INVITED PAPER

MPEG Standards for Compressed Representation of Immersive Audio

This article surveys MPEG-H Audio (universal immersive audio coding) and MPEG-I Immersive Audio (compressed representation for virtual and augmented reality). The article focuses on the differences from other standards as well as the requirements and development process of an MPEG-I 6DoF immersive audio standard.

By SCHUYLER R. QUACKENBUSH, *Senior Member IEEE*, AND JÜRGEN HERRE, *Senior Member IEEE*

ABSTRACT | The term "immersive audio" is frequently used to describe an audio experience that provides the listener the sensation of being fully immersed or "present" in a sound scene. This can be achieved via different presentation modes, such as surround sound (several loudspeakers horizontally arranged around the listener), 3D audio (with loudspeakers at, above, and below listener ear level), and binaural audio to headphones. This article provides an overview of two recent standards that support the bitrate-efficient carriage of high-quality immersive sound. The first is MPEG-H 3D audio, which is a versatile standard that supports multiple immersive sound signal formats (channels, objects, and higher order ambisonics) and is now being adopted in broadcast and streaming applications. The second is MPEG-I immersive audio, an extension of 3D audio, currently under development, which is targeted for virtual and augmented reality applications. This will support rendering of fully user-interactive immersive sound for three degrees of user movement [three degrees of freedom (3DoF)], i.e., yaw, pitch, and roll head movement, and for six degrees of user movement [six degrees of freedom (6DoF)], i.e., 3DoF plus translational x, y, and z user position movements.

KEYWORDS | 3D audio; audio coding; audio compression audio data reduction; augmented reality; immersive audio; MPEG; MPEG-H; MPEG-I; virtual reality.

I. INTRODUCTION

The term "immersive audio" is often used to characterize the latest generation of sound systems that aim at providing an audio experience that conveys to the listener the sensation of being fully immersed into or "present" in a surrounding sound scene. While early sound reproduction systems provided stereophonic sound reproduction over two loudspeakers with an illusion of left-right (and depth) perception to the listener for a limited frontal sound field [1], [2], the second generation added a 360° "surround" experience that extended the presented sound stage to include both to the extreme left and right, as well as sound from behind the listener by adding more loudspeakers from all horizontal directions (e.g., 5.1 and 7.1 [3]–[5]). This already provides a significant degree of user immersion into the sound field. Finally, "3D" sound systems added loudspeakers above and below listener ear level to provide the sensation of full listener immersion (e.g., 5.1+2H, 7.1+4H, and 22.2 [6]–[8]). In this context, 7.1+4H refers to a 7.1 layout with an additional four elevated speakers. Similar immersive effects can be achieved for headphone playback by using binaural reproduction technology that simulates natural sound propagation using head-related transfer functions (HRTFs) or binaural room impulse responses (BRIRs) [9].

Often, the distribution of such immersive 3D audio content happens over bandwidth-limited channels, including streaming, broadcasting, or wireless connections, which requires a bitrate-efficient and high-quality representation of its waveforms and associated metadata. Furthermore, sustainable deployment of immersive audio benefits

Manuscript received February 28, 2020; revised February 16, 2021 and April 7, 2021; accepted April 16, 2021. Date of publication May 28, 2021; date of current version August 20, 2021. *(Corresponding author: Schuyler R. Quackenbush.)*

Schuyler R. Quackenbush is with Audio Research Labs, Scotch Plains, NJ 07078 USA (e-mail: srq@audioresearchlabs.com).

Jürgen Herre is with International Audio Laboratories Erlangen, D-91054 Erlangen, Germany (e-mail: juergen.herre@audiolabs-erlangen.de).

This article has supplementary downloadable material available at https://doi.org/10.1109/JPROC.2021.3075390, provided by the authors.

Digital Object Identifier 10.1109/JPROC.2021.3075390

greatly from widely available and long-term stable open format specifications that enable the economy of scale for content production, distribution, and reproduction technology.

This article provides an overview of two standards from the International Organization for Standardization (ISO), created by its Motion Picture Experts Group (MPEG), that support the bitrate-efficient carriage of high-quality immersive sound. The first is MPEG-H 3D audio, which was standardized in 2015, with a revision (i.e., second edition) in 2019. It is a versatile standard that supports multiple immersive sound signal formats, i.e., channels, objects, and higher order ambisonics (HOA), and is now being adopted in broadcast and streaming applications. The second is MPEG-I immersive audio, an extension of MPEG-H 3D audio, which is currently under development and is targeted for virtual and augmented reality applications.

MPEG-H is described in [11]; however, it is a foundational technology for MPEG-I audio and, thus, requires some description here in order to make this article easily understood by the reader.

II. MPEG-H AUDIO: UNIVERSAL IMMERSIVE AUDIO CODING

The MPEG-H 3D audio specification [10], [11] describes a universal audio coding and rendering environment that is designed to efficiently represent high-quality spatial/immersive audio content for storage and transmission. This is of paramount importance for many types of applications offering 3D audio, such as broadcasting or wireless streaming/download media services.

Since there is no generally accepted 'one-size-fits-all' format for 3D spatial audio, it supports common loudspeaker setups, including mono, stereo, surround, and 3D audio (i.e., setups including loudspeakers above ear level and possibly below ear level). Furthermore, formats that are independent of loudspeaker setups, such as objects and HOA, are supported within the standard.

In addition to efficiently representing high-quality spatial audio content, MPEG-H 3D audio also allows for flexible and optimized rendering over a wide range of reproduction conditions (i.e., loudspeaker setups, headphones, background noise levels in various consumption environments, and so on). Finally, a considerable level of interactivity and customization is available by allowing the user to personalize content playback, including adjusting foreground/background balance, language selection, and dialogue enhancement, i.e., changing dialogue level.

The development of the MPEG-H 3D audio specification within the MPEG audio subgroup was started with early discussions in 2011. A verification test in 2017 assessed the subjective audio quality provided by the technology.

A. Overview and Concepts

The MPEG-H 3D audio architecture supports all three important production paradigms for spatial sound.

1) Channels: Traditionally, spatial sound has been delivered by producing several signals ("channels") that drive loudspeakers positioned in a well-defined geometric setup relative to the listener (e.g., stereo, 5.1, and 22.2). In this way, each channel signal is associated with a specific spatial "meaning." The degree of spatial realism and immersion generally increases with the number of loudspeaker channels. Fig. 1 illustrates the average subjective spatial quality for increasingly rich loudspeaker setups when assuming a 22.2 setup (featuring nine loudspeakers on the upper layers, 11 on the middle one, three on the lower one, and two low-frequency enhancement channels) as the quality reference [12]. A multi-stimulus test with hidden reference and anchor (MUSHRA) was used as the subjective test methodology.

