MPEG-H 3D Audio—The New Standard for Coding of Immersive Spatial Audio

Jürgen Herre, *Senior Member, IEEE*, Johannes Hilpert, Achim Kuntz, and Jan Plogsties

***Abstract*—The science and art of Spatial Audio is concerned with the capture, production, transmission, and reproduction of an immersive sound experience. Recently, a new generation of spatial audio technology has been introduced that employs elevated and lowered loudspeakers and thus surpasses previous 'surround sound' technology without such speakers in terms of listener immersion and potential for spatial realism. In this context, the ISO/MPEG standardization group has started the MPEG-H 3D Audio development effort to facilitate high-quality bitrate-efficient production, transmission and reproduction of such immersive audio material. The underlying format is designed to provide universal means for carriage of channel-based, object-based and Higher Order Ambisonics based input. High quality reproduction is provided 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 the incompatibility between various 3D formats. This paper provides an overview of the MPEG-H 3D Audio project and technology and an assessment of the system capabilities and performance.**

***Index Terms*—Audio coding, audio compression, channel-based, higher order Ambisonics, HOA, MPEG, MPEG-H, object-based, SAOC, spatial audio, USAC.**

I. Introduction

SPATIAL AUDIO denotes the attempt to capture the salient parts of a sound field and reproduce it in some form at other, possible distant places (and times), such that a human listener perceives the spatial characteristics of the original sound scene to a large extent during reproduction. Spatial realism and even immersion are two important goals in this area of research that has been active for many decades, starting with two channel stereo as introduced by Blumlein in 1931 [1], [2] . Later significant extensions to the theme of reproducing spatial audio by an increasing number of loudspeakers include 'surround sound' [3], [4], [5], Ambisonics [6] and wave field synthesis [7]. The latter two aim to deliver an approximation of the original wave field at the point of the listener and inside a certain area, respectively. In order to provide full realism and immersion to the listener, a new generation of technology has started more recently to consider the height dimension by adding elevated (and, sometimes, lower) speakers to the overall setup. Some typical examples of such so-called '3D' loudspeaker setups include 7.1 with two height channels [8], 9.1 [9] and 22.2 [10]. While '3D' loudspeaker setups have been shown to deliver spatial quality surpassing that of traditional '2D' setups [11]–[14], there is currently no agreed-upon 'common denominator' among the multitude of possible loudspeaker setups which could provide interoperability between content producers, equipment manufacturers and consumers in the same way the 5.1 setup has served as the common denominator for surround sound so far.

This paper introduces a new standard for universal and efficient coded representation and rendering of high-quality spatial audio which has been developed recently by the ISO/MPEG standardization group, i.e., the MPEG-H 3D Audio standard [15]. Besides being able to deliver a high amount of immersion, the standard has the potential to unify the plethora of '3D' Audio formats by accepting virtually all known signal formats as input and offering high-quality reproduction for many output formats ranging from 22.2 and beyond down to 5.1, stereo and binaural reproduction—independently of the original encoding format. This unique capability indeed is the potential for overcoming the current incompatibility between various 3D formats.

The paper expands on contributions in [16] and is structured as follows: Section II discusses some important concepts for representation of 3D sound. Section III introduces the architecture of an MPEG-H 3D Audio system and elaborates its components. Section IV focuses on standardization issues of the new technology. Finally, Sections V and VI show the results of a first evaluation of the system performance and conclude the paper.

II. Concepts for 3D Sound Representation

In order to represent immersive spatial sound, a number of concepts have emerged over the years. This section discusses some basic concepts and their properties and puts them into context.

A. Channels

The most well-known (and most common) way of representing spatial sound is to deliver a set of waveforms, called *channel signals*, where each signal is designated to feed a loudspeaker in a known prescribed position relative to the listener position.

Spatial audio material today is mostly produced as channel-based material, where 2.0 and 5.1 are the most

Manuscript received July 30, 2014; revised December 01, 2014; accepted February 26, 2015. Date of publication March 09, 2015; date of current version July 14, 2015. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Lauri Savioja.

J. Herre is with the International Audio Laboratories Erlangen (a joint institution of Friedrich-Alexander University of Erlangen-Nürnberg and Fraunhofer IIS), 91058 Erlangen, Germany, and also with the Fraunhofer IIS, 91058 Erlangen, Germany (e-mail: juergen.herre@audiolabs-erlangen.de).

J. Hilpert and J. Plogsties are with the Fraunhofer IIS, 91058 Erlangen, Germany (e-mail: johannes.hilpert@iis.fraunhofer.de; jan.plogsties@iis.fraunhofer.de).

A. Kuntz is with the International Audio Laboratories Erlangen, 91058 Erlangen, Germany (e-mail: achim.kuntz@iis.fraunhofer.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2015.2411578

1932-4553 © 2015 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

common loudspeaker configurations for two-channel stereo and 'multi-channel' sound, respectively (the number before the period denotes the number of full-range speakers, the number following the period denotes the number of low frequency enhancement channel speakers). Over time, and due to recommendation by ITU-R [3], 5.1 has become the common denominator for surround sound (where all speakers are positioned within a horizontal plane around the listener). In contrast, typical '3D' formats/loudspeaker setups include a varying number of height speakers, such as 7.1 (with two height channels), 9.1 (with four height speakers) and 22.2 (with 9 height and 3 lower speakers). Currently, no single de-facto standard has emerged for general use among the '3D' formats including height.

Sound production (microphone setup and mixing techniques) for channel-based content have been studied for a long time and are well established. Thus, however, sound is produced specifically for a certain reproduction loudspeaker setup, i.e., the produced content is tied to one specific loudspeaker configuration. Trying to reproduce content on a different loudspeaker setup requires additional steps and may result in degraded quality. Specifically, reproducing channel-based content on loudspeaker setups with a lower number of loudspeakers requires *downmixing*, while reproduction on setups with higher number of loudspeaker can be achieved by *upmixing*. In both cases, quality loss may occur due to the conversion to a different setup. As an example, downmixing of signals may cause coloration/comb filtering for coherent signal components. Thus, carefully designed down-/upmixing algorithms have to be employed to achieve high-quality reproduction on arbitrary loudspeaker setups [31], [44]–[46].

In summary, the concept of channel-based spatial audio reproduction is today well-established and finds its limits in the incompatibilities between different formats which is especially salient for the highly immersive '3D' setups. An ideal reproduction system would be able to remove these dependencies.

B. Objects

Alternatively, a spatial sound scene can be described by a number of virtual sources (so-called sound *objects*), each positioned at a certain target object position in space. In contrast to the concept of channels, these object positions can be

a) totally independent from the locations of available loudspeakers, and
b) varying over time for modeling of moving objects, such as a plane flying by over the head of the listener.

