MPEG-H Audio - The New Standard for Universal Spatial / 3D Audio Coding

Jürgen Herre[1], Johannes Hilpert[2], Achim Kuntz[1], and Jan Plogsties[2]

[1] International Audio Laboratories Erlangen, Erlangen, Germany
A Joint Institution of Universität Erlangen-Nürnberg and Fraunhofer IIS

[2] Fraunhofer IIS, Erlangen, Germany

ABSTRACT

Recently, a new generation of spatial audio formats were introduced that include elevated loudspeakers and surpass traditional surround sound formats, such as 5.1, in terms of spatial realism. To facilitate high-quality bitrate-efficient distribution and flexible reproduction of 3D sound, the MPEG standardization group recently started the MPEG-H Audio Coding development for the universal carriage of encoded 3D sound from channel-based, object-based and HOA-based input. High quality reproduction is supported for many output formats from 22.2 and beyond down to 5.1, stereo and binaural reproduction - independently of the original encoding format, thus overcoming incompatibility between various 3D formats. The paper describes the current status of the standardization project and provides an overview of the system architecture, its capabilities and performance.

1. INTRODUCTION

The faithful reproduction of the spatial aspects of recorded sound has been an ongoing topic for a very long time, starting with two-channel stereophony [1,2], continuing with multi-channel ('surround') sound reproduction [3,4,5], Ambisonics [6] and wavefield synthesis (WFS) [7]. While the vast majority of proposed technologies have been using a number of loudspeakers that surround the listener(s) within a horizontal plane, there recently has been a significant move towards adding 'height' or ' lower' loudspeakers above or below the listener's head in order to create an even more enveloping and realistic spatial sound experience. Typical examples of such '3D' loudspeaker setups include 7.1 with two height channels [8], 9.1 [9] and 22.2 [10]. While such loudspeaker setups clearly can deliver higher spatial fidelity than the established 5.1 setup [11,12,13,14], the adoption of 3D setups poses a number of challenges for production, distribution and rendering:

- How can the sound engineer/Tonmeister make best possible use of 3D loudspeaker setups? The answer to this may very well require a learning process similar to that at the transition from stereo to 5.1.

- In contrast to the traditional 2D surround world, where 5.1 is an established standard for content production, distribution and rendering, there is a plethora of concurrent proposals for 3D loudspeaker setups competing in the market. It currently seems quite unclear, whether one predominant format will evolve which eventually can serve – similar to 5.1 for 2D – as a common denominator for content production, digital media and consumer electronics to create a thriving new market.

- How can 3D audio content be distributed efficiently and with the highest quality, such that existing distribution channels (possibly including wireless links) and media can carry the new content?

- How can consumers and consumer electronics manufacturers accept these new formats, given that many consumers may be willing to install just a single 3D loudspeaker setup in their living room with a limited number of speakers. Can they, nonetheless, enjoy content that was produced for, say, 22.2 channels?

Based on such considerations, the ISO/MPEG standardization group has initiated a new work item to address aspects of bitrate-efficient distribution, interoperability and optimal rendering by the new ISO/MPEG-H 3D Audio standard.

This paper describes a snapshot of MPEG-H 3D Audio Reference Model technology [15] as of the writing of this paper, i.e. after the 109th MPEG meeting in July 2014. It is structured as follows: given that coding of multi-channel/surround sound has been present in MPEG Audio for a considerable time, the existing MPEG Audio technology in this field is briefly introduced. Then, the MPEG-H 3D Audio work item is explained and the MPEG-H Reference Model architecture and technology are outlined. Finally, we show results of some recent performance evaluation of the new technology, followed by a number of expected or possible further developments of the Reference Model.

2. PREVIOUS MPEG AUDIO MULTI-CHANNEL CODING TECHNOLOGY

The first commercially-used multi-channel audio coder standardized by MPEG in 1997 is MPEG-2 Advanced Audio Coding (AAC) [16,17], delivering EBU broadcast quality at a bitrate of 320 kbit/s for a 5.1 signal. A significant step forward was the definition of MPEG-4 High Efficiency AAC (HE-AAC) [18] in 2002/2004, which combines AAC technology with bandwidth extension and parametric stereo coding, and thus allows for full audio bandwidth also at lower data rates. For carriage of 5.1 content, HE-AAC delivers quality comparable to that of AAC at a bitrate of 160 kbit/s [19]. Later MPEG standardizations provided generalized means for parametric coding of multi-channel spatial sound: MPEG-D MPEG Surround (MPS, 2006) [20,21] and MPEG-D Spatial Audio Object Coding (SAOC, 2010) [22,23] allow for the highly efficient carriage of multi-channel sound and object signals, respectively. Both codecs can be operated at lower rates (e.g. 48 kbit/s for a 5.1 signal). Finally, MPEG-D Unified Speech and Audio Coding (USAC, 2012) [24,25] combined enhanced AAC coding with state-of-the-art full-band speech coding into an extremely efficient system, allowing carriage of e.g. good quality mono signals at bitrates as low as 8 kbit/s. Incorporating advances in joint stereo coding, USAC is capable of delivering further enhanced performance compared to HE-AAC also for multi-channel signals. For the definition of MPEG-H 3D Audio, it was strongly encouraged to re-use these existing MPEG technology components to address the coding (and, partially, rendering) aspect of the envisioned system. In this way, it was possible to focus the MPEG-H 3D Audio development effort primarily on delivering the missing functionalities rather than on addressing basic coding/compression issues.