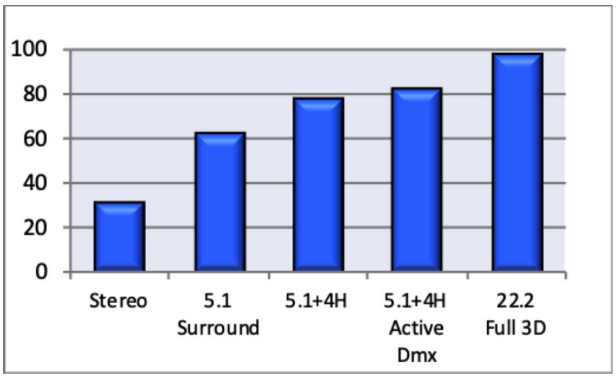


Fig. 1. *Overall sound quality impression (points on a MUSHRA scale from 0 to 100) relative to a 22.2 reference with an increasing number of reproduction channels from stereo to immersive/3D formats, after [12].*

There is a clear improvement from stereo to surround and "3D" (note that, without a 22.2 reference, subjective quality saturated at 5.1+4H). However, one drawback of channel-based production lies in the fact that the produced content asks for a particular loudspeaker setup, and reproduction on other setups that may be available to the user requires additional conversion steps to be performed.

2) Objects: A more recent approach for delivering spatial sound is to produce and deliver spatial audio content as a set of "object" signals with associated metadata specifying the sound source location (and, possibly, other object properties). The locations may be time-varying trajectories to enable moving sound sources, such as a plane flyover. These objects are then reproduced on the user loudspeaker setup (or headphones) by a rendering algorithm. This enables the user to create an interactive/personalized sound experience by adjusting the object characteristics of the rendered output [13]. For example, users could increase or decrease the level of the announcer's commentary or actor's dialog relative to the other audio elements in the audio program. In contrast to the traditional channel paradigm, the object-oriented content representation is agnostic of loudspeaker layouts and provides for greater spatial resolution when presented on a setup with more loudspeakers.

3) Higher Order Ambisonics: An alternative approach to representing spatial audio content is HOA [14], which

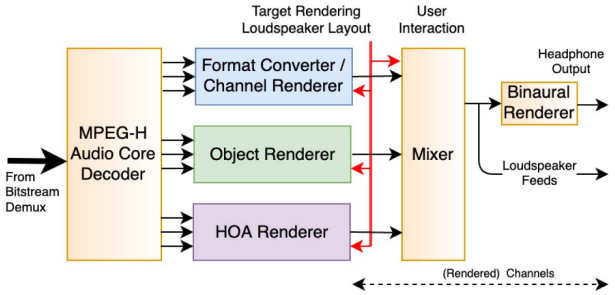


Fig. 2. *Top-level architecture of the MPEG-H 3D audio decoder.*

decomposes the 3D sound field at a particular point in space into spherical harmonics, where the HOA "coefficient" signals are the weighting of the associated harmonics at given time instants. While first-order ambisonics provides limited spatial resolution, higher orders provide increasingly higher resolution and better approximation of the original sound field. Similar to the object-based paradigm, HOA signals are agnostic of loudspeaker layouts and need a renderer in order to be reproduced on a target loudspeaker setup (or on headphones).

Fig. 2 shows a top-level overview of an MPEG-H 3D audio decoder. The main part of the incoming MPEG-H 3D audio bitstream is decoded by the core decoder that reproduces the encoded waveforms that represent either channel signals, object signals, or HOA coefficient signals. These waveforms are then further processed in the three dedicated processing chains handling these paradigms. A so-called format converter processes the channel signals to convert them to the target rendering loudspeaker layout. Also, object signals and HOA components are rendered to the target layout by the object renderer and the HOA renderer, respectively. All three contributions are summed up by a mixer and to generate the final loudspeaker feed signals. Also, the binaural output can be provided. For simplicity, signal postprocessing by dynamic range compression (DRC) and loudness normalization is not shown. This postprocessing produces an output signal with a dynamic range that is appropriate for the consumption environment (e.g., on an airplane) and aligned in terms of subjective loudness when switching between different programs.

The most important building blocks and aspects of the system are discussed in more detail in the following.

B. Waveform Coding

The core part of the codec compresses and represents the waveforms of the channel, object, and HOA signals. To this end, the MPEG-H 3D audio codec is based closely on the technology of the previously developed MPEG Unified Speech and Audio Coding (USAC) [15], [16] system that is the state-of-the-art MPEG codec for compression of mono, stereo, and multichannel audio signals. It provides the best audio coding quality for both audio and speech signals at rates down to 8 kbit/s per channel by combining elements from generic perceptual audio coding and state-of-the-art speech coding. Both codec parts are tightly integrated and can be used adaptively depending on the nature of the input signal at each input frame, in this way choosing the more efficient of both technologies.

For application in MPEG-H 3D audio coding, the USAC technology has been augmented by some tools to further enhance its coding efficiency. Most notably, the newly created intelligent gap filling (IGF) tool [17] provides an enhanced mechanism for noise filling, i.e., for filling spectral regions for which spectral coefficients cannot be transmitted due to a shortage of available bits. In contrast to earlier methods, such as perceptual noise substitution (PNS) [18] or spectral band replication (SBR) [19], the tool parametrically restores portions of the transmitted spectrum while allowing to "intelligently" intermix between transmitted spectral coefficients and coefficients that are parametrically restored from lower frequency spectral regions that have been transmitted in a waveform preserving way. The encoder has control over the assignment and the processing of these spectral regions, or tiles, based on an input signal analysis. In this way, spectral gaps can be filled with spectral coefficients that have a better perceptual match than that pseudorandom noise sequences of conventional noise filling would provide.

Furthermore, a so-called quad-channel element permits the efficient joint encoding of a channel pair (which represents channels that are symmetric to the listener median axis) plus its corresponding pair of top layer speakers. While coding of the main channel pair is done using M/S stereo to avoid binaural unmasking effects [20] and exploit horizontal signal redundancies, vertical signal dependence is exploited by parametric coding of vertically aligned channel pairs [11].

Furthermore, the codec bitstream has been designed to enable seamless transitions (instantaneous rate switching) between several encoded stream versions at different coding rates. Specifically, the compressed audio payload allows fast cue-in as it appears in the context of MPEG dynamic adaptive streaming (DASH) [21] by adding so-called "immediate playout frames" to the syntax. This is particularly advantageous for adaptive streaming over IP networks.

C. Format Conversion

In order to allow the reproduction of encoded channel-based content on any available loudspeaker setup connected to the MPEG-H audio decoder, the format converter maps the encoded channel signals to the target speaker layout. As an example, the decoder may detect a 5.1 surround reproduction loudspeaker setup, while the content has been encoded in 22.2-channel format. Thus, an appropriate high-quality downmix has to be performed in order to enable the best possible listening experience, given the available speaker layout. In this way, the format converter allows universal output formats also for channel-based program material. In order to provide the

best possible user experience, the format converter stage addresses a number of specific aspects, as described in the following.

1) Optimized Downmix for All Setups: The format converter is capable of automatically generating optimized downmix matrices for all target loudspeaker setups, including nonstandard positions (as appear frequently in users' homes) to map the transmitted channel configuration to the output loudspeaker layout.

This is achieved by an iterative search through an internal lookup table of tuned mapping rules for each input channel that happens once during the initialization phase of the format converter module. Each rule describes the mapping of one particular input channel to one or more output loudspeaker channels, possibly accompanied by a specific equalization curve that is to be applied when this particular rule has been selected. The iterative search through the lookup table terminates when the rule with the highest anticipated mapping quality has been found. For each potential input channel, the associated rules have been designed individually based on expert knowledge such that, e.g., excessive use of phantom sources (and, thus, spatial blur) is avoided. Even asymmetric loudspeaker setups can be accommodated. In this way, the rule-based downmix coefficient generation allows to flexibly adapt to different input/output configurations while, at the same time, ensuring high conversion quality due to the use of expert knowledge.