Since object positions do not necessarily coincide with loudspeaker positions, object signals generally need to be rendered to their target positions by appropriate rendering algorithms, see, e.g., Vector Base Amplitude Panning, VBAP [17] as a popular example that performs such a rendering for 2D/3D. From a data point of view, objects consist of object waveform(s) plus associated metadata (object position, gain, etc.). Conversely, for the purpose of coding, channels can be seen as objects that are placed in fixed positions in which loudspeakers reside, such that the associated rendering is simply reproduction by this loudspeaker (note: Independently of the loudspeaker positions, so-called *static objects* can be used to provide sound from certain directions).

In terms of production, a huge amount of object-based content is already present in today's studios as multi-track content containing recorded individual musical instruments, talkers, foley sound effects etc. for which spatial positions are assigned by *panning* tools (i.e., distributing the signal among several nearby loudspeakers, see, e.g., [17]). Nonetheless, in the majority of all cases, the delivery of the final mix is today still done in stereo or multi-channel formats, i.e., in a conventional channel-oriented way rather than in an object-oriented fashion. Converting a studio to delivering object-oriented output would thus include upgrading the mixing desks to deliver both object signals and their associated metadata in a commonly accepted way. A first standardized object-based representation dates back to 1999 (MPEG-4, [40]) but was certainly ahead of its time. Other more recent object-based formats can be found, e.g., in [49] or [50].

In summary, the concept of object-oriented spatial audio is agnostic of actual reproduction loudspeaker setups and thus overcomes the undesirable dependency of the content from the reproduction setup. Nonetheless, major efforts are still needed to make object-based content commonplace [42].

C. Higher Order Ambisonics (HOA)

As a third alternative, a 3D spatial sound scene can be described using *Ambisonics*, i.e., as a number of 'coefficient signals' that represent a spherical expansion of the sound field [6]. As an example, traditional 'first order' Ambisonics decomposition represents the sound field by four signals with varying directional patterns: one with an omni-directional pattern plus three with perpendicular figure-of-eight patterns. Generally, coefficient signals have no direct relationship to channels or objects and are agnostic of the reproduction loudspeaker setup. Ambisonics has a long track record of academic research and is limited in its capability of carrying a high-quality 3D audio sound field and thus was more recently extended towards *Higher Order Ambisonics (HOA)*. HOA provides more coefficient signals and thus an increased spatial selectivity, which allows to render loudspeaker signals with less crosstalk, resulting in reduced timbral artifacts. In contrast to objects, spatial information in HOA is not conveyed in explicit geometric metadata, but in the coefficient signals themselves. Thus, Ambisonics/HOA is not that well suited to allow access to individual objects in a sound scene. It can be shown that Ambisonics in its functionality is closely related to wave field synthesis [18].

In summary, Higher Order Ambisonics has the potential of providing a high-quality description of a 3D spatial sound scene which is agnostic of the reproduction loudspeaker layout. Compared to the channel-based and the object-based concepts, HOA is rather new and not yet widely supported in terms of recording/production equipment.

D. Binaural Rendering

Finally, binaural rendering of sound for headphone playback using a set of *Binaural Room Impulse Responses (BRIRs)* [19] is a valid way of representing and conveying an immersive 3D spatial audio scene to a listener. BRIRs characterize the acoustic transmission from a point in a room to the ears of a listener. They are typically measured with microphones inside the ear canal or created by means of models. The direct portion of the BRIR is referred to as head-related impulse response (HRIR) which is essential for accurate localization and varies from person to person. Headphone playback has constantly gained importance in the age of wireless mobile and multimedia-enabled personal devices. Enhanced playback quality can be achieved when using personalized BRIRs and head-tracked binaural rendering which is, however, not commonly deployed in today's consumer devices yet.

E. Metadata

The concept of metadata accompanying media content is known in production and distribution. Metadata describe properties for content search in archives or controlling the production workflow. The focus here is on audio-related metadata relevant for the playback of content, i.e., for controlling the rendering process and describing the audio content for presentation to the user.

The availability of such metadata in the playback device enables a number of new features, such as user interactivity, dynamic object rendering and adaptation of audio elements to the loudspeaker setup. Typical usage scenarios are: enabling dialog tracks for additional languages replacing the main dialog; changing the commentary level relative to the background level, selecting a preset for visually-impaired, and rotating and zooming in the sound scene.

Audio metadata can be divided into static and dynamic data. Static metadata are considered to be constant for the duration of a program. Examples are a textual description of the audio element, e.g., its dialog language, and the default on/off state of an audio element. Dynamic metadata describe information that change over time, and control the rendering process, e.g., position, gain and spread of a virtual source used in the object renderer. In this way sound can be rendered optimally to each playback scenario, e.g., to loudspeakers in a non-standard configuration or to headphones using binaural rendering as explained in Section II-D.

The metadata elements are associated with one audio element, or they are organized in groups such that they refer to multiple audio elements, e.g., all objects belonging to one sound scene. Specific metadata can be defined to control the way how an application or user interacts with the content. Such elements control, e.g., the allowed range of gain changes, or which audio elements can be played exclusively.

For more complex audio scenes controlling individual audio elements may be impractical for use cases like playback on a tablet. Presets can be defined for that purpose, i.e., predefined metadata configurations for a specific audio rendering, e.g., “Stadium live sound”, “Clear dialog” or “Home team commentary”.

In summary, audio-related metadata open up a range of possibilities to interact with the content and to control audio rendering independent of the playback scenario.

III. System Architecture

This section first briefly reviews existing ISO/MPEG Audio technology that is relevant to MPEG-H 3D Audio coding. Subsequently, the architecture of an MPEG-H 3D Audio system is provided and its components are discussed.

A. Pre-Existing ISO/MPEG Audio Coding Technology

A first commercially-used multi-channel audio coder standardized by MPEG in 1997 is MPEG-2 Advanced Audio Coding (AAC) [20], [21], delivering EBU broadcast quality at a bitrate of 320 kbit/s for a 5.1 signal. This was achieved by adding a number of advanced coding tools to the architecture of MPEG-1 audio codecs in order to provide enhanced performance for transient and tonal items as well as for the coding of several channels. Based on this waveform coder, MPEG-4 High Efficiency AAC (HE-AAC) [22] was created in 2002/2004, which combines the AAC technology with SBR (Spectral Band Replication) bandwidth extension and Parametric Stereo coding, and in this way provides full audio bandwidth also at very low data rates. Bandwidth extension saves on bitrate by omitting transmission of the input signal's high frequency content and resynthesizing it in the decoder based on compactly transmitted parametric side information. Parametric Stereo encodes two input channels by transmitting a single (sum) channel and associated side information which enables a resynthesis of the stereo image at the decoder side. For carriage of 5.1 surround sound, HE-AAC delivers quality comparable to that of AAC at a bitrate of only 160 kbit/s [23].