3. THE MPEG-H 3D AUDIO WORK ITEM

Dating back to early 2011, initial discussions on 3D Audio at MPEG were triggered by the investigation of video coding for devices whose capabilities are beyond those of current HD displays, i.e. Ultra-HD (UHD) displays with 4K or 8K horizontal resolution. With such displays a much closer viewing distance is feasible and the display may fill 55 to 100 degrees of the user's field of view such that there is a greatly enhanced sense of visual envelopment. To complement this technology vision with an appropriate audio component, the notion of 3D audio, including elevated (and possibly lower)

speakers was explored, eventually leading to a 'Call For Proposals' (CfP) for such 3D Audio technologies in January 2013 [26]. The CfP document specified requirements and application scenarios for the new technology together with a development timeline and a number of operating points at which the submitted technologies should demonstrate their performance, ranging from 1.2 Mbit/s down to 256 kbit/s for a 22.2 input. The output was to be rendered on various loudspeaker setups from 22.2 down to 5.1, plus binauralized rendering for virtualized headphone playback. The CfP also specified that evaluation of submissions would be conducted independently for two accepted 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 (4 for CO and 3 for HOA) based on their technical merits to serve as the baseline for further collaborative technical refinement of the specification. Specifically, the winning technology came from Fraunhofer IIS (CO part) and Technicolor/Orange Labs (HOA part). In a next step, both parts were subsequently merged into a single harmonized system. Further improvements are on the way, e.g. binaural rendering. The final stage of the specification, i.e. International standard, is anticipated to be issued at the 111th MPEG meeting in February of 2015.

4. THE MPEG-H REFERENCE MODEL

4.1. General Features and Architecture

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 to this wide range of applications and the different associated playback scenarios are outlined in the following sections.

4.1.1. Flexibility with regard to input formats

A future-proof audio system has to accept multiple formats that are or will become established in music, movie production and broadcast. Generally, multi-channel and 3D audio content falls into the following categories:

- *Channel-based*: Traditionally, spatial audio content (starting from simple two channel stereo) has been delivered as a set of channel signals which are designated to be reproduced by loudspeakers in a precisely defined, fixed target location relative to the listener.

- *Object-based*: More recently, the merits of object-based representation of a sound scene have been embraced by sound producers, e.g. to convey sound effects like the fly-over of a plane or space ship. 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 actual placement of audio objects can vary over time and is not necessarily pre-defined during the sound production process but by rendering it to the target loudspeaker setup at the time of reproduction. This may also include user interactivity.

- *Higher Order Ambisonics (HOA)* is an alternative approach to capture a 3D sound field by transmitting a number of 'coefficient signals' that have no direct relationship to channels or objects.

The following text discusses the role of these format types in the context of MPEG-H 3D Audio.

Channel-based 3D Audio formats:

The improvement offered by 3D sound over traditional 5.1 or 7.1 systems is substantial, since the spatial 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 1 shows the results of a subjective listening test comparing the overall sound quality obtained from 3D systems in comparison to today's stereo and 5.1 formats.

In MPEG-H 3D Audio, the most popular channel-based formats are listed directly in the MPEG specification. Beyond this, other alternative production formats are addressed by including more advanced flexible signalling mechanisms, thus ensuring future proofness.

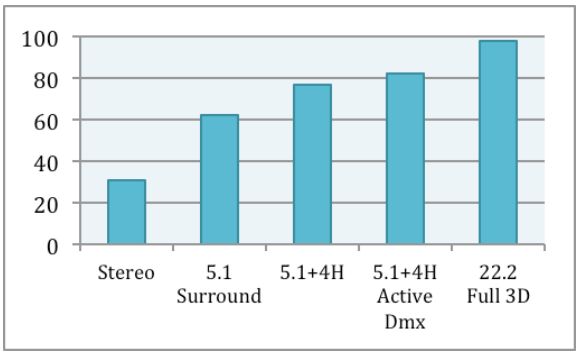

Figure 1 - Overall sound quality impression on a MUSHRA scale from 0 to 100 relative to a 22.2 reference with increasing number of reproduction channels from stereo to surround and immersive / 3D formats, for further details see [11].