2) Advanced Active Downmix: When the optimized downmix coefficients have been found, they can be used as part of an advanced downmix process that is designed to avoid downmix artifacts, such as signal cancellations or comb filtering. Such artifacts could occur when linearly combining input signals that may exhibit significant correlations using static gains. In practice, such effects are quite common for current 3D audio content since it is frequently produced from available 2-D (surround) content by populating the missing channel signals with delayed and filtered copies of the originally available signals.

The MPEG-H 3D audio active downmix process adapts to the properties of the input signal by measuring the correlations between the incoming signals and aligning the phases of these signals, if required. Furthermore, a frequency-selective energy compensation is included that ensures energy preservation after the downmix step and, thus, avoids timbral colorations. Effectively, the compensation algorithm avoids the effects of high interchannel correlation, leaving uncorrelated input unaltered. In addition, a low-complexity active downmix is specified in the MPEG-H standard, which achieves high-quality downmixing while reducing the computational load and processing latency.

3) Downmix Controlled by Artistic Intent: Optionally, downmix matrices can be sent by the content producer or broadcaster along with the bitstream to ensure that the output complies with the artist's intent.

D. Object Rendering

In MPEG-H, objects consist of an encoded monophonic waveform and associated metadata describing how to render these objects to a predefined spatial location. Moreover, an associated object gain also can be transmitted, and both position and gain can be defined as arbitrary dynamic trajectories (see metadata MPEG-H overview [22]).

The MPEG-H object renderer is based on the well-known vector base amplitude panning (VBAP) [23] algorithm and renders the transmitted audio objects to any given output (target) loudspeaker setup. It, thus, processes one decoded audio stream per transmitted audio object, the associated decoded object metadata (e.g., time-varying position data and gains), and the geometry of the target rendering setup.

VBAP positions sound sources by triangulating the 3D surface surrounding the listener. MPEG-H provides an automatic triangulation algorithm that also supports nonstandard/arbitrary target loudspeaker setups. Object locations that are not covered well by the target loudspeaker setup (e.g., locations below the horizontal speaker plane) are supported by the addition of imaginary speakers to provide complete 3D triangle meshes for any setup to the VBAP algorithm [11].

In accordance with the VBAP approach, the MPEG-H 3D audio renderer searches the surrounding loudspeaker triangle that is closest to each object and builds an associated vector base. For this loudspeaker triangle, panning gain values are computed considering overall signal energy preservation. The computed gains are then linearly interpolated between the values computed for different time stamps. Finally, the contributions of each object are summed up to form the final renderer output signals.

E. HOA Decoding and Rendering

HOA describe the audio scene as a 3D acoustic sound field that is represented as a truncated expansion of the wavefield into spherical harmonics [14]. This completely determines the acoustic quantities within a certain source-free region around the listener's position up to an upper frequency limit, beyond which spatial aliasing limits the expansion's accuracy. The time-varying coefficients of the spherical harmonics expansion are called HOA coefficient signals and carry the information of the wavefield that is to be described for transmission or reproduction.

Generally speaking, for a complete 3D expansion of order n, $(n+1)^2$, coefficient signals have to be carried, and many of those signals exhibit—depending on the nature of the sound field—considerable correlations (and, thus, redundancy) between them. Thus, both aspects of redundancy and irrelevance reduction have to be considered to arrive at a bitrate-efficient and yet high-quality representation. To this end, the MPEG-H 3D audio encoder includes a

dedicated toolset for HOA coding, which decomposes the input HOA signals into a different bitrate-efficient internal representation that is used for rate-reduced transmission in the MPEG-H 3D audio bitstream. Conversely, this process has to be reverted in the MPEG-H 3D audio decoder. The underlying algorithmic concepts will be described in the following.

1) Representation of Direct Sound Components: In the first step, a decomposition of the sound field into direct and diffuse sound components is carried out to reduce redundancy in the HOA coefficient signals. Strong sound events that originate from a distinct direction ("direct" components) introduce highly correlated components into many HOA signals, as can be seen from a spherical harmonics expansion of plane waves [24]. In order to identify such redundancies, the encoder performs an analysis of the input signal to detect the presence of significant direct sounds (also called "predominant sounds" in the content of MPEG-H 3D audio) and transmits them separately as parametrically coded plane waves plus associated directional metadata. Then, the direct sound contributions are subtracted from the HOA coefficients for further processing of the remaining ambient sound field components. In this way, a considerable reduction of redundancy can be achieved for input with significant direct sound contributions. Besides the parametric coding of plane wavefield signals, MPEG-H 3D audio also includes a mode for efficiently representing more general directional characteristics.

2) Representation of Ambient Sound Components: In the second step, the remaining ambient sound components are processed. Since localization accuracy is typically not of high perceptual importance for such nondirectional sound field components, their HOA representation order can be reduced without perceptual penalty in order to increase coding efficiency. Nonetheless, this representation still may include high correlations between the HOA coefficient signals. Not only is this disadvantageous from a redundancy point of view but individual perceptual coding of these coefficient signals can also lead to undesirable spatial unmasking of the quantization noise contributions introduced by each of the coefficient signal coders. As a solution to this challenge, the ambience HOA representation is first transformed into a different spatial domain using a spherical Fourier transform. The resulting virtual loudspeaker signals exhibit less correlation and are then used as input to a multichannel codec core for bitstream encoding and transmission. The number of transmitted virtual loudspeaker signals can be chosen according to the available bitrate and perceptual quality considerations.

In the MPEG-H 3D audio decoder, reverse processing takes place as compared to the encoding process. The parametrically represented predominant sound components are resynthesized as HOA contributions. Then, the transmitted virtual loudspeaker channels representing the ambient sound field are mapped back to the HOA domain. Finally, both contributions are added and rendered to the reproduction setup using a generic HOA rendering matrix that is adapted to the target loudspeaker setup layout.

F. Additional Tools & Functions

Besides the core components described so far, the MPEG-H 3D audio decoder also includes a number of additional tools and functions that enhance its performance or applicability in specific situations. These are discussed as follows.

1) Parametric Coding of Objects: In order to provide efficient coding of object-based content at very low bitrates, MPEG-H 3D audio includes a generalization of the MPEG-D Spatial Audio Object Coding (SAOC) scheme [25]. A set of objects to be jointly coded is first downmixed into one or more downmix signals that are then transmitted by regular MPEG-H waveform coding tools. In addition, a compact set of side information, which characterizes the properties of the original object signals, is transmitted alongside the waveform bitstream. Most notably, this side information describes the object level differences (OLDs) and the interobject correlations (IOCs) between the objects in each time/frequency tile. On the decoder side, the transmitted downmix signals are decoded and transformed directly into a set of desired rendered output signals, as needed for the target loudspeaker setup. This is achieved by time-/frequency-dependent matrixing and decorrelation, resulting in an efficient single-step decoding architecture that avoids the need for separate object reconstruction and subsequent rendering.

Since the currently specified MPEG-H 3D audio profiles (see the section on MPEG-H 3D audio performance) target higher bitrates, they do not include parametric object coding.

