG has standardized the ISO/MPEG-H 3D Audio standard.

This paper explains the MPEG-H 3D Audio work item and outlines the MPEG-H 3D Audio codec architecture and technology [5].

The MPEG-3D Audio Work Item

The ISO/MPEG standardization group has initiated a work item on the new ISO/MPEG-H 3D Audio standard in January 2013. The need for a system that is able to handle immersive and object-based audio led to a 'Call For Proposals' (CfP) for such 3D Audio technologies in January 2013 [6]. The CfP specifies requirements and application scenarios for the new technology, a development timeline and operating points at which the submitted technologies are tested for their performance.

The tested bit rates range from 1.2 Mbit/s down to 256 kbit/s for a 22.2 input signal. The output was to be rendered on various loudspeaker setups from 22.2 down to 5.1, plus binaural headphone reproduction. The evaluation of submissions was conducted independently for two input content types, i.e. 'channel and object (CO) based input' and 'Higher Order Ambisonics (HOA)'. At the 105th MPEG meeting in July/August 2013, Reference Model technology was selected from the received submissions. The winning technology came from Fraunhofer IIS (CO part) and Technicolor/Orange Labs (HOA part). Both parts were subsequently merged into a single reference system, further developed by the MPEG Audio group.

The international standard was finalized at the 111th MPEG meeting in February of 2015; the publication by ISO is expected in July 2015.

The MPEG-H 3D Audio Codec

MPEG-H 3D Audio has been designed to meet requirements for delivery of next generation audio content to the user, ranging from highest-quality cable and satellite TV down to streaming to mobile devices.

The main features that make MPEG-H 3D Audio applicable for delivery of next generation audio ranging from highest-quality cable and satellite TV down to streaming to mobile devices are outlined in the following sections.

Flexibility with regard to input formats

MPEG-H 3D Audio allows carriage of multiple 3D audio formats. The different supported immersive sound representations can be categorized as follows:

Channel-based audio: The term channel-based audio refers to the transmission of audio content as a set of channel signals which are designated to be reproduced by loudspeakers in defined, fixed target locations relative to the listener. In MPEG, the most popular channel-based formats are listed directly in the specification. Future-proofness with respect to new channel-based layouts is ensured by including advanced flexible mechanisms for signaling loudspeaker layouts and channel mappings.

Object-based audio: In MPEG-H 3D, also audio objects can be embedded. Audio objects are signals that are to be reproduced as to originate from a specific target location that is specified by associated side information. In contrast to channel signals, the target location of audio objects can vary over time. Object-based content is speaker layout agnostic and has to be rendered to the target loudspeaker layout on the reproduction side.

Higher Order Ambisonics (HOA): HOA is an alternative approach to capture a 3D sound field by transmitting a number of 'coefficient signals' [7]. HOA content is also speaker layout agnostic. In addition to channels and objects, also HOA content can be carried in MPEG-3D Audio.

The MPEG-H 3D Audio codec allows for any combination of channel, object and HOA audio content within one MPEG-H audio bitstream. Thus, the most appropriate representations of different elements of a sound scene can be chosen, e.g. a compact HOA representation for immersive ambient sound plus one or several discrete channels for different language tracks.

Flexibility with regard to reproduction

In contrast to audio production and monitoring, where the setup of loudspeakers is well defined, the setup of loudspeakers in consumers' homes often includes non-ideal placement and differs regarding the number of speakers.

Within MPEG-H 3D Audio, a format converter adapts the content format to the actual real-world speaker setup available on the playback side to provide flexible rendering to different speaker layouts. As media consumption is moving further towards mobile devices as smartphones or tables with headphones, a binaural rendering module was included in the MPEG-H 3D audio decoder. This module aims to convey the spatial impression of immersive audio productions on headphones.

The MPEG-H 3D Audio System Architecture

Figure 1 shows an overview of an MPEG-H 3D Audio decoder and major building blocks of the system.

As a first step, all transmitted audio signals are decoded by the MPEG-H 3D Audio core decoder. Channel-based signals are then mapped to the target reproduction loudspeaker layout using the format converter as outlined below. Object-based signals are rendered to the target reproduction loudspeaker setup by the object renderer using the associated object metadata. Alternatively, it is possible to use an extended Spatial Audio Object Coding (SAOC-3D) to parametrically code channel signals and audio objects. These are rendered to the target reproduction loudspeaker setup using the associated metadata. HOA content is rendered to the target reproduction loudspeaker setup using the associated HOA metadata by the HOA renderer.

In the following, the main technical components of the MPEG-H 3D Audio decoder are described.