Audio Objects:

Using audio objects or embedding of objects as additional audio tracks inside channel-based audio productions and broadcast opens up a range of new applications. Inside an MPEG-H 3D audio bitstream, objects can be embedded that can be selected by the user during playback. Objects allow consumers to have personalized playback options 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 conceivable future broadcasts where several audio elements may be adjusted in level or position to tailor the audio playback experience to the user's liking, as illustrated in the following Figure.

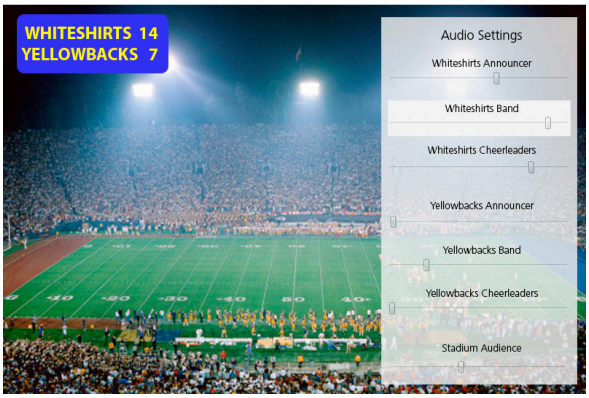

Figure 2 - Thought example of a future interactive American football broadcast

Moreover, audio objects such as dialogue can be controlled individually in terms of their dynamic range, which ensures best intelligibility and supports dedicated reproduction for hearing impaired listeners.

The notion of objects also allows accurate spatial reproduction of sounds in different playback scenarios. Therefore, object metadata that describes the geometric position of the sound sources contained in the objects can be embedded in the bitstream. 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 user's 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.

HOA:

The concept of higher order ambisonics (HOA) provides a way to capture a sound field with a multi-capsule microphone. Manipulating and rendering of such signals requires a simple matrix operation, which will not be discussed in detail in this publication. In addition to channels and objects, also HOA content can be carried in MPEG-H 3D Audio.

4.1.2. Flexibility with regard to reproduction

For audio production and monitoring, the setup of loudspeakers is well defined and established in practice for stereo and 5.1. However, in consumer homes, loudspeaker setups are typically "unconventional" in terms of non-ideal placement and differ regarding the number of speakers. Within MPEG-H 3D Audio, flexible rendering to different speaker layouts is implemented by a format converter that adapts the content format to the actual real-world speaker setup available on the playback side to provide an optimum user experience under the given user conditions. 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 play back audio. Therefore, a binaural rendering component was included in the

MPEG-H 3D audio decoder for dedicated rendering on headphones with the aim of conveying the spatial impression of immersive audio production also on headphones.

Figure 3 shows an overview of an MPEG-H 3D Audio decoder, illustrating all major building blocks of the system:

- As a first step, all transmitted audio signals, be they channels, objects or HOA components, are decoded by an extended USAC stage (USAC-3D).
- Channel signals are mapped to the target reproduction loudspeaker setup using a format converter.
- Object signals are rendered to the target reproduction loudspeaker setup by the object renderer using the associated object metadata.
- Alternatively, signals coded via an extended Spatial Audio Object Coding (SAOC-3D), i.e. parametrically coded channel signals and audio objects, are rendered to the target reproduction loudspeaker setup using the associated metadata.
- Higher Order Ambisonics content is rendered to the target reproduction loudspeaker setup using the associated HOA metadata.

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

4.2. USAC-3D Core Coder and Extensions

The MPEG-H 3D Audio codec architecture is built upon a perceptual codec for compression of the different input signal classes, based on MPEG Unified Speech and Audio Coding (USAC) [24]. USAC is the state-of-the-art MPEG codec for compression of mono to multi-channel audio signals at rates of 8 kbit/s per channel and higher. For the new requirements that arose in the context of 3D audio, this technology has been extended by tools that especially exploit the perceptual effects of 3D reproduction and thereby further enhance the coding efficiency.

The most prominent enhancements are:

- A Quad Channel Element that jointly codes a quadruple of input channels. In a 3D context, inter-channel redundancies and irrelevancies can be exploited in both horizontal and vertical directions. Parametric coding of vertically aligned channel pairs can be carried out while binaural unmasking effects [27] can be avoided in the horizontal plane.
- An enhanced noise filling is provided through Intelligent Gap Filling (IGF). IGF is a tool that parametrically restores portions of the transmitted spectrum using suitable information from spectral tiles that are adjacent in frequency and time. The assignment and the processing of these spectral tiles is controlled by the encoder based on an input signal analysis. Hereby, spectral gaps can be filled with spectral coefficients that perceptually have a better match than pseudo random noise sequences of conventional noise filling would provide.