2) Dynamic Range Control and Loudness Control: In order to enable optimum playback in each consumption environment, the MPEG-H 3D audio decoder also includes a *dynamic range control* (DRC) feature that can be applied to the final output signals and also to individual intermediate sound components, such as objects. The underlying DRC technology is that from the MPEG-D specification [26] and allows to control the dynamic range of the playback in accordance with the background noise conditions of the playback environment (e.g., living room, car, and airplane), effectively providing better subjective user experience in very noisy environments by increasing the level of the quietest portions of the audio program such that they are more easily heard. Furthermore, a loudness normalization function avoids loudness jumps when the user switches between different programs.

3) Binaural Rendering: Besides reproduction on loudspeakers, MPEG-H 3D audio also supports binaural reproduction of spatial sound on headphones. This allows convincing consumption of immersive audio on common

mobile devices, such as mobile phones, handhelds, and portable music players, where headphones would be used.

4) Content-Based Interactivity: The use of audio objects in audio productions opens up the possibility of *user interaction* with the content. To this end, all sound components (channels, objects, and HOA components) that are embedded in the MPEG-H 3D audio bitstream can be selected by the user during playback and adjusted in level offering the possibility of personalized playback. A simple adjustment might be increasing/decreasing the level of the commentary signal relative to the other audio elements according to user preference and, in this way, enhancing the intelligibility of the dialogue. The extent of possible user interaction is under full control of the producer by embedding specific control metadata during the content creation.

G. Performance

As part of the standardization process, MPEG conducts a verification test of all standardized technology. The verification test for MPEG-H 3D audio [27] assessed the performance of a subset of the standardized technology, the low complexity profile. It consisted of four subjective listening tests, each evaluated the audio quality and compression performance at an operating point representing a distinct use case.

The test material consisted of 36 audio items selected to represent typical and critical audio material. The material was either channel-based, channels plus objects, or scene-based, as HOA of a designated order, possibly also including objects.

The subjective tests were conducted at seven test labs. The first three tests (Tests 1–3) were conducted in high-quality listening rooms that were calibrated to conform to the criteria set forth in BS.1116 [28] and also calibrated to be perceptually similar to each other. The fourth test (Test 4) was conducted in acoustically isolated sound booths. Altogether, the test results were based on 190 subjects and nearly 10 000 subjective scores.

Test 1 (Ultra-HD Broadcast): This use case assumed an 8k video broadcast and immersive 22.2-channel and 11.1-channel (as 7.1+4H) audio presentation formats. Since the video would be expected to have a high bit rate, the audio coding bitrate was proportional at 768 kb/s, the highest bit rate amongst the four tests. This test used the "Triple Stimulus Hidden Reference" subjective test methodology (ITU-R BS.1116) [28]. The subjective results showed that the MPEG-H 3D audio low complexity profile operating at 768 kb/s with highly immersive audio content achieved a quality of "Excellent" on the BS.1116 quality scale. It had a mean score (as a BS.1116 "diff score") of −0.31 with a 95% confidence interval of ±0.04, which is well above the −1.0 limit (4.0 out of 5.0 in "absolute scores") recommended in ITU-R BS.1548-4 for "high-quality emission" for broadcast applications.

Table 1 MUSHRA Descriptors and Associated Score Range

Descriptor	Score
Excellent	80-100
Good	60-80
Fair	40-60
Poor	20-40
Bad	0-20

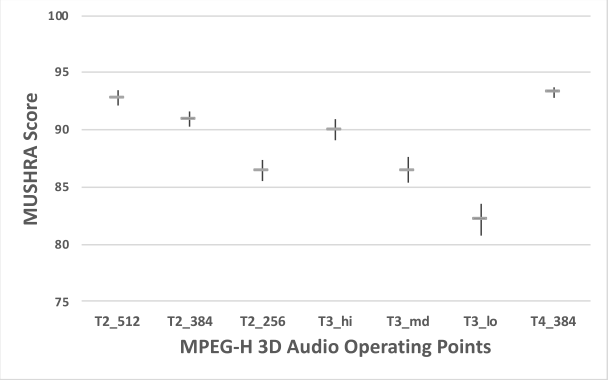


Fig. 3. *Plot of the subjective audio quality of MPEG-H 3D audio for Test 2, Test 3, and Test 4 (for greater visibility, only the MUSHRA scale above 75 points is shown here).*

Test 2 (HD Broadcast): This use case assumed a broadcast program with HD video and immersive audio with 11.1-channel (as 7.1+4H) or 7.1-channel (as 5.1+2H) loudspeaker layouts. All audio was coded at three bit rates: 512, 384, and 256 kb/s.

Test 3 (High-Efficiency Broadcast): As in Test 2, all audio was coded at three bit rates, but this use case assumed a need for greater compression, and the specific bit rates depended on the number of channels in the audio material. Bit rates ranged from 256 (5.1+2H) to 48 kb/s (stereo).

Test 4 (Mobile): This use case assumed that content consumption would be "on the go" (via a mobile device). It used the coded immersive audio content from Test 2, at the 384 kb/s rate, and rendered for headphone presentation using the MPEG-H 3D Audio binauralization engine.

Tests 2–4 all used the "method for the subjective assessment of the intermediate quality level of coding systems" (MUSHRA) [29]. In a MUSHRA test, the correspondence of subjective quality (indicated by descriptor) and the range of subjective score are shown in Table I.

The test results are given in Fig. 3, where the vertical axis is the subjective score (for greater visibility of results, its low end is 75 rather than 0), and the horizontal axis shows the MPEG-H 3D audio operating points tested. In the specific operating point names, the prefix T2_, T3_, and T4_ indicates configurations tested in Test 2, Test 3, and Test 4, respectively, and the numerical suffix indicates the bit rate, in kb/s. A suffix of hi, md, and lo indicates the high, medium, and low bitrates for audio signals in

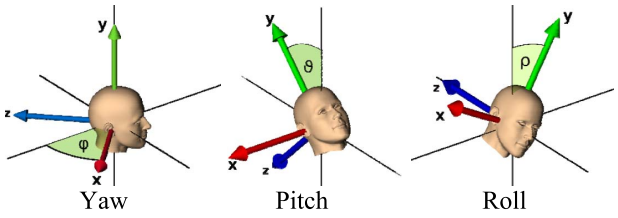


Fig. 4. *Yaw, pitch, and roll rotations, where the yaw is rotation around the +y-axis, the pitch is rotation around the +x-axis, and the roll is rotation around the +z-axis.*

Test 3. The plot shows mean scores as a horizontal tick and 95% confidence intervals on the mean as a vertical stroke.

The subjective results show that, for all bit rates in each of Test 2, Test 3, and Test 4, MPEG-H 3D audio low complexity profile achieved a quality of "Excellent" on the MUSHRA subjective quality scale.

III. MPEG-I IMMERSIVE AUDIO: COMPRESSED REPRESENTATION FOR VIRTUAL AND AUGMENTED REALITY

A. MPEG-I Differs From Other MPEG Standards

MPEG has a long record of standards that support broadcast and streaming of content. In previous MPEG standards, the content was fully conceived at production and the consumer viewed/listened to the presentation from a fixed point of view. This is the environment for MPEG-2 and MPEG-4 audio, visual, and systems delivery of content (i.e., the widely deployed versions of MPEG-4 audio, which supports only channel-based audio). MPEG-H 3D audio is largely the same but additionally supports binauralization of audio for presentation via headphones. This permits the users with the head tracker to turn their heads while viewing the video with the audio sound stage remaining in the correct relationship to the visual scene.