The MPEG-H 3D Audio Core Decoder

The MPEG-H 3D Audio codec architecture is built around a perceptual codec for compression of the waveforms for the different input signal classes: channels, objects, HOA. This coder is based on MPEG Unified Speech and Audio Coding (USAC) [8]. USAC is the state-of-the-art MPEG codec for compression of mono to multi-channel audio signals at data rates of 8 kbit/s per channel and higher. This technology has been extended by tools to, amongst others, exploit the perceptual effects of 3D reproduction. The extensions include new signaling mechanisms for 3D content and loudspeaker layouts, joint coding of quadruples of input channels in a Quad Channel Element and an improved behavior for instantaneous rate switching or fast cue-in as it appears in the context of MPEG Dynamic Adaptive Streaming (DASH) [9] by using so-called 'immediate playout frames'.

Rendering of channel-based content

The MPEG-H 3D Audio decoder contains a 'format converter' module that renders the decoded channel-based signals to numerous loudspeaker setups making use of high-quality downmixes. The format converter in MPEG-H 3D Audio provides the automatic generation of optimized downmix matrices, taking into account non-standard loudspeaker positions. It includes an advanced active downmix algorithm to avoid downmixing artefacts like signal cancellations or comb-filtering that can occur when combining (partially) correlated input signals in a passive downmix [10]. Besides, it also supports optionally transmitted downmix matrices to preserve the artistic intent of a producer or broadcaster and applies equalizer filters for timbre preservation.

The active downmix in the format converter measures correlation properties between input channels that are

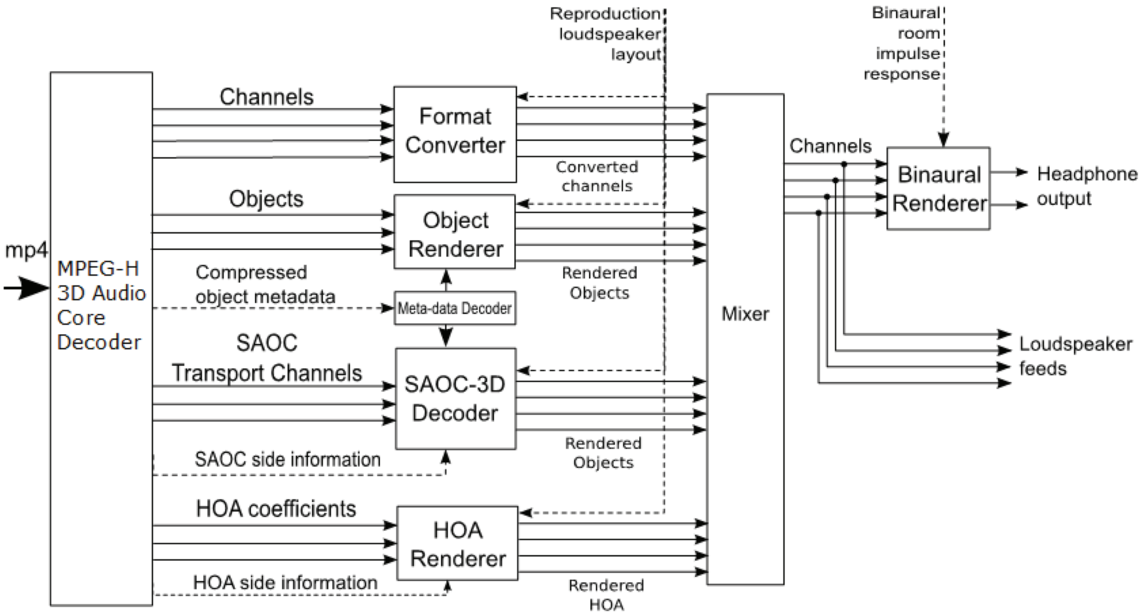

Figure 1: Top level block diagram of the MPEG-H 3D Audio decoder

subsequently combined in the downmix process and aligns the phases of individual input channels if necessary. Secondly, it applies an energy-preserving frequency-dependent normalization to the downmix gains. Uncorrelated input signals remain untouched, thus the algorithm eliminates the artefacts that occur in passive downmixes with only minimum signal adjustments.