Apart from these enhancements in coding efficiency, the USAC-3D core is equipped with new signaling mechanisms for 3D content/loudspeaker layouts and for the type of signals in the compressed stream (audio channel vs. audio object vs. HOA signal).

Another new aspect in the design of the compressed audio payload is an improved behavior for instantaneous rate switching or fast cue-in as it appears in the context of MPEG Dynamic Adaptive Streaming (DASH) [28]. For this purpose, so-called ‘immediate playout frames’ have been added to the syntax that enable gapless transitions from one stream to the other. This is particularly advantageous for adaptive streaming over IP networks.

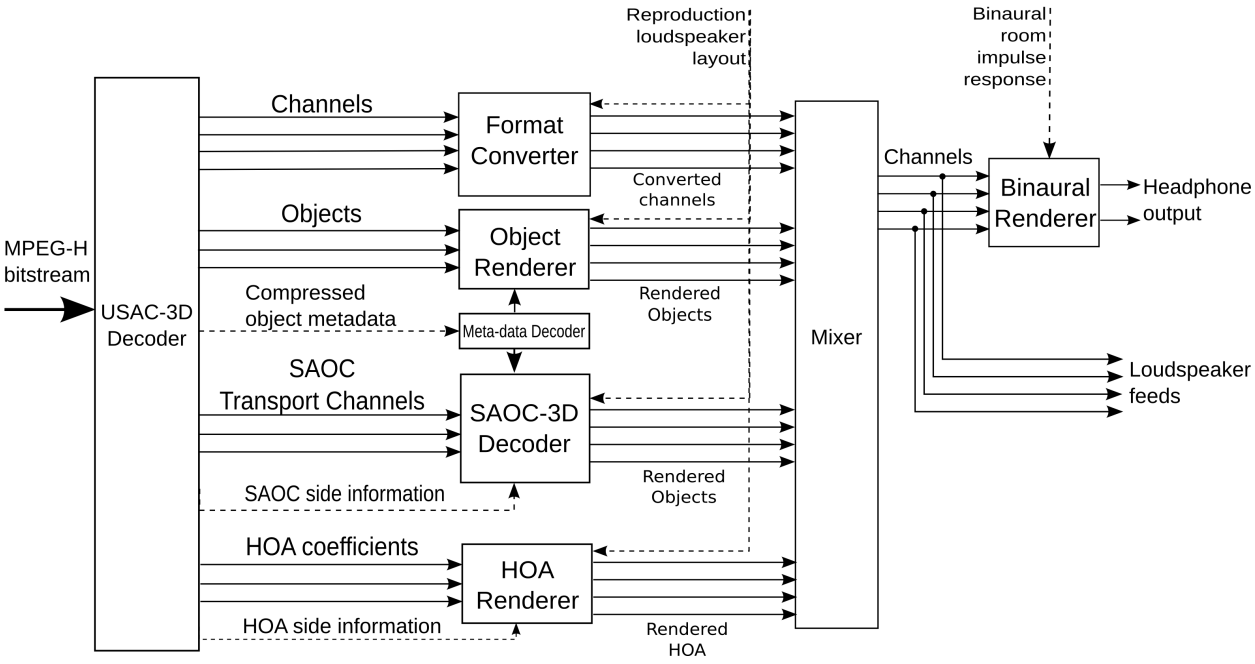

Figure 3. Top level block diagram of MPEG-H 3D Audio decoder

4.3. CO Decoding and Rendering

4.3.1. Format converter

The MPEG-H decoder comprises a so called "format converter" module that converts the decoded raw channel signals to numerous output formats, i.e. for rendering on different loudspeaker setups. This processing block renders high-quality downmixes, for example, when playing back a 22.2 channel program on a 5.1 surround reproduction loudspeaker setup. To produce high output signal quality, the format converter in MPEG-H 3D Audio provides the following features:

- Automatic generation of optimized downmix matrices, taking into account non-standard loudspeaker positions.
- Support for optionally transmitted downmix matrices to preserve the artistic intent of a producer or broadcaster.
- Application of equalizer filters for timbre preservation.
- Advanced active downmix algorithm to avoid downmixing artefacts.

Within the format converter module, there are two major building blocks, i.e. a rules-based initialization block and the active downmix algorithm. Both are described in the following.

Format converter initialization

The first sub-module derives optimized downmix coefficients mapping the channel configuration of the format converter input to the output loudspeaker layout. During the initialization, the system iterates through a set of tuned mapping rules for each input channel. Each rule defines the rendering of one input channel to one or more output channels, potentially complemented by an equalizer curve that is to be applied if the particular mapping rule has been selected. The iteration is terminated at the first rule for which the required output channels are available in the reproduction setup, thus selecting the particular mapping rule. Since the mapping rules have been ordered according to the anticipated mapping quality during the definition of the rules, this process results in selection of the highest-quality

mapping to the loudspeaker channels that are available in the reproduction setup.