Subsequent ISO/MPEG standards provided generalized means for joint parametric coding of multi-channel spatial sound by mapping the input signals/objects to a downmix and associated side information. The decoder can then generate the multi-channel output scene from the transmitted downmix based on the side information. Most importantly, this side information includes the level differences between the input channel signals/objects and the coherence between them in a number of frequency bands. MPEG-D MPEG Surround (MPS, 2006) [24], [25] and MPEG-D Spatial Audio Object Coding (SAOC, 2010) [26], [27] allow for the highly efficient carriage of multi-channel sound and object signals, respectively. In contrast to MPS, SAOC allows the user to interactively change the output scene (i.e., adjust level and position of certain sound objects). Both codecs can operate at very low rates (e.g., 48 kbit/s for a 5.1 signal).

Finally, MPEG-D Unified Speech and Audio Coding (USAC, 2012) [28], [29] was developed by ISO/MPEG by combining enhanced HE-AAC coding with state-of-the-art full-band speech coding (AMR-WB+) and other improvements 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 development 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. As it will be explained in the following, MPEG-H 3D Audio draws from USAC, SAOC and MPEG Surround to achieve its high coding efficiency.

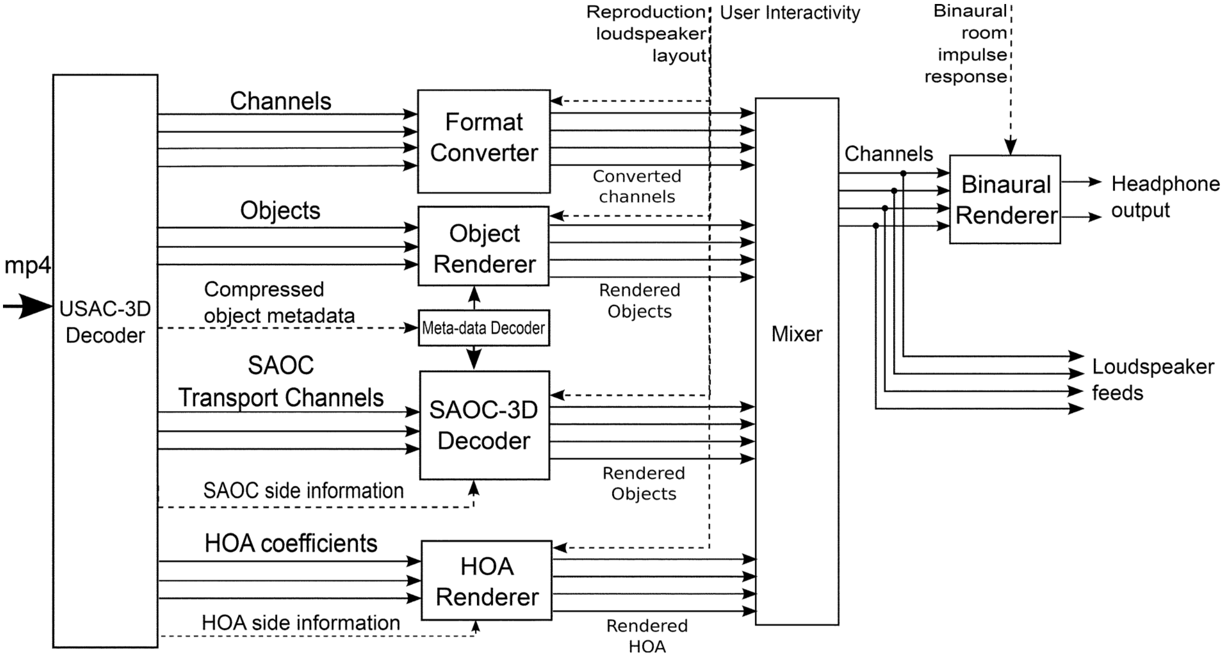

Fig. 1. Top level block diagram of MPEG-H 3D Audio decoder.

B. Overview

Fig. 1 shows a top level diagram of an MPEG-H 3D Audio decoder. The main components are a so-called USAC-3D core decoder, a set of renderers for the different signal classes and mixer.

In a first stage, the different base signals are converted from their data-compressed representation by means of a so-called USAC-3D decoder. Its compression format is explained below.

The different signal classes (waveforms for channel signals and object signals or HOA coefficient signals) are then fed to their associated renderers that map those signals to loudspeaker feeds for the particular reproduction setup that is available at the receiver side. As soon as all rendered signals are available in the reproduction format, they are combined in a mixing stage to form a loudspeaker feed. In case a binaural representation is requested, the reproduction setup is determined by the Binaural Room Impulse Response database of a binaural renderer, and the signal is converted to a virtual 3D scene for headphone reproduction. It is possible to transmit any combination of the different signal types in a single MPEG-H stream, for instance a combination of channel signals with object signals or an HOA scene with objects.

The renderers are:

- A format converter for converting channel signals from their production speaker format to the reproduction speaker layout.
- An object renderer to place static or dynamic object tracks into the reproduction layout.
- An SAOC-3D decoder for objects (or channels) that have been parametrically represented by means of a downmix of the object signals and the parametric side information. This decoder performs both tasks, i.e., parametric decoding and rendering to the target layout in a single step.
- An HOA renderer to convert from the scene based HOA representation to the actual reproduction layout.
- A binaural renderer to convert from a virtual loudspeaker layout to headphone output.
- A distance compensation module in the loudspeaker feed allows for correction of level differences as well as time-alignment of the loudspeaker signals, if the loudspeakers are set up in non-uniform distances to the center of the listening area.

In addition, playback and rendering of the different signal classes can be controlled by a user interface, if the corresponding static metadata marks these signals as enabled for interactivity.

All processing blocks will be discussed in more detail in the following subsections.

C. Core Coder

The core compression scheme for the different signal classes is an evolution of the MPEG Unified Speech and Audio Coder (USAC) [28]. For the new requirements in the context of 3D Audio, this technology has been extended into a so-called USAC-3D codec by tools that exploit the perceptual effects of 3D reproduction and thereby further enhance its coding efficiency.

A Quad Channel Element allows joint coding of a quadruple of input channel signals. This is implemented by combining different stereo coding tools from USAC. Two pairs of stereo signals are jointly coded with separate instances of the Unified Stereo tool [28], each generating a downmix and a residual output signal. In a subsequent step, both downmix signals are jointly coded with one instance of the Complex Prediction Stereo tool [28], both residual signals with a second instance of the same tool. Both joint coding tools strive to maximize decorrelation of their output signals based on inter-channel prediction, resulting in a perceptually optimum panning of quantization noise. The required side information (e.g., prediction coefficients) is compactly transmitted along with the signals. In a 3D context, inter-channel redundancies and irrelevancies can thus be exploited in both, horizontal and vertical room directions at the same time.

An Intelligent Gap Filling tool [15] parametrically restores portions of the transmitted spectral content, using suitable information from spectral tiles that are adjacent in frequency and time. Assignment and processing of these tiles is controlled by the encoder to ensure an optimum perceptual match.

Additional signaling mechanisms have been incorporated into the codec specification in order to describe the content format with 3D loudspeaker layout and to mark the different signal types for proper routing and rendering in the MPEG-H decoder.