Rendering of object-based content

The rendering of audio objects on arbitrary trajectories is realized by an object renderer that applies Vector Base Amplitude Panning (VBAP) [11]. As input the renderer expects the geometry data of the target rendering setup, one decoded audio stream per transmitted audio object and object metadata associated with the transmitted objects, e.g. time-varying position data and gains. The MPEG-H 3D Audio object renderer provides an automatic triangulation algorithm for arbitrary target configurations. The triangulation makes use of imaginary loudspeakers to provide complete 3D triangle meshes for any setup to the VBAP algorithm: The imaginary loudspeakers extend the loudspeaker setup in regions where physical loudspeakers are missing to cover the complete hull around the listener. The signal contributions rendered by VBAP to the imaginary loudspeakers are downmixed to the physically existing loudspeakers.

SAOC-3D decoding and rendering

In MPEG-H 3D Audio the original Spatial Audio Object Coding (SAOC) codec [12, 13] has been enhanced with the possibility to use more than two downmix channels and direct decoding/rendering arbitrary output speaker setups. Some SAOC tools that have been found unnecessary within MPEG-H 3D Audio have been excluded, such as residual coding.

HOA decoding and rendering

Higher order ambisonics (HOA) is based on a truncated expansion of the wave-field into spherical harmonics. The time-varying coefficients of the spherical harmonics expansion are called HOA coefficients and carry the information of the wave field that is to be transmitted or reproduced.

Instead of directly transmitting the HOA coefficients, MPEG-H 3D Audio applies a two-stage coding to improve the coding performance of the system. These two stages have to be reverted in the MPEG-H 3D Audio decoder in opposite order.

In the HOA encoder the sound field is decomposed into predominant and ambient sound components. Predominant components mainly contain directional sounds. Predominant components are transmitted as audio streams together with associated time-variant parametric information (direction of the directional components, activity of the directional components in the field). The ambient sound components mostly contain non-directional sound. Details of the spatial properties of this part of the field are considered less important and the resolution of the ambient component is typically reduced to improve the coding efficiency. As the HOA representation of the ambient component may exhibit high correlations between the HOA coefficients and this can lead to undesired spatial unmasking of the coding noise, the HOA representation is decorrelated by transforming it into a different spatial domain before the coding step.

The HOA rendering itself consists of a simple matrix multiplication of the multichannel HOA representation and a rendering matrix.

Loudness and Dynamic Range Control

Within MPEG-H 3D Audio, comprehensive loudness related measures according to ITU-R BS.1770-3 [14] or EBU R128 [15] can be embedded into the bitstream for the purpose of loudness normalization. The decoder normalizes the audio signal to a desired target level to achieve consistent loudness over several subsequent program items such that manual volume adjustments by the user can be avoided or regulatory requirements can be met.

MPEG-H includes a framework of dynamic range control (DRC). Multiple individual DRC gain sequences can be signaled that allow encoder-controlled dynamic range processing in the playback device for a variety of devices and listening conditions, including home and mobile use cases. Individual DRC processing can be applied to the elements contained in an MPEG-H 3D Audio bitstream, e.g. to optimize intelligibility in adverse listening situations. The MPEG DRC concept also provides improved clipping prevention and peak limiting.

Personalized Playback and Interactivity

A special metadata scheme in MPEG-H 3D audio allows for personalized playback options ranging from simple adjustments, such as increasing or decreasing the level of dialog relative to the other audio content, to broadcasts where several audio objects may be adjusted in level or position to adapt the audio playback experience to the user's liking. An overview of MPEG-H audio metadata is provided in [16].

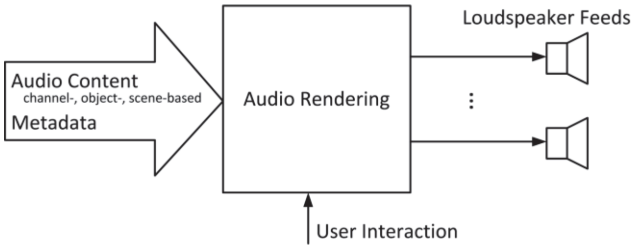

Figure 2: Concept of rendering audio tracks to a reproduction loudspeaker layout, taking into account associated metadata and possible user interactions.

With the metadata definition, MPEG-H 3D Audio supports several use-cases for audio interactivity and object-based audio, such as changing the position of sound events, changing the language of a program, enabling of additional dialog tracks, choosing between content versions and automatic screen-related audio scene scaling.

The metadata consists of descriptive metadata containing information about the existence of objects inside the bitstream and high-level properties of audio elements. It also contains restrictive metadata that defines how interaction is possible or enabled by the content creator. Structural metadata allows for grouping and combination of audio content. Each object-based audio track has in addition associated dynamic object metadata that describes the temporal change of the object position, a linear gain, and a spread angle that describes the energy distribution of an object in azimuth and elevation direction and a dynamic object priority value.