The rules have been designed individually for each potential input channel incorporating expert knowledge, e.g to avoid excessive use of phantom sources when rendering to the available target loudspeakers. Thus the rules-based generation of downmix coefficient allows for a flexible system that can adapt to different input/output configurations, while at the same time ensuring a high output signal quality by making use of the expert knowledge contained in the mapping rules. Note that the initialization algorithm compensates for non-standard loudspeaker positions of the reproduction setup, aiming at the best reproduction quality even for asymmetric loudspeaker setups.

Active downmix algorithm

Once the downmix coefficients have been derived, they are applied to the input signals in the actual downmix process. MPEG-H 3D Audio uses an advanced active downmix algorithm to avoid downmix artefacts like signal cancellations or comb-filtering that can occur when combining (partially) correlated input signals in a passive downmix, i.e. when linearly combining the input signals, weighted with static gains. Note that high signal correlations between 3D audio signals are quite common in practice since a large portion of 3D content is typically derived from 2D legacy content (or 3D content with smaller loudspeaker setups) e.g. by filling the additional 3D channels with delayed and filtered copies of the original signals.

The active downmix in the MPEG-H 3D Audio decoder adapts to the input signals in two ways to avoid the issues outlined above for passive downmix algorithms: Firstly, it measures the correlation properties between input channels that are subsequently combined in the downmix process and aligns the phases of individual input channels if necessary. Secondly, it applies a frequency dependent energy-normalization to the downmix gains that preserves the energy of the input signals that have been weighted by the downmix coefficients. The active downmix algorithm is designed such that it leaves uncorrelated input signals untouched, thus eliminating the artefacts that occur in passive downmixes with only minimum signal adjustments.

4.3.2. Object renderer

In MPEG-H 3D, transmitted metadata allows for rendering audio objects into predefined spatial positions. Time-varying position data enables the rendering of objects on arbitrary trajectories. Additionally, time-varying gains can be signaled individually for each audio object. An overview of MPEG-H audio metadata is provided in [32].

The object renderer applies Vector Base Amplitude Panning (VBAP, [29]) to render the transmitted audio objects to the given output channel configuration. As input the renderer expects

- Geometry data of the target rendering setup.
- One decoded audio stream per transmitted audio object.
- Decoded object metadata associated with the transmitted objects, e.g. time-varying position data and gains.

As presented in the following, VBAP relies on a triangulation of the 3D surface surrounding the listener. The MPEG-H 3D Audio object renderer thus provides an automatic triangulation algorithm for arbitrary target configurations. Since not all target loudspeaker setups are complete 3D setups, e.g. most setups lack loudspeakers below the horizontal plane, the triangulation introduces imaginary loudspeakers to provide complete 3D triangle meshes for any setup to the VBAP algorithm.

The MPEG-H 3D Audio object rendering algorithm performs the following steps to render the transmitted audio objects to the selected target setup:

- Search for the triangle the current object position falls into.
- Build a vector base $L = [l_1, l_2, l_3]$ out of the three unit vectors pointing towards the vertices of the selected loudspeaker triangle.
- Compute the panning gains vector $G = [g_1, g_2, g_3]^T$ for transmitted object position P according to $P = L*G \rightarrow G = L^{-1}* P$
- Normalize G to preserve energy and apply the transmitted object gain g:

- $G_{norm} = gG / sqrt(g_1^2+g_2^2+g_3^2)$
- Linearly interpolate between the current panning gains and the gains computed from the object metadata received for the previous time stamp..
- Compute output signals by mixing the input signals through application of the interpolated gains into the output channels.
- Add output signal contributions of all rendered objects.

4.3.3. SAOC-3D decoding and rendering

In order to serve as a technology component for 3D audio coding, the original Spatial Audio Object Coding (SAOC) codec [22, 23] has been enhanced into SAOC-3D with the following extensions:

- While SAOC supports only up to two downmix channels, SAOC-3D supports more (in principle an arbitrary number of) downmix channels.
- While rendering to multi-channel output has been possible with SAOC only by using MPEG Surround (MPS) as a rendering engine, SAOC-3D performs direct decoding/rendering to multichannel/3D output with arbitrary output speaker setups. This includes a revised approach towards decorrelation of output signals.
- Some SAOC tools that have been found unnecessary within the MPEG-H 3D Audio system have been excluded. As an example, residual coding has not been retained, since carriage of signals with very high quality can already be achieved through encoding them as discrete channel or object signals.