MPEG-I immersive audio departs from the framework of passive consumption of media from a fixed user position. In MPEG-I, the user can move around in the world created by the media presentation, with head movement or both head movement and body movement in virtual space, where we assume that audio presentation is done via headphones. Head movement is designated as three degrees of freedom (3DoF), with axes yaw, pitch, and roll; both head movement and full-body movement are designated as six degrees of freedom (6DoF) with axes yaw, pitch, roll, x, y, and z. The yaw, pitch, and roll movements are illustrated in Fig. 4. Three DoF requires rotation of the sound image on playback, while 6DoF requires rotation and translation of the sound image on playback.

The world created by the media presentation may be a fully virtual world, in which MPEG-I decoder and renderer present all audio–visual aspects of a synthetic world, or it may be an augmented reality world, in which synthetic audio–visual objects are superimposed in the user's real world. In this case, a head-mounted display (HMD) is a "see-through" display on which video "objects" can be projected.

B. Audio Rendering for VR/AR

MPEG-2 and MPEG-4 streams contain compressed media and perhaps some metadata. MPEG-H 3D audio streams contain compressed media and considerably more metadata. 3D audio supports highly immersive audio presentations, e.g., 22.2 channels, in which loudspeakers may surround the user, and metadata can have an impact on audio sound objects, e.g., to dynamically animate their location. However, all aspects of the audio are conceived and mixed at the production side, and the decoder merely plays out the compressed and transmitted audio, as with MPEG-2 and MPEG-4 movies.

In contrast, MPEG-I streams contain all compressed media needed for a virtual environment and also all metadata necessary to describe a virtual environment. In the case of audio, this metadata would describe all aspects of sound-emitting objects in the environment and also all aspects of the environment that impact the user's perception of sound in that environment. For example, the user walks into a virtual restaurant, and there are many dinner guests talking. However, the dining area is small and loud, so the metadata must describe not only the other guests' positions and orientations but also the positions of the floor, walls, and ceiling of the room and the acoustic reflectivity of each. Furthermore, as the user walks from the restaurant door to their table, the audio must change since the user's position relative to other guests and the wall surfaces change.

When MPEG-I describes a virtual world, the compressed media and metadata associated with that world must be sufficiently rich that the audio–visual presentation gives the user the feeling of actually being present in the world. This sense of *presence* is a fundamental goal for MPEG-I technology.

C. MPEG-I 3DoF Audio

In 2017 MPEG specified Omnidirectional Media Format (OMAF) [30]. The "omnidirectional" in the name derives from its function: the user can sit in the viewing position and see and hear the content while turning their head in any direction. It specifies the coding, encapsulation, presentation, and consumption of video, audio, image, and text, as well as the signaling and transport over DASH and MPEG media transport (MMT) [31]. OMAF has several profiles that support varying video capability, all with 360° viewing. The audio is provided by MPEG-H 3D audio with loudness and DRC tools and, of course, with binauralization tools for presentation via headphones.

The OMAF specification opens the door to VR movies, in which a user can step into the movies and explore with a 360° "room-sized" screen. It also supports VR sports and

concerts, in which a user can be a virtual participant at the live event.

D. Requirements for MPEG-I 6DoF Audio

The steps for developing an MPEG standard are typically to: 1) create the requirements that the standard must fulfill; 2) develop an architecture that matches the requirements; 3) issue a Call for Proposals (CfP) for technology that meets the requirements; and 4) further develop the submitted and selected technology such that requirements are met and the desired level of performance is achieved.

There are more than 25 requirements for MPEG-I immersive audio [32], but they can be summarized into a smaller number of main topics that MPEG-I immersive audio will support, as listed in the following:

1) spatial sound reproduction, to give the user a perceived experience that is consistent with the user's 6DoF movement in the environment;
2) efficient representation and compression of media and metadata;
3) presentation via headphones or loudspeakers;
4) sound source models with directivity and spatial extent;
5) convincing rendering of the room or environment acoustics;
6) occlusion of sound sources (e.g., due to room or environment geometry);
7) Doppler shift of audio associated with fast-moving sources;
8) locally captured audio (e.g., the user's voice) rendered to be realistic given the acoustic environment;
9) live audio of other users (e.g., multiparticipant virtual world) with latency suitable for conversation.

The dominant requirement is the realistic rendering of sound sources given the user's current virtual acoustic environment such that the user experiences a sense of actually being present in the virtual world. Of course, an equally important requirement is the need for compression of the data needed for the sounds and the acoustic rendering.

The use cases leading to the MPEG-I immersive audio requirements typically consider video presented via a head-mounted display (HMD) and audio presented via headphones, but the presentation could also be via HMD and loudspeakers or even an immersive video display and loudspeakers. Besides the typical use case of a user exploring a virtual world, there is also a case for a multiuser experience, termed "social VR." This could be as simple as several friends on a virtual couch watching a virtual soccer match. The friends watch a shared virtual TV, and MPEG-I immersive audio permits them to have natural conversations between themselves.

E. Developing MPEG-I 6DoF Immersive Audio

MPEG experts have developed architecture and requirements for MPEG-I immersive audio, and the development of technology that maps onto the architecture and meets the requirements will begin in 2021 and is expected to be completed in 2023. This specification will require compression and transmission of audio signals and virtual acoustic environment information, and the associated reconstruction of audio and rendering with regard to the user position and orientation within the virtual acoustic environment. Since MPEG only recently completed the MPEG-H 3D audio specification, and this audio codec achieves the best performance within the suite of MPEG audio standards (from stereo through 22.2 channels), it was decided that the representation of all audio material would be via MPEG-H 3D audio. This frees MPEG-I immersive audio to concentrate on the representation and compression of metadata needed to create a virtual audio scene and on the rendering of the scene for the user's position and orientation. In this respect, MPEG-I immersive audio will use MPEG-H 3D audio as its audio media compression engine and add additional metadata to that compressed data stream.

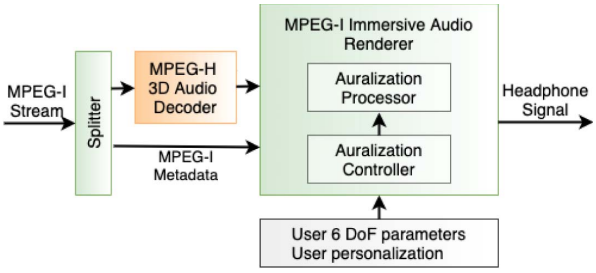


Fig. 5. *Architecture of MPEG-I immersive audio decoding/rendering.*

A simplified block diagram of the MPEG-I immersive audio architecture is shown in Fig. 5. In the architecture, the MPEG-I compressed stream is at the left and is split into an MPEG-H 3D audio stream and an MPEG-I metadata stream. The former is decoded by an MPEG-H 3D audio decoder to produce the audio signals needed to render the virtual scene. At the right-hand side of the figure is the MPEG-I immersive audio renderer. It takes as inputs (on the left-hand side) the audio signals, and signal and scene metadata and (on the bottom) user 6DoF position and orientation (and other interaction information), as well as user personalization, such as HRFT and headphone equalization filters. It produces as output the user's headphone signal. Within the immersive audio renderer, there may be additional structures: the "auralization controller" and the "auralization processor." In this envisioned structure, the processor implements audio processing (e.g., filtering) at the sampling rate, while the controller sets up the audio processing (e.g., which filter), at a rate responsive to changes in the user environment, such as user 6DoF motion. This is a simplified block diagram, which does not show aspects of social VR or of rendering tool extensibility.