D. Renderers

The system comprises renderers for the different spatial sound representations that can be transported in MPEG-H 3D audio bitstreams. All renderers adapt their processing to the particular reproduction loudspeaker layout that is known to the decoder device. Their processing principles are described in the following.

1) Multi-Channel Content Rendering: Multi-channel audio content is generally produced for playback over a specific reproduction loudspeaker layout, i.e., in a specific ‘format’. However, due to the large number of existing and upcoming multi-channel formats in the market, the loudspeaker layout installed with the MPEG-H decoder might deviate substantially from the production loudspeaker setup. Furthermore, loudspeakers in typical domestic listening environments are usually not set up at the ideal nominal positions defined by a standardized reproduction format, but may be rather displaced due to aesthetic considerations, limited space, mount restrictions, or simply indifference or ignorance of the user.

The system thus includes a “format converter” processing block that renders audio content transmitted in a specific multichannel format to the target format defined by the actual reproduction speaker layout. It performs an active downmix avoiding the well-known downmix artifacts like signal coloration, signal cancellations, or uncontrolled signal boost known from simple downmixes when adding (partially) correlated, yet temporally unaligned signals as, e.g., reported in [30]. The active downmix algorithm in the MPEG-H channel renderer operates in a subband domain and employs an adaptive phase-alignment of the input signals as well as an adaptive downmix normalization to preserve the input signal powers. The algorithm further exploits measures of the downmix input correlations to avoid unnecessary modifications of uncorrelated input signals. For a detailed description of the active downmix algorithm see [31].

Since any unpredictable combination of transmitted format and target format may occur, the format converter features an algorithm for the automatic generation of downmix matrices that is adapted to the current target format. This downmix matrix design procedure incorporates expert knowledge for the optimal mapping of each input channel to the available output channels. It further takes into account the actual loudspeaker positions, compensating for deviations from nominal layout geometries as, e.g., defined in [3].

Especially in broadcast applications, a producer or content provider may want to retain control over the decoder downmix process, e.g., for artistic reasons. For such use cases, the MPEG-H bitstream syntax allows to transmit downmix coefficients for multiple target setups in an efficient way, that are applied in the channel renderer instead of the decoder generated downmix gains. In addition, a bitstream element enables the enforcement of passive (i.e., not signal adaptive) downmix processing, if required, e.g., to emulate legacy downmix behavior.

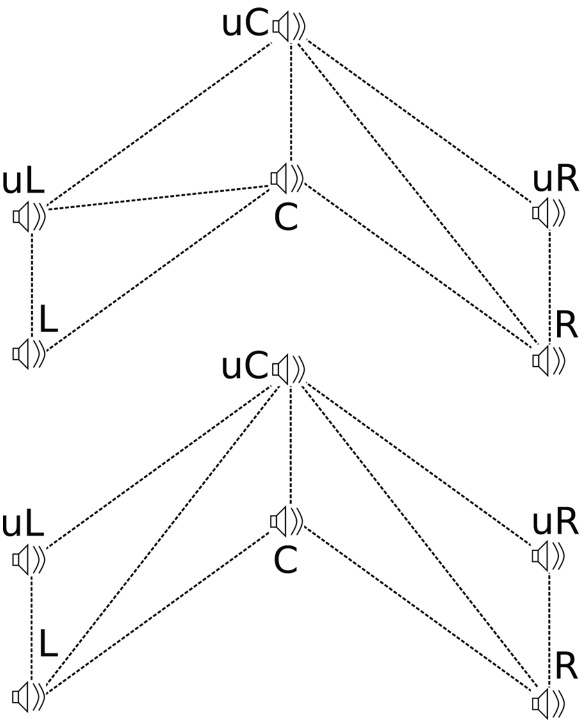

Fig. 2. Triangulation for L(eft), C(enter), R(ight) loudspeakers complemented by an upper layer of corresponding loudspeakers (uL, uC, uR). Result of the MPEG-H triangulation algorithm (lower sketch) compared to generic triangulation result (upper sketch): The MPEG-H triangulation algorithm yields a left-right and front-back symmetric division of the hull into triangles.

2) Object Content Rendering: Audio objects transmitted with potentially time-variant position metadata allow the precise positioning of sound events in spatial sound scenes. They are rendered in the MPEG-H 3D Audio decoder by a refined Vector Base Amplitude Panning (VBAP) algorithm. VBAP is built on the idea of controlling the perceived direction of a panned virtual source as a weighted linear combination of the sound contributions of up to three loudspeakers, whose positions in space span a vector base. The desired panning weights can be derived by solving a system of linear equations and normalizing the panning coefficients for energy preservation. For an extensive study on VBAP see [17].

The VBAP object renderer in MPEG-H 3D Audio is complemented by a Delaunay triangulation algorithm that provides triangle meshes adapted to the specific geometry of the reproduction loudspeaker setup [32]. Two extensions have been included over a generic triangulation and VBAP rendering to improve the perceptual rendering result, especially for arbitrary loudspeaker setups:

Firstly, the triangulation algorithm has been designed such that it yields a left-right and front-back symmetric division of the loudspeaker hull around the listener into triangle meshes, thus avoiding asymmetric rendering of symmetrically placed sound objects, see Fig. 2.

Secondly, in order to prevent uneven source movements and to avoid the need to restrict object coordinates to the regions supported by the physical loudspeaker setup, imaginary loudspeakers are added to the target setup in regions where the resulting triangle mesh would not cover the full sphere around the listener. During rendering, VBAP is applied to the loudspeaker setup which is extended by the imaginary loudspeakers, and the VBAP panning gains of the imaginary loudspeakers are downmixed to those of the physically available loudspeakers. The downmixing gains for mapping virtual to available loudspeakers are derived by distributing the virtual loudspeakers' energy equally to the neighboring loudspeakers, where the neighbors are obtained from the triangulation algorithm noted above.

One prominent use case for the added imaginary loudspeakers are reproduction layouts that only consist of loudspeakers in the horizontal plane: In this example an imaginary loudspeaker is added at the “Voice-of-God” position above the center of the listening area, resulting in smooth perceived movements, e.g., for fly-over sound objects.

MPEG-H further features a gradual spread parameter that gives the content creator an additional degree of freedom to express artistic intents. It allows to spread the energy of objects over multiple loudspeakers, thus creating the perception of audio sources with increased extent. Large spread parameters can be used to render unlocalized sound, e.g., in order to render the effect of sound objects that move through the listener. The spread algorithm in MPEG-H 3D Audio is based on Multiple Direction Amplitude Panning (MDAP) [43].

Positional metadata used for rendering of objects can change dynamically in short intervals, e.g., 2048 audio samples. As this would result in a relatively high bitrate, a data compression method is applied. For random-access support, a full transmission of the complete set of dynamic element metadata happens on a regular basis, i.e., intra-coded metadata. In between random access points, quantized differential metadata is transmitted along with a variable number of polyline points to accurately describe the geometric data [35].

3) SAOC-3D: In the context of 3D audio coding, the original Spatial Audio Object Coding (SAOC) codec [26], [27] has been enhanced into an SAOC-3D scheme which compresses and renders both channel and object signals in a very bitrate-efficient way.

SAOC-3D has been derived by incorporating the following extensions into the original scheme: Firstly, while the original SAOC only supports up to two downmix channels, SAOC-3D can map the multi-object input to an arbitrary number of downmix channels (and associated side information). Secondly, rendering to multi-channel output is done directly in contrast to classic SAOC which has been using MPEG Surround as a multi-channel output processor. Finally, some tools from the original SAOC specification were dropped, since they have been found unnecessary in the context of the MPEG-H 3D Audio system. As an example, the residual coding tool has not been retained, since carriage of channel or object signals with very high quality can be achieved through encoding them as discrete channel or object signals, i.e., without resorting to SAOC-3D.

4) HOA Coding and Rendering: HOA as a wavefield-based representation of spatial sound describes the physical properties of a sound field as a mathematically well motivated series expansion, i.e., in form of the HOA coefficients. However, this representation is not optimally applicable to the two basic principles of audio coding: redundancy reduction and irrelevance reduction. Therefore, instead of directly coding the HOA coefficient channels with a multi-channel audio codec, two preprocessing steps are applied in the MPEG-H encoder that are reverted in the decoder in opposite order.

Firstly, the sound field is decomposed into direct sound components and ambience components to reduce redundancy: Sound events emanating from a distinct direction result in highly correlated signals in a multitude of HOA components, as is obvious from the spherical harmonics expansion of plane waves [33]. In order to exploit and reduce this redundancy, the encoder performs an analysis to detect directional sounds and subtracts them from the HOA coefficients of the encoder input. They are transmitted separately as plane waves with accompanying direction metadata. In the decoder, an HOA coefficients representation is synthesized for the plane waves and added to the HOA coefficients for the ambience sound field components. Note that parametric coding of plane wave components in sound fields has also been proposed, e.g., in [47], [48].

In addition to the parametric coding of plane wave field components, MPEG-H further offers a mode for parametric coding of field components with more involved directional patterns.

Secondly, to improve the irrelevancy reduction, the HOA coefficients are transformed into virtual loudspeaker signals by means of a spherical Fourier transform. This processing step allows the multi-channel core audio codec to apply a psychoacoustic model to reduce perceptually irrelevant information during the coding process. In addition, the transformation from the HOA representation to a set of virtual loudspeaker signals in general further reduces correlations between the audio signals fed into the multichannel core audio coder, resulting in improved redundancy reduction and thus improved coding efficiency. Similarly as for the coding of multichannel audio, the number of used transport channels basically depends on the available bitrate as well as the perceptual quality demanded for the output signals of the multichannel core audio decoder.

In the MPEG-H 3D Audio decoder, the HOA sound field representation is reconstructed by synthesizing the direct sound components and transforming the virtual loudspeaker representation back to HOA coefficients. The HOA coefficients representation is then rendered to the reproduction setup using a generic HOA renderer with a rendering matrix adapted to the target loudspeaker setup geometry.

5) Binaural Renderer: The binaural rendering in MPEG-H 3D Audio is carried out as a post-processing step by efficiently converting the decoded signal into a binaural downmix signal that provides an immersive sound experience when listening over headphones.

Typically, the binaural renderer is fed with a BRIR database for rendering virtual loudspeakers in the form of FIR filter coefficients. These are parameterized such that they can be used within the binaural renderer’s signal processing blocks.

In the binaural renderer the output of the mixer is processed, see Fig. 3. The BRIR database is considered to be stored locally at the decoder side and is fed to the decoder via a dedicated interface.

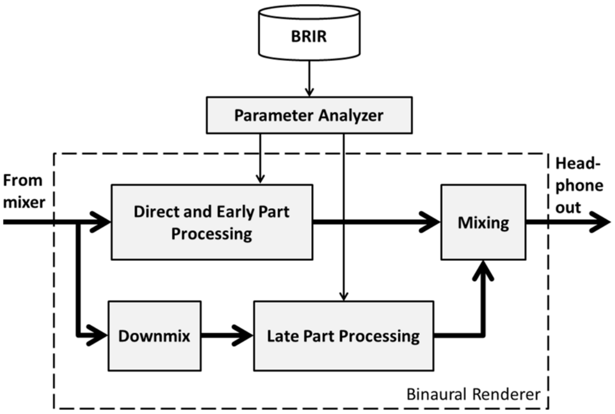

Fig. 3. Conceptual diagram of the binaural renderer.

Two binaural rendering methods are defined in the standard: Time domain (TD) and frequency-domain (FD) binaural renderer. Depending on whether the output of the mixer stage is in time or QMF domain one or the other is computationally more efficient. Both types have in common that they process the early part of a BRIR separately from the late part. The early part is convolved in FFT domain (TD binaural renderer) or QMF domain (FD binaural renderer) with the multi-channel output signal of the mixer. The late part of the BRIR is used for processing a downmixed version of the mixer output channels for complexity reduction. In case of the TD-binaural renderer, the signal is convolved with a common so called “diffuse” filter modelling the late reflections and reverberation of the BRIRs. For the FD binaural renderer, a sparse frequency domain reverberator [34] is used to process a stereo downmix of the mixer output channels to form the binaural diffuse sound. The signals of the early binaural processing and the diffuse part are combined to form the binaural output signals. It is noteworthy that in both FD and TD binaural renderers a cut-off point in time within the BRIRs is identified for each frequency band where energy is considered to be sufficiently low. No processing is done for the parts after the cut-off point. This decreases the computational complexity of the binaural rendering.

In essence, the goal of the BRIR parameterization and binaural rendering is that the signal should sound perceptually similar to a signal convolved with the BRIR database. However, the complexity of the binaural processing in MPEG-H 3D Audio is less than 20% of the straight-forward convolution method.

6) DRC/Loudness Processing: Inherited from MPEG-D Audio, MPEG-H 3D Audio provides comprehensive dynamic range control (DRC) and loudness processing functionality. It enables adaption of the decoding process to the listening situation, resulting in specific characteristics of the output signals, namely appropriate dynamic range as well as appropriate long-term loudness. The functionality is, e.g., required to allow content providers to ensure homogenous playback loudness and to conform to loudness regulations. DRC processing in the MPEG-H 3D Audio decoder facilitates a playback device to tailor the output signals' dynamic range to different listening environments. The two key components of DRC/loudness processing in the system are a flexible syntax to include various DRC metadata and loudness measures in the bitstream (see [35] for a non-exhaustive list), and the DRC/loudness processing blocks in the decoder.

Encoders can embed long-term loudness measures in the MPEG-H 3D Audio bitstream that are obtained from single-channel or multi-channel loudness models as defined in [36]. In the MPEG-H 3D Audio decoder, the deviation of the transmitted loudness measures from a desired target loudness is determined to derive loudness normalization gains that are applied to the decoded audio signals. The desired target loudness can heavily depend on the listening situation. For an elaborate treatment of loudness normalization for different listening situations see [37]. Since a high desired target loudness may result in high loudness normalization gains in the decoder, clipping could occur during the loudness normalization. Thus, the MPEG-H 3D Audio decoder may apply a dynamic range compression (see, e.g., [38] for introduction) prior to the application of the normalization gain to reduce the peak values of the audio signal.

The encoder guided DRC functionality enables the adaption of the MPEG-H 3D Audio decoder to various listening conditions, characterized, e.g., by different background noise levels, maximum desired or allowed peak sound pressure levels, and the resulting dynamic range, that can be reproduced in a listening environment depending on the reproduction device.

Multiple DRC gain sequences can be transmitted, tailored either to a complete sound scene, or to individual elements/element groups of the audio scene. The first option enables the transmission of DRC data at especially low bitrate costs, whereas the second option allows for individual dynamics processing, e.g., to realize ducking functionality for voice-over applications, speech intelligibility improvements in adverse listening environments, or the like. DRC sequences can be transmitted as full-band gains or, if the transmission bit-budget permits, as frequency dependent multi-band gains.

One specific DRC processing block at the end of the processing chain implements a guided clipping prevention gain stage, saving the computational complexity of an otherwise potentially required peak-limiter, thereby reducing the power requirements, e.g., on mobile devices.

7) Interactivity Features: The ability to carry and interpret content-related metadata enables efficient transmission of sound scenes that can be interactively manipulated at the receiving end.

Interactivity, in particular signal volume and panning changes, can be assigned to audio objects, channel beds and HOA scenes. Besides parameters for the amount and type of interactivity, descriptive metadata carries information about the content, e.g., the language, of a signal. This static metadata is only transmitted once at the beginning of an audio file or, e.g., at random-access points. The bitstream syntax is designed in a bit-efficient manner.

A large variety of use cases for interactivity can be addressed, like

- personalized selection of commentary version and language
- transmission and optional reproduction of optional audio tracks, e.g., audio description, team radio in car races
- adjustment and balance of dialog vs. background level

A discussion of these use cases is given in [39]. A complete description of all metadata elements and their processing can be found in [35].

This kind of interactivity can not only enhance a personalized user experience but also provide significant alleviation features for the visual and hearing impaired.

IV. Standardization

Starting in early 2011, initial discussions on ‘3D Audio’ at ISO/MPEG were stimulated by investigations 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, 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 [40]. At the 105th MPEG meeting in July/August 2013, a Reference Model technology was selected from the received submissions (4 for CO, i.e., ‘channel and object content’ and 3 for HOA content) based on their technical merits to serve as the baseline for further collaborative technical refinement of the specification. Specifically, the winning technologies came from Fraunhofer IIS (channel and objects part) and Technicolor/Orange Labs (HOA part). In a next step, both parts were subsequently merged into a single harmonized system.

The MPEG-H 3D Audio standardization time-line was designed to consolidate technology in July 2014. Until then, further refinements to the system were discussed in the MPEG audio group and implemented. Due to the increasing interest in MPEG-H 3D Audio in broadcast application standards like ATSC and DVB, the further development timeline of the specification foresees a status of International Standard by February 2015.

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 operating 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 April/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. Eventually, there is opportunity for collaborative further improvement.

V. Performance Evaluation

Seven submissions were received by MPEG for the call for proposals for MPEG-H 3D, four for CO and three for HOA. Proponents were to deliver encoded representations of the test signals (bitstreams) and decoded waveforms for each test.

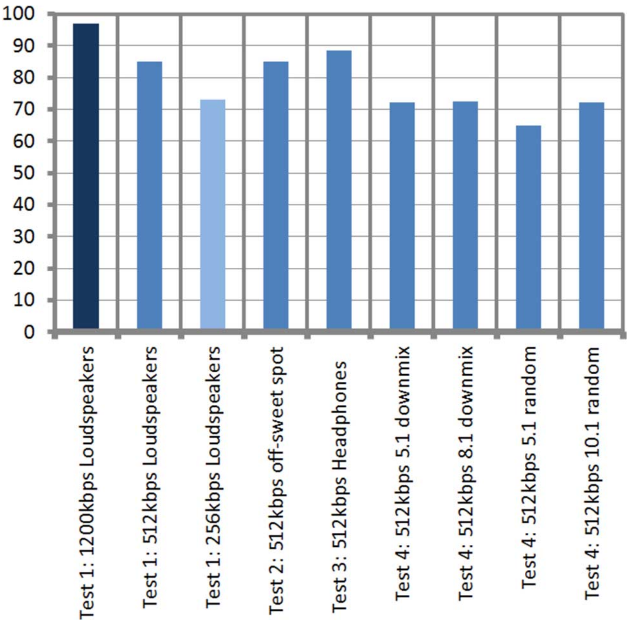

Fig. 4. Overview of the listening test results for the Reference Model of MPEG-H 3D Audio for channel and objects content; the total mean MUSHRA score is shown; the confidence intervals were smaller than 2.5 points for every test point. Note that the results presented here are obtained in four different tests.

Before selecting the reference model technology for the continuation of the standardization all systems have undergone rigorous testing. First, 24 test items were selected ranging from typical movie content to signals assumed to be challenging for the coding and rending task. Twelve of the signals contained CO content ranging from 9.1 to 22.2 channels and up to 31 objects. Twelve HOA test items were selected for testing HOA submissions including recorded and synthetic scenes with Ambisonics orders 3 to 6.

A series of listening tests was performed according to the MUSHRA methodology. More than 40,000 answers from a total of 10 test labs were collected. These expert listeners were evaluating the basic audio quality, i.e., their task was to rate any difference of the system under test including timbre, spatial impression, codec artefacts and aspects of creative intent. Several test configurations were chosen to measure the performance of each system at different points of operation:

- Test 1: Rendering to 22.2 loudspeakers in sweet spot
  - Bit rates: 1.2 Mbit/s, 512 kbit/s, and 256 kbit/s
  - Objective: Demonstrate very high quality for reproduction on large reproduction setups
- Test 2: Rendering to 22.2 loudspeakers; off sweet spot
  - Bit rate: 512 kbit/s
  - Objective: Verify results of Test 1 for non-optimum listener position
- Test 3: Binaural Rendering to headphones
  - Bit rate: 512 kbit/s
  - Objective: Demonstrate ability for spatial rendering to headphones
- Test 4: Rendering to alternative speaker configuration
  - Bit rate: 512 kbit/s
  - Speaker configurations: 5.1, 8.1, Random 5.1, Random 10.1

- Objective: Demonstrate ability to perform high-quality rendering for reproduction setups with a lower number of speakers or non-standard loudspeaker layout

The selection was based on a Figure of Merit calculated from the test result and the complexity for binaural rendering. After evaluation the systems submitted by Fraunhofer IIS and Technicolor/Orange Labs were selected as the Reference Model for CO and HOA, respectively.

The listening test results of the selected system for CO at all test points are depicted in Fig. 4.

For CO content, the results show that excellent quality, i.e., more than 80 points on a scale according to MUSHRA [41] can be achieved for bitrates of 512 kbit/s and 1200 kbit/s for playback on the original loudspeaker layout. For a bit rate of 256 kbit/s, "Good" quality, i.e., between 60 and 80 points can be achieved.

In case of binaural rendering of CO content to headphones, no undue degradation of quality due to the computationally-optimized binaural processing can be observed.

Furthermore, coding the content at 512 kbit/s and rendering the decoded signals to a lower number of loudspeaker channels yields results in the "Good" quality range.

VI. Conclusions

This paper introduced the ISO/MPEG-H 3D Audio technology which is designed to provide high-quality bitrate-efficient carriage of immersive spatial audio content. The format is universal in that it accepts all common input types (channel-oriented, object-oriented and HOA-based) and delivers optimized playback on loudspeaker setups from 22.2 and more down to 5.1 and stereo, as well as binaural playback on headphones. Extensions for very low bitrates have been investigated during a second phase of the standardization process and led to promising results. It is anticipated that the new standard substantially contributes to bridging the compatibility gap between the various immersive reproduction loudspeaker setups and formats.

References

[1] R. Alexander, *The Inventor of Stereo: The Life and Works of Alan Dower Blumlein*, ser. ISBN 978–0240516288. Oxford, U.K.: Focal, 2000.

[2] A. D. Blumlein, "Improvements in and Relating to Sound-Transmission, Sound-Recording and Sound Reproducing Systems," British patent 394 325, 1931.

[3] *Multichannel Stereophonic Sound System With and Without Accompanying Picture*, ITU-R Rec.-BS.775–2, Int. Telecom Union, Geneva, Switzerland, 2006.

[4] F. Rumsey, *Spatial Audio*, ser. ISBN 0 240 51623 0. Oxford, U.K.: Focal Press, 2001.

[5] A. Silzle and T. Bachmann, "How to find future audio formats?," in *Proc. VDT-Symp.*, Hohenkammer, Germany, 2009.

[6] M. A. Gerzon, "Perophony: with-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 3–10, 1973.

[7] J. Ahrens, "Analytic methods of sound field synthesis," in *T-Labs Series in Telecommunication Services*, ser. ISBN 978-3-642-25742-1. Berlin/Heidelberg, Germany: Springer, 2012.

[8] C. Chabanne, M. McCallus, C. Robinson, and N. Tsingos, "Surround sound with height in games using dolby pro logic IIz," in *Proc. 129th AES Conv., Paper Number 8248*, San Francisco, CA, USA, Nov. 2010.

[9] B. V. Daele, "The immersive sound format: Requirements and challenges for tools and workflow," in *Proc. Int. Conf. Spatial Audio (ICSA)*, Erlangen, Germany, 2014.

[10] K. Hamasaki, K. Matsui, I. Sawaya, and H. Okubo, "The 22.2 multichannel sounds and its reproduction at home and personal environment," in *Proc. AES 43rd Int. Conf. Audio for Wirelessly Networked Personal Devices*, Pohang, Korea, Sep. 2011.

[11] A. Silzle *et al.*, "Investigation on the quality of 3D sound reproduction," in *Proc. Int. Conf. Spatial Audio (ICSA)*, Detmold, Germany, 2011.

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

[13] K. Hamasaki *et al.*, "Effectiveness of height information for reproducing presence and reality in multichannel audio system," in *Proc. 120th AES Conv.*, Paris, France, 2006.

[14] S. Kim, Y. W. Lee, and V. Pulkki, "New 10.2-channel vertical surround system (10.2-VSS); comparison study of perceived audio quality in various multichannel sound systems with height loudspeakers," in *Proc. 129th AES Conv.*, San Francisco, CA, USA, 2010.

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

[16] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H audio—the new standard for universal spatial/3D audio coding," in *Proc. 137th AES Conv.*, Los Angeles, CA, USA, 2014.

[17] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, Jun. 1997.

[18] S. Spors and J. Ahrens, "A comparison of wave field synthesis and higher-order Ambisonics with respect to physical properties and spatial sampling," in *Paper 7556, 125th AES Conv.*, San Francisco, CA, USA, Oct. 2008.

[19] *Technology of Binaural Listening*, J. Blauert, Ed. Berlin/Heidelberg, Germany: Springer-Verlag, 2013.

[20] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, "ISO/IEC MPEG-2 advanced audio coding," *J. AES*, vol. 45/10, pp. 789–814, Oct. 1997.

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

[22] J. Herre and M. Dietz, "Standards in a nutshell: MPEG-4 high-efficiency AAC coding," *IEEE Signal Process. Mag.*, vol. 25, no. 3, pp. 137–142, May 2008.

[23] *EBU Evaluations of Multichannel Audio Codecs* , EBU-Tech. 3324, UBU, Geneva, Switzerland, Sep. 2007 [Online]. Available: https://tech.ebu.ch/docs/tech/tech3324.pdf

[24] J. Hilpert and S. Disch, "Standards in a nutshell: The MPEG surround audio coding standard," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 148–152, Jan. 2009.

[25] *MPEG-D (MPEG Audio Technologies), Part 1: MPEG Surround*, , 2007, ISO/IEC 23003–1:2007.

[26] J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilpert, L. Villemoes, L. Terentiv, C. Falch, A. Hölzer, M. L. Valero, B. Resch, H. Mundt H, and H. Oh, "MPEG spatial audio object coding—the ISO/MPEG standard for efficient coding of interactive audio scenes," *J. AES*, vol. 60, no. 9, pp. 655–673, Sep. 2012.

[27] *MPEG-D (MPEG Audio Technologies), Part 2: Spatial Audio Object Coding*, ISO/IEC 23003–1:2010, 2010.

[28] M. Neuendorf *et al.*, "The ISO/MPEG unified speech and audio coding standard—consistent high quality for all content types and at all bit rates," *J. Aud. Eng. Soc.*, vol. 61, no. 12, pp. 956–977, Dec. 2013.

[29] *MPEG-D (MPEG Audio Technologies), Part 3: Unified Speech and Audio Coding*, ISO/IEC 23003-1:2012, 2012, .

[30] S. K. Zielinski, F. Rumsey, and S. Bech, "Effects of down-mix algorithms on quality of surround sound," *J. Audio Eng. Soc.*, vol. 51, no. 9, pp. 780–798, 2003.

[31] J. Vilkamo, A. Kuntz, and S. Füg, "Reduction of spectral artifacts in multichannel downmixing with adaptive phase alignment," *J. Audio Eng. Soc.*, vol. 62, no. 7/8, pp. 516–526, Jul. 2014.

[32] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," in *Proc. ACM Trans. Math. Software (TOMS)*, New York, NY, USA, Dec. 1996, vol. 22, pp. 469–483.