4.4. HOA Decoding and Rendering

Higher order ambisonics (HOA) builds on the idea of a field based representation of an audio scene. More mathematically stated, it is based on a truncated expansion of the wave field into spherical harmonics, which determines the acoustic wave field quantities within a certain source free region around the listener's position up to an upper frequency limit beyond which spatial aliasing occurs. 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 transmitting the HOA coefficients directly in a bitstream representation, MPEG-H 3D Audio applies a two-stage coding process to the HOA data to improve the coding performance of the system, namely spatial coding of the HOA components and multichannel perceptual coding. These two stages have to be reverted in the MPEG-H 3D Audio decoder in reverse order as shown below in Section 4.4.3

The spatial coding block for the HOA representation applies two basic principles: decomposition of the input field and decorrelation of the signals prior to transmission in the core coder, both of which are described in the following.

4.4.1. Decomposition of the sound field in encoder

In the HOA encoder the sound field determined by the HOA coefficients is decomposed into predominant and ambient sound components. At the same time, parametric side-information is generated that signals the time-varying activity of the different sound-field components to the decoder.

Predominant components mainly contain directional sounds and are coded as plane wave contributions that travel through the wave field of interest in a certain direction. The number of predominant components can vary over time as well as their direction. They are transmitted as audio streams together with the associated time-variant parametric information (direction of the directional components, activity of the directional components in the field).

The remaining part of the HOA input, which has not been captured by the predominant component, is the ambient component of the sound field to code. It mostly contains non-directional sound components. Details of the spatial properties of this part of the field are considered less important. Therefore the spatial resolution of the ambient component is typically reduced by limiting the HOA order to improve the coding efficiency.

4.4.2. Encoder signal component decorrelation

The predominant sound components are represented as plane wave signals with associated directions. Thus sound events emanating from uncorrelated sound sources in different directions lead to uncorrelated audio streams to transmit.

However, the HOA representation of the ambient component may exhibit high correlations between the HOA coefficients. This can lead to undesired spatial unmasking of the coding noise since the quantization noise introduced by the perceptual coder is uncorrelated between the coder channels, thus resulting in different spatial properties of the desired signal and the quantization noise during reproduction. The HOA representation is therefore decorrelated by transforming it into a different spatial domain to avoid the spatial unmasking of the coding noise. Note that this spatial decorrelation step and its inverse operation in the decoder is equivalent to the mid-side coding principle applied to stereo coding of correlated signals, e.g. when coding a phantom source using a stereo audio coder.

4.4.3. MPEG-H 3D Audio decoder HOA rendering

The MPEG-H 3D Audio decoder transmitted HOA content is first decoded into a HOA representation by the following processing steps:

- Multichannel USAC 3D core decoding.
- Inverse decorrelation of ambient sound, i.e. transformation from decorrelated representation to a HOA coefficients representation.
- Synthesis of a HOA coefficients representation of the predominant sound components.
- HOA composition (superposition of HOA representations of predominant and ambient components).

In a subsequent processing step the composed HOA representation is rendered to the target loudspeaker configuration using a generic HOA renderer. The HOA rendering itself consists of a simple matrix multiplication of the multichannel HOA representation and a rendering matrix.

The HOA rendering matrix has to be generated at the time of initialization or when the HOA order or the reproduction setup change. It is a matrix that mixes the contribution of each HOA component to the available loudspeakers using mixing gains that result in the best field approximation of that HOA component in a region around the listener. One main design characteristic of the HOA rendering matrix is the energy preservation. This describes the characteristics that the HOA signal's loudness is preserved independent of the speaker setup and that constant amplitude spatial sweeps can be perceived equally loud after rendering.

4.5. Loudness and Dynamic Range Processing

4.5.1. Loudness normalization

One of the essential features for a next generation audio delivery is proper loudness signaling and normalization. Within MPEG-H 3D Audio, comprehensive loudness related measures according to ITU-R BS.1770-3 [30] or EBU R128 [31] 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. 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.

4.5.2. Dynamic range control

Looking at different target playback devices and listening environments, the control of the dynamic range is vital. 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 improved clipping prevention and peak limiting.

5. PERFORMANCE EVALUATION

For MPEG-H 3D Audio, several candidate technologies have undergone rigorous testing to select the best coding and rendering system for immersive audio. 24

test items were chosen to represent typical and critical audio material. During the performance evaluation more than 40000 answers from a total of 10 test labs were collected. Four main test cases were defined to characterize the system in different operation points:

- Test 1: Rendering to 9.0 - 22.2 loudspeakers
  Objective: Demonstrate very high quality for reproduction on large reproduction setups. Three bit rates: 1.2 Mbit/s, 512 kbit/s, 256 kbit/s

- Test 2: Listening at four "off sweet spot" positions
  Objective: Verify results from Test 1 for non-optimum listener positions. Bit rate 512 kbit/s

- Test 3: Binaural rendering to headphones
  Objective: Demonstrate ability for convincing headphone rendering. Bit rate 512 kbit/s

- Test 4: Rendering to alternative speaker configurations
  Objective: Demonstrate ability to perform high-quality rendering to smaller and non-standard reproduction setups: 5.1, 8.1 and, with two loudspeaker setups that were *randomly selected subsets* of the 22.2 setup, one with 5 and one with 10 loudspeakers ('Random 5', 'Random 10'). Bit rate 512 kbit/s

All tests were carried out with the MUSHRA test methodology in high quality listening rooms and used the original format, i.e. 9.0, 11.1, 14.0 or 22.2 signal as their reference. There were 12 test items for each of the two input categories CO and HOA.

After the evaluation of all test results, the system submitted by Fraunhofer IIS was selected as the reference model for MPEG-H 3D Audio CO and the Technicolor/Orange Labs submission for HOA processing since these systems performed better or equal than submitted competing systems. The pooled results for the selected RM system for 'channels and objects' are shown in Figure 4.

As can be seen from the above test results,

- bitrate was the major factor that determined the achieved subjective audio quality. At 1.2 Mbit/s and 512 kbit/s, the reference model technology delivered, on average, excellent quality, and produced good sound quality at a bitrate of 256 kbit/s.

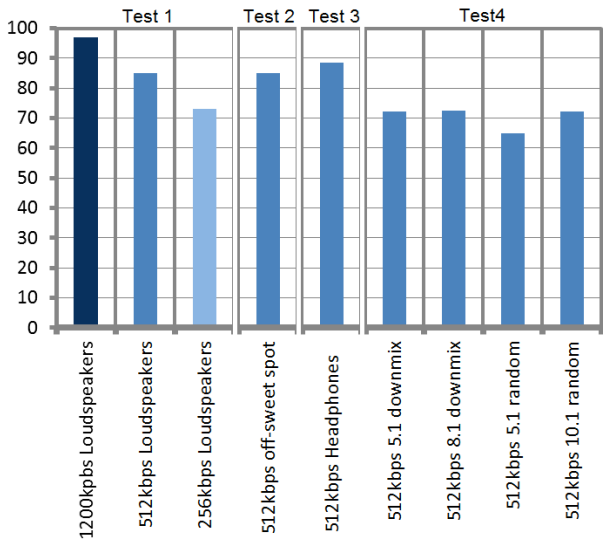

Figure 4: Summary of Reference Model listening test results for channel and objects content; the total mean MUSHRA score of each test is shown. Confidence intervals were smaller than 2.5 points in every case. Note that the results are obtained in separate tests. Same shading indicates same bitrates.

- 'off sweet spot' listening test did not reveal any additional problematic effects which would degrade sound quality.

- Test 3 showed adequate binaural quality at 512 kbit/s without undue degradation due to coding/decoding or simplified/optimized binaural processing.

6. FURTHER EVOLUTION AND OUTLOOK

The MPEG-H 3D standardization time-line is designed to consolidate technology by July 2014. Until then, further developments were under discussion in the MPEG audio group.

One activity was the merge of the CO codec and the HOA codec (both systems had originally been developed as separate architectures in response to the Call for Proposals). The architecture of the merged system, as depicted in Figure 3, has been defined and implemented between the beginning of February and July 2014.

An independent standardization timeline has been defined for a so-called "Phase 2" of MPEG-H 3D Audio. The associated part of the Call for Proposals asked for technology proposals to extend the operation range of the 3D Audio codec to even lower rates. Specifically, proponents were asked to submit coded material at bitrates of 48, 64, 96 and 128 kbit/s for 22.2 channels (or a full HOA encoding) by May 2014. A selection of technology for a Phase 2 reference model was made at the 109th MPEG meeting in July 2014. For the CO input category, the winning technology was provided by Fraunhofer IIS and was based on Phase 1 technology together with an MPEG Surround extension for the lowest bitrates. For the HOA input category, a merge of the systems of Technicolor and Qualcomm is performed. Finally, there is opportunity for collaborative further improvement.

Due to the increasing interest in MPEG-H 3D Audio in broadcast application standards like ATSC and DVB, the timeline of Version 1 is designed such that the specification is expected to be International Standard by February 2015.

7. CONCLUSIONS

In order to facilitate high-quality bitrate-efficient distribution and flexible reproduction of 3D sound, the MPEG standardization group recently started the development effort of MPEG-H Audio Coding which allows for the universal carriage of encoded 3D sound from channel-based, object-based and HOA-based sound formats. Reproduction is supported for many output setups ranging from 22.2 and beyond down to 5.1, stereo and binaural reproduction. Depending on the available output setup, the encoded material is rendered to yield highest spatial audio quality, thus overcoming the incompatibility between various 3D (re)production formats. Moreover, MPEG-H Audio is a unified system for carriage of channel-oriented, object-oriented and Higher Order Ambisonics based high quality content. This paper described the current status of the standardization project and provided an overview of the system architecture, its technology, capabilities and current performance. Further improvements and extensions, such as the ability to operate at very low data rates, or the integration into transport systems are on the way.