As stated previously, it is the goal of MPEG-I to achieve the sense that the user is present in the virtual world.

A component of this is that audio sound sources are perceived to be coming from (i.e., colocated with) their associated visual sources. It is well known [33] that visual cues provide a strong influence on the perceived localization of associated sound sources. It has also been shown [34] that full-body motion in the physical world, as with 6DoF, provides strong cues as to what a user should perceive in a virtual world.

For these reasons, it was decided that the evaluation of possible MPEG-I immersive audio technology should be done as an audio–visual experience that permits full 6DoF of user motion in the physical world. The audio–visual experience is presented via a high-quality HMD and high-quality headphones. The HMD has a head tracker that permits the user's 6DoF parameters to be delivered to the audio–visual presentation computer. It was decided that the visual presentation would be via Unity [35] with the option of custom scripts and visuals being developed for assessing MPEG-I immersive audio technology. The audio presentation is done via Max/MSP [36], in which Max directs the flow of audio to the user's headphones and in which candidate MPEG-I immersive audio technology is implemented as a max external. The platform that renders visuals to the HMD and audio to the user's headphones is referred to as the MPEG-I audio evaluation platform (AEP) [37].

In order to assess the full performance of candidate MPEG-I immersive audio technology, there needs to be appropriately rich test material. This would be a Unity visual scene containing sound sources associated with any combination of channels, objects, and HOA audio. The simplest to consider is a scene with one or more audio object sources, where each object may have directivity and the environment may have reflections (room walls, floor, and ceiling), and furthermore, the environment may have occluding elements (e.g., walls) that may interfere with sound propagating from an audio object to the user's position. In addition, the test material may have moving objects that could move behind occluding elements (with respect to the user's position) and also could have a speed sufficient to be rendered with the Doppler shift. Beyond objects, channels can be used to bring existing channel-based audio material into a scene. This could be simple two-channel (stereo) audio that permits an audio object to have some perceived width, which is often appropriate. It could also be an immersive channel bed, e.g., 7.1+4H, that could provide either an ambience or ambience plus perceived audio sound sources embedded within the channel bed signal. Finally, an audio source can be an HOA signal, which represents an entire audio scene.

There is an additional category of scenes in which there are multiple HOA signals, designated as "multipoint HOA." In a 3DoF user experience, it is appropriate that the HOA capture point is also the point (x, y, z) at which the user is situated. However, in a 6DoF experience, it is possible to have scenes with multiple HOA capture points (i.e., with different x, y, and z locations) with the intent that the user could walk from a location of one HOA capture point to another HOA capture point and experience a meaningful and consistent audio presentation, e.g., based on the two HOA signals. In addition, a single HOA source can be rendered in the usual way when the user is "near" the HOA capture point ("internal" rendering, with sounds surrounding the user and at a distance), or it can be rendered as a point source when the user is distant from the HOA capture point ("external" rendering), with a graceful transition from internal to external.

Since all proponents responding to a CfP must encode and represent the audio media and metadata associated with each scene, there must be a clear expression of the structure and meaning of the components of a scene. As stated earlier, all audio signals are encoded using MPEG-H 3D audio. An MPEG-I immersive audio encoder must take as input the scene metadata, carry the MPEG-H 3D audio compressed audio signals, and produce as output a proponent-specific stream that can be processed by the proponent immersive audio decoder, which produces, e.g., headphone audio. For this purpose, the MPEG-I 6DoF audio encoder input format (EIF) has been created.

The EIF describes the structure and representation of the scene metadata information that the MPEG-I immersive audio encoder must read and compress. It is like a scene graph, where each item of test material is one scene. As previously mentioned, a scene might be a restaurant with descriptions of the dimensions and acoustic properties of walls, ceiling, and floor, along with location and size of tables and position and orientation of dinner guests (as potential audio sources). In addition, there may be a band playing at the far end of the room. The scene acoustic metadata can be static and treated as a "one-time" transmission, while the audio signals of the band instruments and singer vocals would be continuous audio signals.

As the next step in the standardization process, a CfP on MPEG-I immersive audio technology is about to be issued, in which a prospective respondent receives information on: 1) requirements; 2) EIF; 3) AEP; and 4) test material, which is described in the CfP itself. Furthermore, the CfP will lay out the timeline and logistics for responding to the CfP and how the submissions will be evaluated. One important aspect of the evaluation of submissions is conducting real-time subjective performance tests. As already mentioned, this will be done using the AEP.

A block diagram of the AEP is shown in Fig. 6. In the figure, to the right-hand side of the vertical dashed line is the AEP, which is a computer running Unity and Max/MSP. Unity displays the video in the HMD, and the head tracker in the HMD, responding to the positioning beacons, connects back to Unity, which then sends user position and orientation information to Max/MSP via a Unity extOSC message. On this computer platform, Max/MSP is able to support a number of max externals running in parallel. In the AEP, each max external is a candidate immersive audio decoding and rendering engine running in real time.

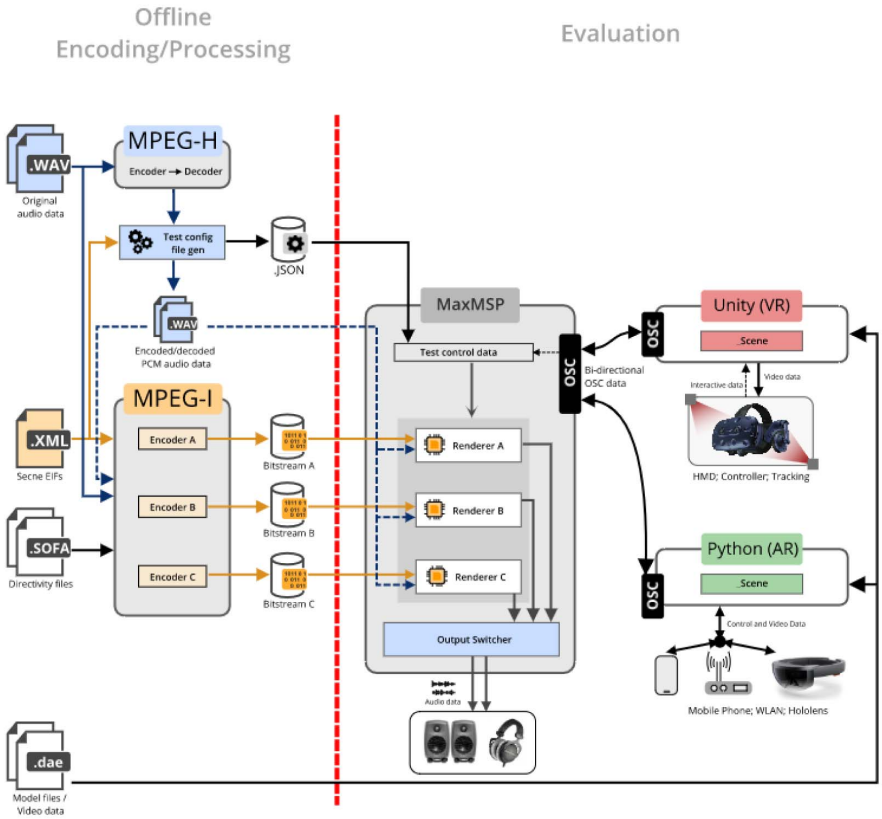


Fig. 6. ***MPEG-I AEP.***

Each of these will be supplied by a proponent as part of the response to the CfP. Custom scripting on the Unity engine permits the user to switch from one proponent technology to another, where the output audio is rendered seamlessly as the AEP switches audio playback from one proponent renderer to another.