The metadata is defined in a way such that content creators can configure possible user interaction. Therefore, different user interaction categories and interaction ranges are defined:

On-Off Interactivity: A group of audio tracks can interactively be switched on or off. The content of the group is either played back or discarded.

Gain Interactivity: The level/gain of a group of audio tracks can interactively be changed. The amount of possible gain change can be restricted by metadata fields.

Position Interactivity: The position of a group of objects can interactively be changed. The ranges for azimuth and elevation offset, as well as a distance change factor can be restricted by metadata fields.

In addition, it is possible to define presets to control the user interaction. Presets can be used to offer pre-configured combinations of groups for a more convenient selection by the user.

Summary and Conclusions

With a new generation of audio systems emerging, featuring reproduction setups including height loudspeakers, as well as different representations of 3D audio, the need for an audio codec arose that is capable of meeting the new requirements. Thus ISO initiated the work on MPEG-H 3D Audio, resulting in an international standard that allows for flexible coding of channel, object and HOA content and combinations thereof.

The now finalized MPEG-H audio codec includes renderers that can generate output signals for arbitrary loudspeaker setups as well as binauralized headphone output. Further, the decoder allows for personalization of the presentation by the listener: The decoder output can be tailored to different reproduction devices, different listening environments, preferred language settings, as well as the preferred mix balance between ambience, dialogue or commentary tracks. In addition, DRC and loudness processing ensures adaption of the dynamic range to the listening situation and consistent loudness over different program items.

Literature

[1] Silzle, A. et al.: Investigation on the Quality of 3D Sound Reproduction. International Conference on Spatial Audio (ICSA). Detmold, Germany, 2011.

[2] Chabanne, C. et al: Surround Sound with Height in Games Using Dolby Pro Logic IIz, 129th AES Convention, San Francisco, USA, November 2010.

[3] Daele, B. V.: The Immersive Sound Format: Requirements and Challenges for Tools and Workflow, International Conference on Spatial Audio (ICSA), Erlangen, Germany, 2014.

[4] Hamasaki, K. et al: The 22.2 Multichannel Sounds and its Reproduction at Home and Personal Environment, AES 43rd International Conference on Audio for Wirelessly Networked Personal Devices, Pohang, Korea, September 2011.

[5] ISO/IEC JTC1/SC29/WG11 N14747, Text of ISO/MPEG 23008-3/DIS 3D Audio, Sapporo, July 2014.

[6] ISO/IEC JTC1/SC29/WG11 N13411: Call for Proposal for 3D Audio, Geneva, January 2013.

[7] Spors, S. and Ahrens J.: A comparison of wave field synthesis and higher-order ambisonics with respect to physical properties and spatial sampling, 125th AES Convention, San Francisco, USA, October 2008.

[8] Neuendorf, M. et al: The ISO/MPEG Unified Speech and Audio Coding Standard - Consistent High Quality for All Content Types and at All Bit Rates, Journal of the AES, Vol. 61, No. 12, December 2013, pp. 956-977.

[9] ISO/IEC 23009-1:2012(E), Information technology - Dynamic adaptive streaming over HTTP (DASH) - Part 1: Media presentation description and segment formats, 2012.

[10] J. Vilkamo, A. Kuntz, S. Füg: Reduction of Spectral Artifacts in Multichannel Downmixing with Adaptive Phase Alignment, Journal of the Audio Engineering Society, Volume 62, Issue 7/8, July 2014, pp. 516-526

[11] Pulkki, V.: Virtual sound source positioning using vector base amplitude panning. Journal of the Audio Engineering Society, Volume 45, Issue 6, June 1997, pp. 456-466.

[12] Herre, J. et al: MPEG Spatial Audio Object Coding – The ISO/MPEG Standard for Efficient Coding of Interactive Audio Scenes, Journal of the AES, Vol. 60, No. 9, September 2012, pp. 655-673.

[13] ISO/IEC 23003-1:2010, MPEG-D (MPEG audio technologies), Part 2: Spatial Audio Object Coding, 2010.

[14] ITU-R, Recommendation-BS1770.3. Algorithms to measure audio programme loudness and true-peak audio level, International Telecommunications Union, Geneva, Suisse, 2012.

[15] European Broadcasting Union (EBU), Recommendation R128. Loudness Normalization and permitted maximum Levels of Audio Signals, Geneva, Suisse, 2011.

[16] Füg, S. et al.: “Design, Coding and Processing of Metadata for Object-Based Interactive Audio”, 137th AES Convention, Los Angeles, USA, October 2014.