8. REFERENCES

[1] Alexander, R.: The Inventor of Stereo: The Life and Works of Alan Dower Blumlein. Focal Press, 2000, ISBN 978-0240516288.

[2] Blumlein, A.D.: Improvements in and relating to sound-transmission, sound-recording and sound reproducing systems. 1931, British Patent 394 325.

[3] ITU-R, Recommendation-BS.775-2, Multichannel stereophonic sound system with and without accompanying picture. 2006, Intern. Telecom Union, Geneva, Suisse.

[4] Rumsey, F., Spatial Audio. 2001, Focal Press, Oxford. ISBN 0 240 51623 0.

[5] Silzle, A. and Bachmann, T.: How to Find Future Audio Formats? VDT-Symposium, 2009, Hohenkammer, Germany.

[6] Gerzon, M.A., Perophony: With-Height Sound Reproduction. J. Audio Eng. Soc., 1973. Issue 21(1): p. 3-10.

[7] Ahrens, J., Analytic Methods of Sound Field Synthesis. T-Labs Series in Telecommunication Services. 2012, Springer, Berlin, Heidelberg. ISBN 978-3-642-25742-1.

[8] Chabanne, C., McCallus, M., Robinson, C., Tsingos, N.: Surround Sound with Height in Games Using Dolby Pro Logic IIz, 129th AES Convention, Paper Number 8248, San Francisco, CA, USA, November 2010.

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

[10] Hamasaki, K., Matsui, K., Sawaya, I., and Okubo, H.: 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.

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

[12] Hiyama, K., Komiyama, S., and Hamasaki, K.: The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field. 113th AES convention. 2002. Los Angeles, USA.

[13] Hamasaki, K., et al.: Effectiveness of Height Information for Reproducing Presence and Reality in Multichannel Audio System. 120th AES Convention. 2006. Paris, France.

[14] Kim, S., Lee, Y.W., and Pulkki, V.: New 10.2-channel Vertical Surround System (10.2-VSS); Comparison study of perceived audio quality in various multichannel sound systems with height loudspeakers. 129th AES Convention. 2010. San Francisco, USA.

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

[16] Bosi, M., Brandenburg, K., Quackenbush, S.: ISO/IEC MPEG-2 Advanced Audio Coding, Journal of the AES, Vol. 45/10, October 1997; pp. 789-814.

[17] ISO/IEC JTC1/SC29/WG11 MPEG, International Standard ISO/IEC 13818-7, Generic Coding of Moving Pictures and Associated Audio: Advanced Audio Coding, 1997.

[18] Herre, J., Dietz, M.: Standards in a Nutshell: MPEG-4 High-Efficiency AAC Coding, IEEE Signal Processing Magazine, Vol. 25, Iss. 3, 2008; pp. 137-142.

[19] EBU Evaluations of Multichannel Audio Codecs. EBU-Tech. 3324. Geneva, September 2007, available at https://tech.ebu.ch/docs/tech/tech3324.pdf.

[20] Hilpert, J., Disch, S.: Standards in a Nutshell: The MPEG Surround Audio Coding Standard, IEEE Signal Processing Magazine, Vol. 26, Iss. 1, 2009; pp. 148-152.

[21] ISO/IEC 23003-1:2007, MPEG-D (MPEG audio technologies), Part 1: MPEG Surround, 2007.

[22] Herre, J., Purnhagen, H., Koppens, J., Hellmuth, O., Engdegård, J., Hilpert, J., Villemoes, L., Terentiv L., Falch, C., Hölzer, A., Valero, M.L., Resch, B., Mundt, H., and Oh, H.: 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.

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

[24] Neuendorf, M.; Multrus, M.; Rettelbach, N. 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.

[25] ISO/IEC 23003-1:2012, MPEG-D (MPEG audio technologies), Part 3: Unified Speech and Audio Coding, 2012.

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

[27] Blauert, J. Spatial hearing: The psychophysics of human sound localization, revised edition; MIT Press, 1997.

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

[29] Pulkki, V.: Virtual sound source positioning using vector base amplitude panning. Journal of the Audio Engineering Society, Volume 45, Issue6, June 1997; pp. 456-466.

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

[31] European Broadcasting Union (EBU), Recommendation R128. Lautheitsaussteuerung, Normalisierung und zulässiger Maximalpegel von Audiosignalen, 2011, Geneva, Suisse.

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