The left-hand side of the figure shows the processing of MPEG-I immersive audio streams. In fact, at the time of the CfP, a stream format has not yet been determined, but what is known is that MPEG-H 3D audio will be the codec for all audio signals. Hence, these audio signals are common to all proponent renderers. In the AEP, audio signals can be "precoded": original audio signals are encoded and then decoded using MEPG-H 3D audio, and these signals are then supplied to Max/MSP and to the individual max externals. In contrast, each proponent will have the freedom to compress the signal and scene metadata as they wish, and these data in a compressed format will be available to the proponent's max external.

As already stated, the AEP permits the subject to seamlessly switch between the audios produced by the proponent's max external renderers. While a MUSHRA subjective test methodology could be used, it was decided to use the A-B comparative test methodology [38]. Controlling that switching is via a graphical user interface that is viewable as part of the Unity visual rending to the HMD. With the appropriate controller button press, the AB-style panel shown in Fig. 7 becomes visible. This panel permits the subject to select between two proponent systems to hear, A or B, and to make a comparative scoring via the sliders, where descriptors of subjective quality are shown above the slider. A full or incomplete set of AB comparisons will be made, and a Thurstone V analysis [39] of the comparison scores permits a rigorous statistically based differentiation of systems under test.

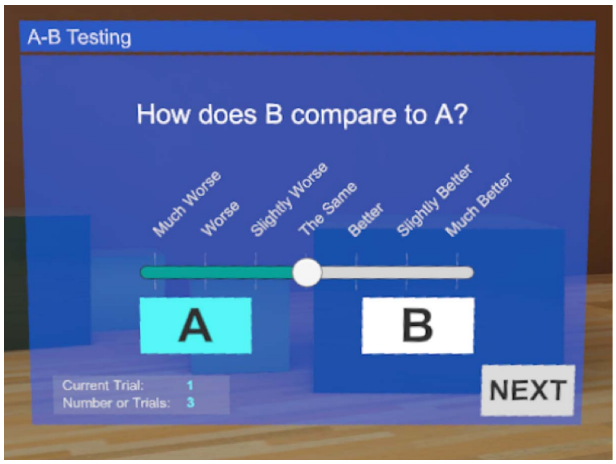


Fig. 7. ***AB-style evaluation panel.***

While A-B comparative methodology will be used in the CfP, other test methodologies may be employed later in the standardization process, such as MUUSHRA or absolute category rating [40], using attributes such as plausibility, externalization, consistency, and basic audio quality.

After proponents have submitted their technology in response to the CfP, all submitted technology will be assessed using the AEP. Subjects from many test labs will participate in the CfP subjective listening test although, in this case, they will have an audio–visual experience. For each test scene, they will give each proponent system a score, from which a mean score over all subjects and all test scenes can be calculated for each proponent system. The subjective performance is a very important factor in selecting the "best" proponent system, but other aspects of the systems under test also must be considered: the "motion-to-sound" latency or the time from a user's head turn to the time that the rendering of the sound field corresponding to the new head position is presented to the headphones must be below the perceivable threshold; the size of the scene's compressed metadata that must be transmitted in order to render a scene; and the complexity, both in terms of computation and platform memory. All these factors (and more not mentioned here) must be considered when selecting the best technology.

After this "base" technology has been selected, the MPEG standardization process will engage in a series of core experiments (CEs), in which MPEG experts conduct controlled experiments in attempts to improve the performance of, or add additional functionality to, the base technology. For example, a CE could compare the performance of the base technology to the base technology with one module replaced by another "enhanced" module. Typically, performance is measured by subjective quality, but it could also be complexity or compression. If the base with the enhanced module provides better performance, then this becomes the new base technology. In this way, the CE processes provide continuous improvements in performance.

When the developed technology is deemed of sufficiently high performance, the CE process will be concluded. At this point, the standard is frozen, except for review and clarification of the specification text, or fixing bugs in the companion source code, and MPEG undertakes to document the performance of the standard by way of a formal subjective test, or verification test. This is similar to the subjective test in the CfP but will have only the final MPEG-I immersive audio technology being assessed. The test will use the AEP, but the several systems under test will instead be MPEG-I immersive audio at meaningful operating points.

Getting to a final MPEG-I immersive audio standard will take some time. Even the task of constructing the AEP, creating associated test scenes, and all supporting programs and scripts have been a significant effort. It is anticipated that MPEG-I immersive audio will become International Standard in 2023.

IV. CONCLUSION

This article discusses two recent MPEG standards for the bitrate-efficient representation of immersive audio. The MPEG-H audio specification has been designed as a universal coding and rendering technology, offering unprecedented flexibility in content representation (channels, objects, and HOA) and presentation (supported standard and nonstandard loudspeaker setups, binaural rendering, interactivity, and dynamic range/loudness control). The standard is currently in its deployment phase for broadcasting and 3D audio services. At the same time, it serves as the basis for the subsequent MPEG-I audio standard that addresses efficient representation and rendering of VR and AR audios. For the rendering of 3DoF immersive content, MPEG-H already provides a full solution, while additional technology has to be developed for the 6DoF use case. Upon completion in 2023, MPEG-I immersive audio will establish the first long-term, stable format for a compressed representation of audio content for 6DoF VR and AR rendering and presentation. It will be used for consumer applications, such as broadcasting, streaming, and social VR. ■

Acknowledgment

The authors would like to thank Dr. Achim Kuntz for his careful review of drafts of this article.

REFERENCES

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

[2] A. D. Blumlein, "Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems," British Patent 394 325, Dec. 14, 1931.

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

[4] F. Rumsey, *Spatial Audio*. 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, pp. 1–15.

[6] 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.*, San Francisco, CA, USA, Nov. 2010, pp. 1–11, Paper 8248.

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

[8] 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 Wirelessly Netw. Pers. Devices*, Pohang, South Korea, Sep. 2011.

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

[10] *Information Technology—High Efficiency Coding and Media Delivery in Heterogeneous Environments—Part 3: 3D Audio*, Standard ISO/IEC 23008-3:2019, 2019.

[11] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H audio—The new standard for coding of immersive spatial audio," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 770–779, Aug. 2015.

[12] A. Silzle, S. George, E. A. P. Habets, and T. Bachmann, "Investigation on the quality of 3D sound reproduction," in *Proc. Int. Conf. Spatial Audio (ICSA)*, Detmold, Germany, 2011, p. 334.

[13] R. L. Bleidt *et al.*, "Development of the MPEG-H TV audio system for ATSC 3.0," *IEEE Trans. Broadcast.*, vol. 63, no. 1, pp. 202–236, Mar. 2017.

[14] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement and Virtual Reality*. New York, NY, USA: Springer, 2019.