[33] E. G. Williams*, Fourier Acoustics*. New York, NY, USA: Academic, 1999.

[34] J. Vilkamo, B. Neugebauer, and J. Plogsties, "Sparse frequency-domain reverberator," *J. Audio Eng. Soc.*, vol. 59, no. 12, pp. 936–943, Dec. 2011.

[35] S. Füg, A. Hölzer, C. Borß, C. Ertel, M. Kratschmer, and J. Plogsties, "Design, coding and processing of metadata for object-based interactive audio," in *Proc. 137th AES Conv.*, Los Angeles, CA, USA, 2014.

[36] *Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level*, ITU-R, Rec. -BS1770.3, Int. Telecom Union, Geneva, Switzerland, 2012.

[37] *Practical Guidelines for Distribution Systems in Accordance With EBU R 128,* , EBU-TECH Document 3344, European Broadcast Union, Geneva, Switzerland, 2011.

[38] U. Zölzer*, Digital Audio Signal Processing*. Chichester, U.K.: Wiley, 1997.

[39] S. Meltzer, M. Neuendorf, and D. Sen, "MPEG-H 3D audio—The next generation audio system," in *Proc. IBC Conf. 2014 Amsterdam.ISO/IEC JTC1/SC29/WG11 N13411: Call for Proposal for 3D Audio*, Geneva, Switzerland, Jan. 2013.

[40] *Coding of Audio-Visual Objects*, ISO/IEC JTC1/SC29/WG11 (MPEG), International Standard ISO/IEC IS 14496, 1999.

[41] *Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*, ITU-R Rec. BS.1534-1, Int. Telecomm. Union, Geneva, Switzerland, 2003.

[42] C. Q. Robinson, N. Tsingos, and S. Metha, "Scalable format and tools to extend the possibilities of cinema audio," *SMPTE Motion Imaging J.*, vol. 121, no. 8, pp. 63–69, Nov. 2012.

[43] V. Pulkki, "Generic panning tools for MAX/MSP," in *Proc. Int. Comput. Music Conf.*, Berlin, Germany, 2000.

[44] S. Zielinski, F. Rumsey, and S. Bech, "Effects of down-mix algorithms on quality of surround sound," *J. Audio Eng. Soc.*, vol. 51, no. 9, pp. 780–789, Sep. 2003.

[45] C. Avendano and J.-M. Jot, "A frequency-domain approach to multichannel upmix," *J. Audio Eng. Soc.*, vol. 52, no. 7/8, pp. 740–749, Jul. 2004.

[46] C. Faller, "Multiple-loudspeaker playback of stereo signals," *J. Audio Eng. Soc.*, vol. 54, no. 11, pp. 1051–1064, Nov. 2006.

[47] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, Jun. 2007.

[48] S. Berge and N. Barrett, "High angular resolution planewave expansion," in *Proc. 2nd Int. Symp. Ambisonics and Spherical Acoust.*, Paris, France, 2010.

[49] *Audio Definition Model—Metadata Specification*, EBUTech 3364, Eur. Broadcasting Union, Geneva, Switzerland, Jan. 2014 [Online]. Available: https://tech.ebu.ch/docs/tech/tech3364.pdf

[50] Dolby Atmos—Next Generation Audio for Cinema (White Paper) 2013 [Online]. Available: http://www.dolby.com/uploadedFiles/Assets/US/Doc/Professional/Dolby-Atmos-Next-Generation-Audio-for-Cinema.pdf

Jürgen Herre received his Dipl.-Ing. degree in electrical engineering from the University of Erlangen-Nürnberg, Erlangen, Germany. He then joined the Fraunhofer Institute for Integrated Circuits (IIS) in Erlangen, Germany, in 1989. Since then he has been involved in the development of perceptual coding algorithms for high quality audio, including the well-known ISO/MPEG-Audio Layer III coder (aka "MP3"). In 1995, Dr. Herre joined Bell Laboratories for a PostDoc term working on the development of MPEG-2 Advanced Audio Coding (AAC). By the end of '96 he went back to Fraunhofer to work on the development of more advanced multimedia technology including MPEG-4, MPEG-7, and MPEG-D, currently as the Chief Executive Scientist for the Audio/Multimedia activities at Fraunhofer IIS, Erlangen. In September 2010, Dr. Herre was appointed Full Professor at the University of Erlangen and the International Audio Laboratories Erlangen.

Prof. Herre is a fellow of the Audio Engineering Society, co-chair of the AES Technical Committee on Coding of Audio Signals and vice chair of the AES Technical Council. He is a member of the IEEE Technical Committee on Audio and Acoustic Signal Processing, served as an associate editor of the IEEE Transactions on Speech and Audio Processing and is an active member of the MPEG audio subgroup.

Johannes Hilpert received a Dipl.-Ing. degree in electrical engineering from the University of Erlangen-Nürnberg, Germany in 1994. Upon graduation he joined the Fraunhofer Institute for Integrated Circuits (IIS) in Erlangen, Germany, where he worked on perceptual audio measurement and MPEG perceptual audio codecs such as MP3 and AAC. Starting in 2000, he headed the team for real-time audio coding algorithms on digital signal processors and since 2001 he is in charge of the Audio Coding and Multimedia Software Group. Since 2013, he has the position of a Chief Engineer of the Audio & Multimedia Division. His recent research topics are parametric multi-channel and multi-object audio coding and 3D audio. Mr. Hilpert is a co-editor of the ISO/MPEG MPEG Surround and SAOC standards.

Achim Kuntz received a Dipl.-Ing. degree in electrical engineering from the University of Erlangen-Nürnberg in 2002. He then joined the Telecommunications Laboratory at the same university carrying out studies on spatial sound reproduction, multidimensional signals and wave field analysis techniques, resulting in his doctoral degree in 2008. He is Senior Scientist at the Fraunhofer Institute for Integrated Circuits (IIS) in Erlangen, Germany, and member of the International Audio Laboratories Erlangen. His current field of interest comprises perceptually motivated signal processing algorithms for audio applications including the acquisition, manipulation, coding and reproduction of spatial sound. He is active in the standardization of audio codecs within the MPEG audio subgroup.

Jan Plogsties studied electrical engineering at the University of Technology in Dresden, Germany. He received his M.Sc. in Audio and Acoustics at the Aalborg University, Denmark in 1998. He worked as a Researcher on binaural sound reproduction and perception at the acoustics labs at Aalborg. In 2001, he joined Fraunhofer IIS in Erlangen, Germany, working on MPEG-4 systems and spatial audio reproduction and audio coding. He was involved in the coordination of the European FP6 project CAROUSSO. He published several conference and journal papers. For many years he was a Lecturer at the ARD.ZDF medienakademie. More recently he contributed to major international standardization bodies (ISO MPEG, 3GPP). Since 2009, he has been heading a research and development team dealing with virtual acoustics and sound reproduction on mobile devices.