[15] *Information Technology—MPEG Audio Technologies—Part 3: Unified Speech and Audio Coding*, Standard ISO/IEC 23003-3:2012, 2012.

[16] M. Neuendorf *et al.*, "The ISO/MPEG unified speech and audio coding standard—Consistent high quality for all content types and at all bit

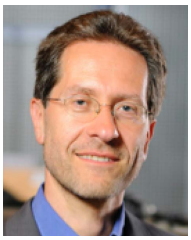
rates," *J. Audio Eng. Soc.*, vol. 61, no. 12, pp. 956–977, Dec. 2013.
[17] S. Disch *et al.*, "Intelligent gap filling in perceptual transform coding of audio," in *Proc. 141st AES Conv.*, Los Angeles, CA, USA, Oct. 2016, Paper 9661.
[18] J. Herre and D. Schultz, "Extending the MPEG-4 AAC codec by perceptual noise substitution," in *Proc. 104th AES Conv.*, Amsterdam, The Netherlands, 1998.
[19] P. Ekstrand, "Bandwidth extension of audio signals by spectral band replication," in *Proc. 1st IEEE Benelux Workshop Model Based Process. Coding Audio (MPCA)*, Leuven, Belgium, Nov. 2002, pp. 73–79.
[20] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1997.
[21] *Information Technology—Dynamic Adaptive Streaming Over HTTP (DASH)—Part 1: Media Presentation Description and Segment Formats*, Standard ISO/IEC 23009-1:2019.
[22] S. Füg *et al.*, "Design, coding and processing of metadata for object-based interactive audio," in *Proc. 137th AES Conv.*, Los Angeles, CA, USA, 2014.
[23] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, Jun. 1997.
[24] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *J. Acoust. Soc. Amer.*, vol. 116, no. 4, pp. 2149–2157, Oct. 2004, doi: 10.1121/1.1792643.
[25] A. Murtaza, J. Herre, J. Paulus, L. Terentiv, H. Fuchs, and S. Disch, "ISO/MPEG-H 3D audio: SAOC-3D decoding and rendering," in *Proc. 139th AES Conv.*, New York, NY, USA, 2015, Paper 9434.
[26] *Information Technology—MPEG Audio Technologies—Part 4: Dynamic Range Control*, Standard ISO/IEC 23003-4:2015.
[27] *N16584 MPEG-H 3D Audio Verification Test Report*. [Online]. Available: https://mpeg.chiariglione.org/standards/mpeg-h/3d-audio/mpeg-h-3d-audio-verification-test-report
[28] *Methods for the Subjective Assessment of Small Impairments in Audio Systems*, document ITU-R Rec. BS BS.1116-3, Feb. 2015.
[29] *Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems Also Known as Multi Stimulus Test With Hidden Reference and Anchor (MUSHRA)*, document ITU-R Rec. BS.1534-3, Oct. 2015.
[30] *Information Technology—Coded Representation of Immersive Media—Part 2: Omnidirectional Media Format*, Standard ISO/IEC 23090-2:2019.
[31] *Information Technology—High Efficiency Coding and Media Delivery in Heterogeneous Environments—Part 1: MPEG Media Transport (MMT)*, Standard ISO/IEC DIS 23008-1.
[32] *N18158 MPEG-I Audio Architecture and Requirements*. [Online]. Available: https://mpeg.chiariglione.org/standards/mpeg-i/immersive-audio-coding
[33] D. Alais and D. Burr, "The ventriloquist effect results from near-optimal bimodal integration," *Current Biol.*, vol. 14, no. 3, pp. 257–262, Feb. 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/14761661
[34] O. Rummukainen, T. Robotham, S. Schlecht, A. Plinge, J. Herre, and E. Habets, "Audio quality evaluation in virtual reality: Multiple stimulus ranking with behavior tracking," in *Proc. Conf. Audio Virtual Augmented Reality*, Redmond, WA, USA, Aug. 2018, pp. 1–10.
[35] [Online]. Available: https://unity.com
[36] [Online]. Available: https://cycling74.com/
[37] T. Robotham, O. Rummukainen, J. Herre, and E. A. P. Habets, "Evaluation of binaural renderers in virtual reality environments: Platform and examples," in *Proc. 145th AES Conv.*, New York, NY, USA, 2018, pp. 1–5.
[38] *General Methods for the Subjective Assessment of Sound Quality*, document ITU-T Rec. BS.1284.
[39] M. Perez-Ortiz and R. K. Mantiuk, "A practical guide and software for analysing pairwise comparison experiments," 2017, *arXiv:1712.03686*. [Online]. Available: http://arxiv.org/abs/1712.03686
[40] *Methods for Objective and Subjective Assessment of Quality*, document ITU-T Rec. P.800, Aug. 1996.

ABOUT THE AUTHORS

Schuyler R. Quackenbush (Senior Member, IEEE) received the B.S. degree from Princeton University, Princeton, NJ, USA, in 1975, and the M.S. and Ph.D. degrees from the Georgia Institute of Technology, Atlanta, GA, USA, in 1980 and 1985, respectively, all in electrical engineering.

In 1986, he joined AT&T Bell Laboratories, Murray Hill, NJ, USA, as a Member of Technical Staff and subsequently joined AT&T Laboratories, Florham Park, NJ, USA. In 2002, he founded Audio Research Labs, Scotch Plains, NJ, USA, an audio technology consulting company. In 2006, he was a cofounder of Lightspeed Audio Labs, Tinton Falls, NJ, USA, which creates Internet-based platforms for the collaborative creation of music. He is currently a Principal Consultant with Audio Research Labs. He is active in the area of standardization of audio coding algorithms and is the Convenor of the International Standards Organization MPEG Audio Coding Working Group. He was one of the authors of the ISO/IEC MPEG Advanced Audio Coding standard. His research interests are audio and speech processing.

Dr. Quackenbush has served as a member of the IEEE Technical Committee on Audio and Acoustic Signal Processing. He is also a Fellow of the Audio Engineering Society (AES) and the Co-Chair of the AES Technical Committee on Coding of Audio Signals.



Jürgen Herre (Senior Member, IEEE) joined the Fraunhofer Institute for Integrated Circuits (IIS), 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-coder (aka "MP3"). In 1995, he joined Bell Laboratories for a postdoctoral term working on the development of MPEG-2 Advanced Audio Coding (AAC). By the end of 1996, he went back to Fraunhofer IIS, Erlangen, to work on the development of more advanced multimedia technology, including MPEG-4, MPEG-7, and MPEG-D, where he is currently the Chief Executive Scientist of the Audio and Media Technologies Division. In September 2010, he was appointed as a Professor at the University of Erlangen-Nuremberg, Erlangen, and the International Audio Laboratories Erlangen, Erlangen.

Dr. Herre has served as a member of the IEEE Technical Committee on Audio and Acoustic Signal Processing. He is also a Fellow of the Audio Engineering Society. He is also an Active Member of the MPEG Audio Subgroup. He was a recipient of two Fraunhofer Awards in 1992 and 2004, respectively, the Eduard-Rhein Award in 2015, and the IEEE 2020 Industrial Innovation Award. He is also the Vice-Chair of the AES Technical Committee on Coding of Audio Signals and the Vice-Chair of the AES Technical Council. He has served as an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING.