J. Herre and S. Disch, "MPEG-I Immersive Audio Reference Model For The Virtual/Augmented Reality Audio Standard"
J. Audio Eng. Soc., vol. 71, no. 5, pp. 229–240, (2023 May).
DOI: https://doi.org/10.17743/jaes.2022.0074

MPEG-I Immersive Audio – Reference Model For The Virtual/Augmented Reality Audio Standard

JÜRGEN HERRE,[1] *AES Fellow* , **AND SASCHA DISCH**[2]
(juergen.herre@audiolabs-erlangen.de) (sascha.disch@iis.fraunhofer.de)

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

MPEG-I Immersive Audio is a forthcoming standard that is under development within the MPEG Audio group (ISO/IEC JTC1/SC29/WG6) to provide a compressed representation and rendering of audio for Virtual and Augmented Reality (VR/AR) applications with six degrees of freedom (6DoF). MPEG-I Immersive Audio supports bitrate-efficient and high-quality storage/transmission of complex virtual scenes including sources with spatial extent and distinct radiation characteristics (like musical instruments) as well as geometry description of acoustically relevant elements (e.g., walls, doors, occluders). The rendering process includes detailed modeling of room acoustics and complex acoustic phenomena such as occlusion and diffraction due to acoustic obstacles and Doppler effects as well as interactivity with the user. Based on many contributions, this paper reports on the state of the MPEG-I Immersive Audio standardization process and its first technical Reference Model architecture. MPEG-I Immersive Audio establishes the first long-term stable audio format specification in the field of VR/AR and can be used for many consumer applications such as broadcasting, streaming, social VR/AR, or Metaverse technology.

1 INTRODUCTION

In the area of new media technologies, Virtual and Augmented Reality (VR/AR) have been growing tremendously over the past few years and are on the way to wide-spread consumer application. Many essential key technology components for VR/AR have become available broadly step by step, such as

- Plenty of computing power even on mobile devices (including *Graphics Processing Unit [GPU]*)
- Good and affordable *Head Mounted Displays (HMDs)* and tracking devices [1–3]
- Convenient access to *personalized ear impulse responses* [4–6]
- Public availability of *VR/AR engines* [7,8]

Although the development of immersive games certainly has been an initial key driving force in development of VR/AR technology, a much wider range of useful and attractive applications is now considered, including edutainment, virtual concerts, virtual cultural exhibitions/performances, virtual conferences, social VR (i.e., a joint social experience of a group of people in a common virtual space, communicating and sharing a joint experience like a sports event, media presentation, etc.) all the way to next-generation broadcasting services that transmit VR/AR programs.

For many of these applications, however, it is of vital importance that the VR/AR content to be created can be represented in an open and long-term stable format that will still play back seamlessly in, say, 30 years from now, just the way we are used to from mp3 compressed audio files (or any of its standardized successor formats), to preserve the investment to be made into the new content creation chain.

Based on such considerations, the ISO/MPEG standardization group has initiated a new work item to address aspects of high-quality compressed representation and efficient rendering of VR/AR media several years ago. The audio part of this emerging MPEG-I standard is based on the latest predecessor technology (the ISO/MPEG-H 3D Audio standard) and encompasses a broad range of VR/AR-specific technology components. In the end, the new MPEG-I Immersive Audio standard will enable a compressed representation for efficient storage, streaming, and broadcasting of virtual content for many different applications including wireless transmission to mobile devices.

Along with other parts of MPEG-I (i.e., Part 12, "Immersive Video"; Part 5, "Visual Volumetric Video-Based Coding (V3C) and Video-Based Point Cloud Compression"; and Part 2, "Systems Support"), the suite of standards supports a complete audio-visual VR or AR presentation in which the user can navigate and interact with the virtual environment in 6DoF, that is with free spatial navigation (x, y, z) and user head orientation (yaw, pitch, roll).

This paper provides an overview of the ISO/MPEG-I Immersive Audio standardization process with special focus on the selected first Reference Model (RM) (i.e., baseline technology) as of the writing of this paper, i.e., after the 139th MPEG meeting in July 2022. It is structured as follows: Given that coding of spatial audio has been present in MPEG Audio for a considerable period of time, the existing MPEG Audio technology background in this field is briefly introduced. Then, the concept of the MPEG-I Immersive Audio work item is explained and its technical development environment introduced. Next, the process and outcome of the performance evaluation of several competitive VR/AR candidate systems are outlined. The selected best technology is described as the RM in terms of technology and capabilities. Finally, an outlook on the further development of the standard and some conclusions are provided.

2 MPEG SPATIAL AUDIO HISTORY

Starting around 1990, the standards provided by the MPEG Audio group have been successfully defining the state of the art in perceptual audio coding for more than three decades. Over time, the scope of the developed technologies expanded from simple coding of mono or stereo audio toward the representation and interactive rendering of spatial audio and VR/AR audio. This can be seen as a natural step in terms of immersivity and user interactivity. The most important spatial audio technologies are listed in the following:

The first widely used multichannel audio coder standardized by MPEG in 1997 is MPEG-2 Advanced Audio Coding (AAC) [9,10], delivering EBU broadcast quality at a bitrate of 320 kbit/s for a 5.1 signal. A significant step forward was the definition of MPEG-4 High Efficiency AAC (HE-AAC) [11] in 2002/2004, which combines AAC technology with bandwidth extension and parametric stereo coding and thus allows for full audio bandwidth at lower data rates. For carriage of 5.1 content, HE-AAC delivers quality comparable with that of AAC at a bitrate of 160 kbit/s [12]. Later MPEG standardizations provided generalized means for parametric coding of multichannel spatial sound: MPEG-D MPEG Surround (MPS, 2006) [13,14] and MPEG-D Spatial Audio Object Coding (SAOC, 2010) [15,16] allow for the highly efficient carriage of multichannel sound and object signals, respectively. Both codecs can be operated at lower rates (e.g., 48 kbit/s for a 5.1 signal).

MPEG-D Unified Speech and Audio Coding (USAC, 2012) [17,18] combined enhanced AAC coding with state-of-the-art full-band speech coding into an extremely efficient system, allowing carriage of, e.g., good-quality mono signals at bitrates as low as 8 kbit/s. Incorporating advances in joint stereo coding, USAC is capable of delivering further enhanced performance compared with HE-AAC also for multichannel signals. The MPEG-H 3D Audio codec [19,20] provides efficient and universal representation and rendering of all known production paradigms (channels, objects, and Higher-Order Ambisonics [HOA]) on arbitrary loudspeaker setups and binaural output. Although its waveform coding core is based on MPEG-D USAC, the novel unified combination of these production paradigms, together with the versatile rendering capabilities, create a universal spatial audio system.

3 THE MPEG-I IMMERSIVE AUDIO WORK ITEM

Dating back to early 2017, initial discussions on developing a specification for Audio for Virtual and Augmented Reality applications at ISO/MPEG Audio were triggered by the general MPEG-I discussions that address both visual and audio representations as well as carriage of VR/AR content in a joint bitstream or computer file. It was decided to split the project into two phases with increasing level of sophistication:

Phase 1 addresses VR/AR applications with three degrees of freedom (3DoF), i.e., representation and rendering of VR/AR material with respect to the user's arbitrary head movements along all three possible axes ("pitch, yaw, roll").

Phase 2 additionally allows free navigational movement of the user in all three axes of the 3D coordinate space within the virtual scene (called six degrees of freedom [6DoF]). In this way, the user can explore virtual scenes extensively and scene details that are hidden behind visible objects are revealed. Although 3DoF VR merely requires the reproduction of media on a sphere around the user's head, 6DoF VR rendering requires detailed knowledge on the entire scene and rendering of the scene components that are perceptible from the current user position and orientation.

For the audio part, it was found that the MPEG-H 3D Audio specification already included all required technical means for covering Phase 1 (3DoF Audio): It provides a high-quality representation of the associated audio waveforms (channels, objects, and HOA) [19,20] and a high-quality 3D rendering process, including the rotation of the soundscape as it is required for head-tracked binaural audio rendering to headphones. In 2017, MPEG specified the Omnidirectional Media Format [21] (OMAF) that allows VR movies, in which a user can view with a 360-degree "room-sized" screen and can watch, e.g., VR sports and concerts as a virtual participant.

Thus, all further audio standardization activities focused on Phase 2 (6DoF Audio) technology from this point. As it is common for ISO MPEG standards development, a set of requirements and a generic architecture were devised first [22]. Then, a "Call for Proposals (CfP)" was issued as an open call for technology proposals that are evaluated against each other to determine the best winning technol-

ogy (see Sec. 5.2). This involves provision of a VR/AR test bed, creation of suitable and critical test content, and selection of a methodology for reliable subjective quality evaluation of the rendered content and a set of procedures for conducting the competitive evaluation and determining its winner. These aspects are described in more detail in the next two sections. Then, the winning technologies are combined into an RM, as described in Sec. 6, both in a textual and a software representation, which forms the basis of all subsequent further technical improvements.

4 DEVELOPMENT ENVIRONMENT

Although the development of previous MPEG Audio coding standards merely required an encoding and decoding process/software that is applied to original audio waveforms (e.g., WAV files) in an offline fashion, the development and evaluation of an audio standard for representation and rendering of VR/AR audio is a much more complex project. Such a project is not feasible without a number of preparation steps that enable to set up of suitable VR/AR test material, process it by the technology under development, present it to subjects and evaluate the technology's performance. Compared with previous MPEG audio standards, the two most notable differences are as follows:

- VR/AR is an inherently multimodal experience that relies on the coherent simultaneous presentation of both audio and visual components. Additionally, it involves the user's sense of self-movement and body position (proprioception); and
- VR/AR is not merely presented to the user/subject for static consumption. Instead, it is a very dynamic/interactive experience due to user movement (3DoF/6DoF) or many other types of user interactions with the content. This requires that the technology reacts in real-time to the user, i.e., implements a real-time (rather than offline) processing/rendering.

The remainder of this section describes the different component of the development and evaluation environment that have been conceived within the MPEG Audio group to provide an suitable environment for the standards development process.

4.1 Audio Scene Description

For VR and AR, a description of the (virtual and/or real) acoustic world has to be provided as a ground truth for all subsequent processing. Although it was recognized within MPEG Audio that there might be other efforts for defining more or less universal scene description languages, it was found that there was none that described all the necessary acoustic features in sufficient detail and rigor. Consequently, as a first key ingredient, an *Encoder Input Format (EIF)* specification was created as a means of expressing detailed and specific properties of the intended audio scene [23]. The EIF specification was developed over a period of several years as an XML-based description language that includes constructs such as

- *Sources* using waveforms from channels, objects, and HOA as they are already supported in MPEG-H Audio in combination with a rich palette of time-varying properties (position/orientation, size, radiation characteristics, rendering options). A source with size can be either be *homogeneous* (i.e., with a constant sound characteristics over its extent) or *heterogeneous* (i.e., with varying sound characteristics in different extent regions, defined by different associated waveforms).
- *Acoustic Environments* (AEs) describing the late reverberation properties of scene parts like rooms or outside environments.
- *Geometry* describes acoustic enclosures, walls, occluders, etc. together with their acoustic material properties (transmission, reflection/scattering, dissipation, etc.). These constructs can be used to accurately compute early reflections, diffraction, and occlusion phenomena.
- *User Interactivity* specifications enable the user to interact with the virtual environment in various ways through input from controller devices, 6DoF movement, or other external processes like physical simulation engines.

A special challenge arises in the context of AR scene description wherein the acoustic parameters of the real listening room should be taken into account for the rendering process (e.g., such that the virtual acoustics of an augmented component match the acoustics of the real listening space). Because these parameters are unknown at the time of scene authoring, they need to be conveyed to the renderer at run time. To this end, and because the kind of information that needs to be represented is similar to the scene description, a subset of the EIF specification was chosen to define a *Listener Space Description File (LSDF)* that contains the individual listening space specific information and is read by the renderer at initialization time. Examples for such information are the shape and location of the listening room walls and the room's acoustic parameters.

4.2 Audio Evaluation Platform

The second key ingredient for developing and evaluating VR/AR technology is the availability of a reasonably simple but effective VR/AR test bed. To this end, the *Audio Evaluation Platform (AEP)* was proposed [24] and collaboratively refined further [25].

The AEP (see Fig. 1) is a real-time 6DoF VR/AR evaluation environment that uses a Personal Computer (PC) with a high-performance graphics card. HMDs are used for visual rendering and tracking—HTC Vive Pro for VR and Microsoft HoloLens(2) for AR. The visual component is represented by the Unity software that renders computer graphics generated scene content and interfaces with the HMD. For real-time audio processing, Max/MSP is used

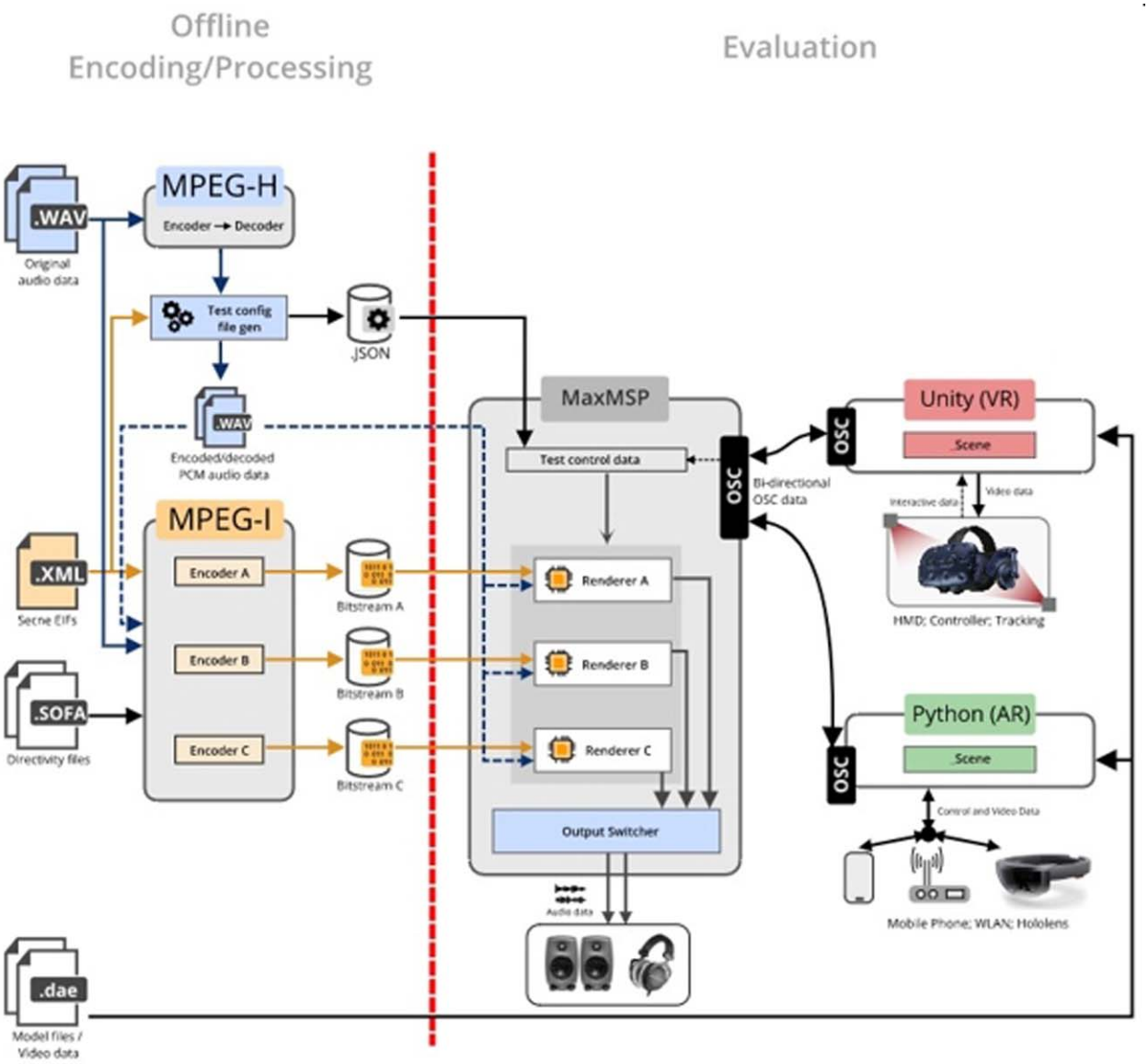

Fig. 1. Overview of the Audio Evaluation Platform (reproduced from [25]).

as a host software in which several experimental audio rendering technologies can run as Max external binaries. Visual (Unity) and audio (Max/MPS) communicate via Open Sound Control (OSC) messages.

All individual audio renderers receive the common (preen/decoded) waveforms belonging to the scene and general control data such as user position/orientation or controller data. Each renderer to be evaluated receives its individual metadata bitstream, which has been created in an offline way by their corresponding encoders from the EIF and the waveforms and delivers its audio output (e.g., two-channel binaural audio) for assessment. For comparative evaluation, the monitored renderer output can be switched on-the-fly by a virtual Graphical User Interface (GUI) to allow direct comparison. This requires that all renderer plug-ins run correctly in real time and in parallel on the host computer. The overall sequence of a complete test session is configured by a .JSON script.

4.3 Test Content

The definition of the EIF specification and the AEP enabled the creation of test content that includes visual (Unity) and audio (EIF) components and is used to exercise the technology under development by evoking certain acoustic phenomena that have to be represented/rendered faithfully/in a plausible way. Examples include fixed and moving audio sources at individual locations and distances, Doppler effects, outdoor and indoor environments with different acoustics and transitions between them, sources with size and directional radiation patterns, diffraction and occlusion through various acoustically effective obstacles, etc.

Over time, the MPEG Audio group developed an inventory of essential acoustic attributes of the envisioned MPEG-I Audio functionalities as well as a pool of test scenes that address these attributes. As an example, Fig. 2 shows the "Park" test scene, which can be used for evaluation of several acoustic attributes:

Fig. 2. The "Park" test scene with a sized fountain, and bushes, columns and walls as occluders.

- The center of the scene features a fountain that is a sound source with considerable *size (extent)*. Plausible rendering of this sound source with consistent audio and visual parts is evaluated including the rendering behavior when *stepping into the fountain* from the outside.
- Depending on the user position, the sound of the fountain can be *occluded fully* (behind the distant stone walls) or *partially* (behind the bushes).
- Also, for areas in which occlusion happens, the natural *diffraction* effect needs to be rendered adequately (behind walls, bushes, and pillars).

Eventually, a selection of such scenes served as a test set for comprehensive and critical quality evaluation in the competitive phase of the standards development as explained in the subsequent sections.

4.4 Binaural Rendering

Because the majority of VR/AR applications is expected to use binaural rendering to headphones, Head-Related Impulse Responses (HRIRs) or Binaural Room Impulse Responses (BRIRs) [26] are used for convolution with the rendered sound. Either generic or individually measured impulse responses could be applied, depending on availability. To support all possibilities, existing MPEG Audio standards that provide binaural rendering (MPEG Surround [13,14], MPEG-D SAOC [15,16], and MPEG-H 3D Audio [19,20]) support an interface to load impulse responses into the decoder/renderer rather than mandating a certain impulse response set. This concept was carried forward also with MPEG-I Immersive Audio. For practical experimentation and quality evaluation in the standardization process, the FABIAN HRIR database [27,28] with a diffuse field equalization was chosen because of its good perceptual characteristics and high spatial sampling density.

5 PERFORMANCE EVALUATION

For the development of a technical standard, it is indispensable to rigorously evaluate competing technology proposals/alternatives to choose the best available technology for standardization. This requires solid and well-understood methods for quality assessment. Specifically in the field of audio, subjective quality evaluation by listening tests is the state-of-the-art evaluation method.

5.1 Subjective Test Methodology

Compared with previous MPEG audio standards, the quality evaluation of calls for a different test paradigm beyond testing merely subjective audio quality. Quality evaluation of VR/AR needs to account for three involved human senses and their multimodal interaction, i.e., audio/hearing, visual/seeing, and the sense of body motion/proprioception. It is important to understand that in contrast to previous MPEG Audio listening tests, no well-defined reference condition (formerly: uncompressed audio) exists for auralized VR/AR audio content. Three subjective test methodologies were proposed in the course of the preparation effort and can be run as software on the AEP:

- An virtualized extrapolation of the well-known MUSHRA test methodology [29] and the recently finalized ITU-R Recommendation BS.2132 [30] provides a multistimulus presentation/multiway comparison of several renderer outputs via GUI [31] (see Fig. 3) and does not make use of an explicit reference condition. Note that operation without reference leads to a different use of the verbally anchored 100-point scale for the same conditions (see Figure 12 in [32]). Because of computational limitations (all renderers evaluated on one evaluation screen need to run seamlessly in parallel), this test method was not employed for the initial (CfP) comparative quality evaluation but rather, in the subsequent collaborative improvement phase of the standard.
- As an alternative, a virtualized simple pairwise comparison of two renderer conditions (A/B test) on a seven-point ITU comparison scale [33] was proposed (see Fig. 4). To summarize many individual comparison results into an overall quality score for each tested renderer output, the “Thurston Case V” model [34] is applied, which discards the actual value of the comparison score in favor of just the score sign and delivers an overall score in units of *Just Objectionable Differences (JODs)*. Because of its low computational demands (only two systems have to run in real-time at once), the A/B comparison test methodology was used in the competitive MPEG-I Audio large-scale comparative quality evaluation.
- Finally, a multi-attribute assessment method according to ITU-R Recommendation BS.2132 [30] has been proposed that gathers individual subject scores for a set of perceptual attributes of each condition (Plausibility, Consistency, Externalization, Basic Audio Quality). All renderer outputs are presented separately (i.e., without direct comparison to other conditions). It is not fully understood how

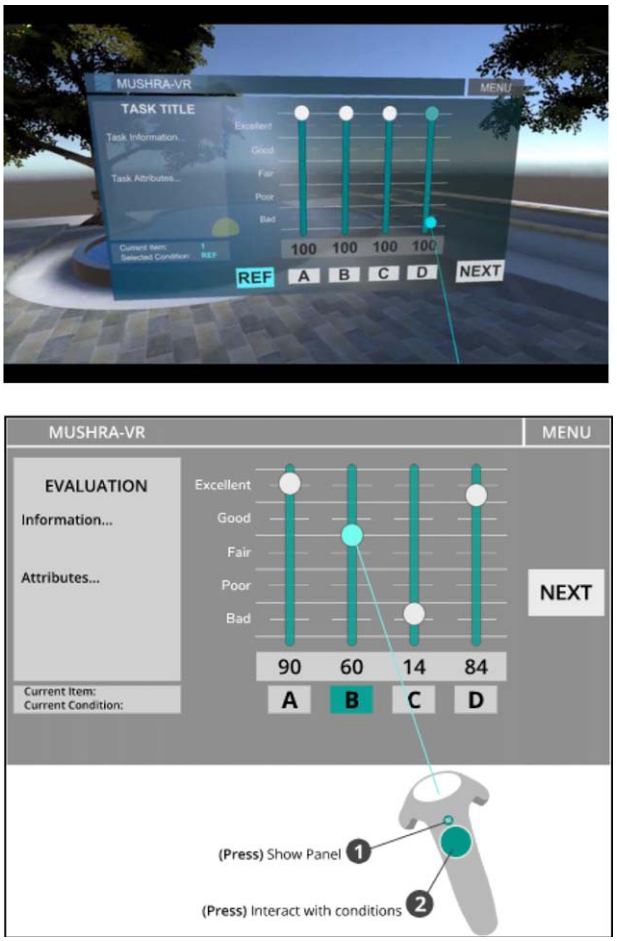

Fig. 3. Multistimulus grading panel. Top view: appearance in the scene (this predecessor version still has an "REF" button). Bottom view: Panel view and controller usage.

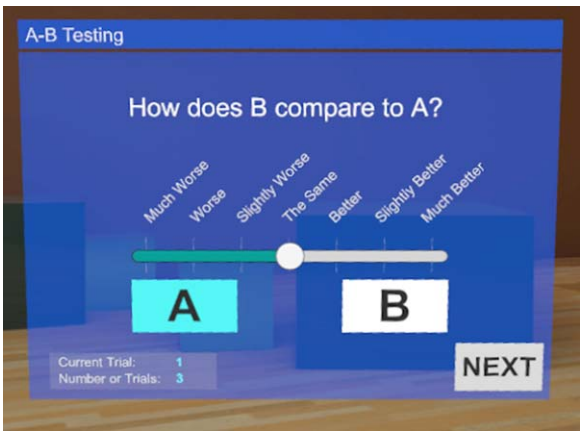

Fig. 4. A/B comparison grading interface (reproduced from [36]).

a subjectively valid overall figure of merit can be derived from the single attribute scores. This test methodology can also be used in the collaborative improvement phase of the standard.

5.2 MPEG-I Audio Call for Proposals

After the preparatory phase in which the MPEG-I Audio requirements, architecture, and the previously described components (EIF, AEP, test content and subjective quality evaluation methodology) were developed, a competitive phase followed: In the MPEG-I CfP [35], an open call to the outside world for MPEG-I Audio candidate technology was issued that included the encoding process, delivery of compressed scene metadata bitstreams, and renderer executables. Up to two proposals could be submitted by each proposing party. The proposals were evaluated according to previously defined evaluation criteria against each other, and the winners of the competitive evaluation then form the MPEG-I Audio RM, i.e., the baseline for all further technology development.

Three tests were specified to evaluate different technology aspects:

- Test 1: VR Baseline Technology
  Objective: Identify best "base" technology for RM0 that is common for VR and AR
  Details: Use HTC Vive Pro and headphones. The test material (14 test scenes) contains object, channel, and 3DoF HOA sound sources (i.e., sound sources with customary HOA decoding) as it is already available in MPEG-H Audio. Besides the overall winner by virtue of best subjective quality, there is the possibility for an additional low complexity or low bitrate category winner.
- Test 2: AR Test
  Objective: Identify best AR-specific technology
  Details: Use Microsoft HoloLens (2) and headphones. The test material (seven test scenes) contains object and channel sound sources. The winning AR specific technology will be merged into the Test 1 "baseline technology".
- Test 3: 6DoF HOA Test
  Objective: Identify best 6DoF HOA technology
  Details: Use HTC Vive Pro and headphones. The test material (four test scenes) contains HOA sources with interior/exterior rendering and multipoint HOA. The winning specific technology will be merged into the Test 1 "baseline technology".

The CfP process followed a strict time schedule: In April 2021, the CfP Document [35] was issued, followed later by test and evaluation guidelines and supplementary documentation. The delivery deadline of technology submissions (bitstreams and audio renderer executable) was November 10, 2021. Subsequently, the subjective tests were to be conducted using the AEP in a number of test laboratories that had volunteered for this effort. After the finalization of the subjective evaluation, consolidation of the results and selection of the RM technology was carried out in January 2022.

A total of twelve laboratories located across Europe, Asia, and North America registered as test sites for the CfP subjective evaluation. Fourteen working

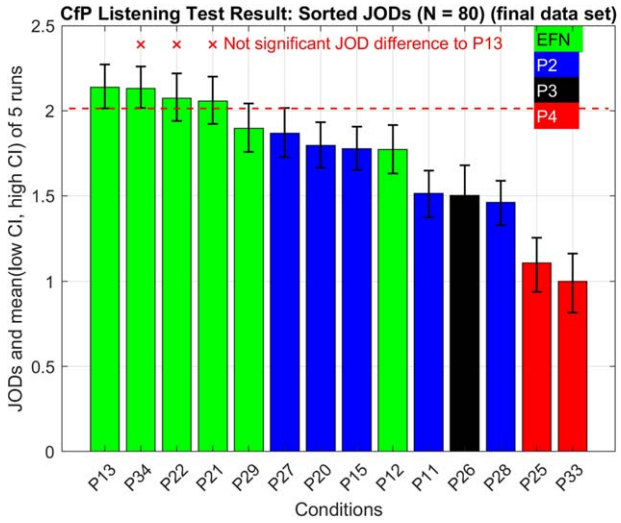

Fig. 5. Sorted system performance on a Just-Objectionable-Differences (JODs) scale for Test 1 after [37]. One JOD unit corresponds to 75% probability of one condition being better than another. Joint submissions are color-marked, and P2, P3, and P4 denote submissions other that from Ericsson-Fraunhofer-Nokia (EFN).

technology proposals were submitted to the CfP evaluation and tested by the test sites in individualized test sessions.

5.3 Evaluation Results

Subjective evaluations were conducted for all three tests. According to the previously defined evaluation procedures, the main criterion/figure of merit for the competition was subjective quality as measured by the A/B pairwise comparison test and summarized into a single figure of merit using the Thurston Case V model (see Sec. 5.1). The evaluation results were as follows [37]:

- Test 1: VR Baseline Technology(12 Test Labs, 80 valid subjects)
  Objective: Identify best "base" technology for RM0 that is common for VR and AR
  Outcome: The four top-ranked submissions were slight variations of a joint submission by Ericsson/Fraunhofer IIS/AudioLabs/Nokia (EFN, green in Fig. 5). No "category winner" was found to fulfill the low-complexity category winner criterion (complexity $\leq$20% of the best-performing proposal at still good quality). The criterion for a "low bitrate category winner" (bitstream size $\leq$5/12 of the best performing proposal at still good quality) was met by a joint submission of Dolby, Philips and Qualcomm (P27), see Fig. 5.
- Test 2: AR Test (6 Test Labs, 37 valid subjects)
  Objective: Identify best AR-specific technology
  Outcome: The best submission was a variation of the joint submission by Ericsson/Fraunhofer IIS/AudioLabs/Nokia. The winning AR-specific technology was merged into the Test 1 "baseline technology".
- Test 3: 6DoF HOA Test (9 Test Labs, 54 valid subjects)
  Objective: Identify best 6DoF HOA technology
  Outcome: The best submission was a variation of the joint submission by Ericsson/Fraunhofer IIS/AudioLabs/Nokia. The winning specific technology was merged into the Test 1 "baseline technology."

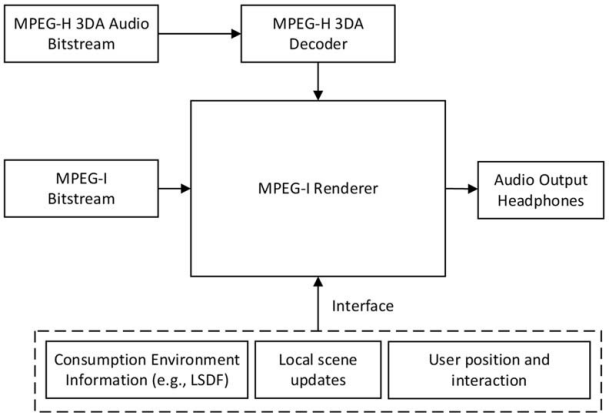

Fig. 6. MPEG-I Audio Top Level Architecture.

All winning technology components from Ericsson/Fraunhofer IIS/AudioLabs/Nokia were put together and provided to MPEG Audio as first version of the MPEG-I Immersive Audio Reference Model in textual form ("Working Draft") and software ("Reference Software") in April 2022. Some Test 1 low-bitrate technology components are merged into this first version by October 2022.

6 THE MPEG-I AUDIO REFERENCE MODEL

MPEG Audio RMs describe the entire compression chain, comprising (waveform and metadata) encoder, the bitstream format, and the decoder/renderer. While at MPEG, the encoding step is traditionally left "non-normative" to allow compatible improvements in encoding, both bitstream format and decoder/renderer are fixed and constitute the normative part of the RM, thus ensuring interoperability across different implementations. In the following, the first version of the MPEG-I Immersive Audio RM is described from the renderer side, i.e., the VR/AR-specific technology as of July 2022 (further additions may happen during Core Experiments, see Sec. 7).

6.1 Renderer Overview

Fig. 6 shows the top-level architecture of the MPEG-I Immersive Audio decoding/rendering. The waveforms belonging to an VR/AR scene are taken from an MPEG-H 3D Audio bitstream and decoded into pulse code modulation (PCM) signals, which forms one input to the MPEG-I Renderer. The second input to the renderer is the MPEG-I Au-

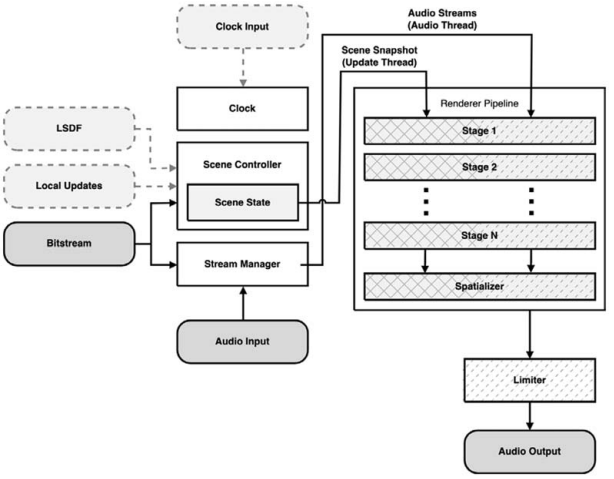

Fig. 7. Overview of MPEG-I Audio Renderer.

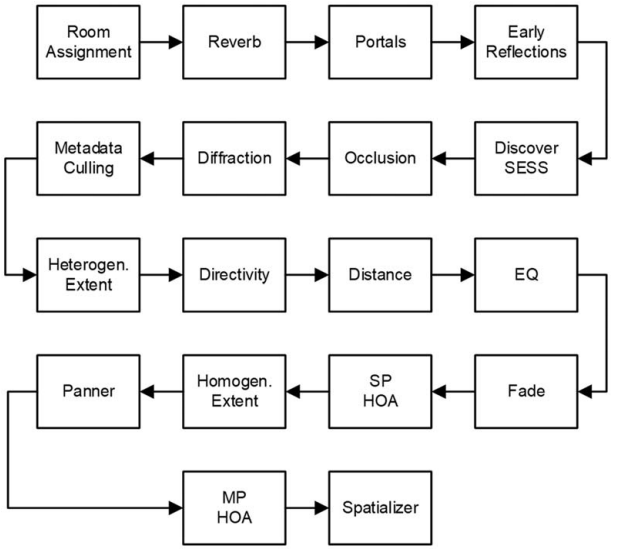

Fig. 8. MPEG-I Audio Renderer Pipeline and its Renderer Stages.

dio metadata bitstream which describes all further VR/AR scene information. Additionally, the renderer also receives dynamic updates on the user position and orientation, on scene updates and for AR, information on the consumption environment (describing e.g., its acoustic and geometric properties). The latter can be described in a *LSDF* file.

The core part of the scheme is the MPEG-I Audio Renderer, which is further detailed in Fig. 7.

The renderer has as inputs the bitstream, information on local updates (including user position and orientation), and for AR, information on the consumption environment (LSDF). It includes the *Scene State* as a central element that describes the current state of all 6DoF metadata including audio elements and geometry. The *Scene Controller* handles all internal and external changes to the Scene State data. The *Stream Manager* is an interface for the renderer to access audio streams that are associated with an audio element. Finally, a clock component is available as a time reference to the system.

The actual rendering happens by serial processing of audio elements in the *Renderer Pipeline*, which comprises several *Renderer Stages* that represent different aspects of virtual acoustic processing in a predefined order. Specifically, the concept is that *Render Items (RIs)* that represent audio elements in the renderer pipeline propagate through this pipeline and are processed or modified on their way. RIs may originate directly from audio sources (primary RIs) but also from other secondary sources like mirror image sources for modeling early reflections or substitute sources for diffraction rendering (secondary RIs).

Most modifications, like gain or equalizer (EQ) changes, happen in the parametric metadata domain and alter metadata rather than directly processing the actual audio samples. Audio processing is mostly carried out at the end of the pipeline and during spatialization. Therefore, these functions are implemented in two threads: The operations that regard the snapshot of the scene state are carried out at the low control rate in a control thread (e.g., control rate 50 Hz), while operations on the actual audio streams run in an audio thread at the audio block rate (nonoverlapping blocks of 256 time domain PCM samples at $f_s = 48$ kHz).

One can say that the task of the Renderer is to auralize the current Scene State as it has been updated by the Scene Controller. At the end of the audio processing, a spatializer renders all RIs to their target locations (for binaural output, this means applying the corresponding HRIRs/BRIRs). Finally, the output is subjected to a limiter in order to avoid clipping in extreme circumstances (e.g., listening to a very loud audio source in close proximity). The renderer pipeline is described next.

6.2 Renderer Pipeline and Stages

The stages of the Renderer pipeline are shown in Fig. 8. Each stage has access to the list of RIs and may create/add or delete/inactivate RIs in this list. Although carried out sequentially in a certain execution order, the nature of the various rendering stages is quite different, and subsequent stages do not necessarily work on the same rendering aspects. Although some set up aspects of further processing without touching any audio sample, others (typically later in the pipeline) carry out actual audio processing tasks. All relevant acoustic effects are implemented in this processing chain, which is subsequently described:

- Room Assignment
  The room assignment stage updates metadata of RIs to reflect the current scene configuration regarding its AEs. This is necessary for a relevance check by the subsequent Late Reverb and Portal stages as well as others. A default AE can be applied in scene parts without explicitly defined AE.
- Reverb
  The reverb stage generates late reverberation for AEs using efficient Feedback Delay Network (FDN) reverberators. In the configuration used in the Audio Reference Model, the FDN is parametrized to use

15 delay lines and an optimized feedback matrix. For AR rendering, the renderer optimizes the FDN parameters based on the LSDF of the listening space (e.g., such that the virtual acoustics of the augmented components matches the acoustics of the real listening environment).

- Portals
  The portals stage models the propagation of late reverb from one AE into another AE (practically: from one reverberant room to a connected one). This is done by substituting the entire reverberant room by a sound source with a homogeneous extent (see next) that has the room's enclosure and volume.
- Early Reflection
  The early reflection stage calculates specular surface reflections (e.g., from walls or other acoustically effective scene elements), which lead to the early reflections preceding the late (stochastic) reverberation. The reflections are derived from the surface geometry and properties by an image source model (including source visibility check) in which reflection and transmission filters model the sound propagation.
- Discover SESS
  This is a helper stage for rendering *Spatially Extended Sound Sources (SESSs)*. Its output is used in later stages like Occlusion and Homogeneous Extent (see next).
- Occlusion
  The occlusion stage carries out a ray-cast from the listener position to the acoustically relevant geometries (obstacles) to determine nonoccluded sound paths. The stage also supports opaque/partially transparent objects and extended sound sources (sources with size).
- Diffraction
  The diffraction stage derives one or more relevant diffraction paths from geometry information and line-of-sight occlusion checks to model sound propagation with diffraction, e.g., around corners by adding secondary RIs at the diffraction edge closest to the listener.
- Metadata Culling
  The metadata culling stage deactivates RIs that become inaudible due to very low gain (attenuation or occlusion) and thus reduces computational complexity.
- Heterogeneous Extent
  The heterogeneous stage renders sources with so-called heterogeneous extent with two to nine source waveform channels, as well as HOA sources with extent and both an interior and exterior representation (e.g., for HOA sources that represent an extended spatial area, and where a listener can move outside and inside that area and perceive the sound field accordingly).
- Directivity
  The directivity stage implements source directivity, i.e., a direction-dependent radiation pattern as it is found in many natural sound sources like a human talker or musical instruments. The stage adds the corresponding direction-dependent equalization by adding them to the relevant RIs' central EQ fields that accumulate all EQ effects to be applied later on by the EQ stage (see next).
- Distance
  The distance stage implements the effects of the sound travelling over a certain distance from the source to the listener including three aspects: First, a physically accurate *Propagation Delay* is modeled using a variable delay line with subsample spline interpolation and smoothing. Second and third, the signal attenuations due to the geometrical spreading of the source energy (e.g., the energy of a point source decays with 1/r) and due to air absorption are determined and accumulated in the central EQ fields of the RIs.
- Fade
  The fade stage applies fade-in/out ramps over time on audio signals when their RIs are activated/deactivated, e.g., by metadata culling or teleportation. This fade-in/out operation happens in a synchronized way between several stages.
- Single-Point HOA
  The single-point HOA stage loads HOA sources in Equivalent Spatial Domain (ESD) format and transforms them to HOA format for subsequent processing with a Magnitude Least-Squares (MagLS) decoder to render binaural signals.
- Homogeneous Extent
  The homogeneous extent stage efficiently renders SESS with a size and spatially uniform sound characteristic directly to binaural output. This is done by first determining the binaural cues that the SESS generates at the ear signals and then imposing theses cues onto the source input signal, optionally using decorrelation processing [38].
- Panner
  The panner stage positions ("pans") the late reverb outputs of the reverb stage into the output sound image in head tracked mode using Vector Base Amplitude Panning (VBAP) [39] with additional controls, such as spatial spread.
- Multi-Point HOA
  The multipoint HOA stage renders a set of HOA sources into a joint sound image and creates a 6DoF listening experience with binaural output from them. To this end, spatial metadata is used from an encoded bitstream.
- Spatializer
  The spatializer stage positions the point source RIs into the binaural output signals. To this end, it selects the nearest HRIR in the provided HRIR dataset for a given RI position and convolves its associated audio waveform with the HRIR stereo filter through a fast partitioned convolution. A cross-fade is used to ensure smooth filter changes between positions.

6.3 Bitstream

MPEG-I Immersive Audio extends the existing MHAS stream transport format, introduced with MPEG-H 3D Audio [19,20], to additionally carry MPEG-I 6DoF scene data. Three newly defined MHAS packets integrate MPEG-I 6DoF scene data in the form of a Scene Config packet, a Scene Update packet, and a Scene Payload packet.

The lightweight *SceneConfigpacket* provides all relevant initialization information to the renderer. The *Scene Update packet* communicates scene updates and update conditions that are known when the stream starts and also interactive scene changes that are unknown at stream startup. The *Scene Payload packet* accommodates the main load of MPEG-I metadata. It contains directivity data, geometry descriptions, and all other metadata for controlling rendering stages such as Reverb, Early Reflections, or Diffraction.

The three-packet concept allows VR/AR implementations for different use cases like client–server applications or broadcasting. The distribution of all necessary payloads among several Scene Payload packets can be flexibly adapted to the use case at hand, e.g., to interleave different MHAS packet types within a stream. For stream linking and synchronization, existing MHAS mechanisms can be used.

7 FURTHER EVOLUTION AND OUTLOOK

After the initial preparatory phase, the competitive phase (CfP) of the MPEG-I Immersive Audio standardization provided the technology baseline (RM) for all further development. Furthermore, at the time of writing, a merge process for the low-bitrate technology winning in the CfP evaluation is in progress, including technology aspects such as efficient metadata compression, voxel-based occlusion and diffraction [40] and compact reverb parametrization.

Also, like in all ISO/MPEG standards, a collaborative phase for further refinement and extension of the RM follows. This is conducted by means of "Core Experiments." Each Core Experiment addresses a specific functionality, either existing or still missing in the RM, and can be submitted by any party inside the MPEG Audio group. Following a well-defined procedure, the submitted proposal has to demonstrate its merit relative to the existing RM performance in order to be admitted to the updated RM (standards text and reference software). This involves cross-checks by independent parties.

Clearly, a number of required or desired functionalities could not be tested during the MPEG-I CfP evaluations due to limitations in time and resources. These missing functionalities include the following:

- Rendering on loudspeakers (as opposed to headphones)
- Client-server based streaming operation with a backchannel from the client to the server
- Social VR (meaning that several participants can interactively enjoy, e.g., sports events together in a virtual room while being able to communicate in real-time with each other. This involves low-delay communication capabilities in both low-bitrate audio coding and VR/AR rendering.
- Handling of large-scale scene landscapes (e.g., entire cities with many buildings and rooms) that can be loaded/updated dynamically as "sub-scenes").

All technology refinement is expected to be finalized in 2023, and the final result will be tested in a *Verification Test* to benchmark its performance. This is the basis for including the MPEG-I Immersive Audio specification into upcoming VR/AR applications and application standards.

8 CONCLUSIONS

To facilitate high-quality bitrate-efficient representation of audio for VR/AR in 6DoF applications, the ISO/MPEG Audio group has been working on developing the MPEG-I Immersive Audio specification since 2017. After several years of extensive preparations for establishing the technical requirements and a suitable VR/AR real-time development and evaluation environment, 14 technology proposals were evaluated in a large-scale subjective testing effort involving 12 test laboratories worldwide. As a result, the first RM architecture was established from the best-performing proposals (Fraunhofer IIS/AudioLabs, Nokia, Ericsson) and low-bitrate technologies (Dolby, Philips, Qualcomm). This architecture is the basis for further collaborative refinement of the standard until its anticipated finalization in 2023. Based on MPEG-H 3D Audio, the developed technology offers a compact representation of VR/AR audio and an efficient high-quality rendering process that includes accurate modeling of many complex acoustic phenomena.

MPEG-I Immersive Audio establishes the first long-term stable audio format specification in the field of VR/AR and can be used for many consumer applications like broadcasting, streaming, social VR, or Metaverse technology.

9 ACKNOWLEDGEMENTS

The authors would like to express their sincere thanks to Simon Schwär, Andreas Silzle, Tracy Harris, Nils Peters, and Alexander Adami for helping to prepare this article. The presented technical work is the result of a large-scale effort of the teams at Ericsson, Fraunhofer IIS/International Audio Laboratories Erlangen, and Nokia. Additional technology contributions come from the teams at Dolby Laboratories, Philips, and Qualcomm. Special acknowledgement goes to Dr. Schuyler Quackenbush for his diligent leadership of the standardization process and to the entire ISO/MPEG Audio group (ISO/IEC JTC1/SC29/WG6).

10 REFERENCES

[1] HTC, "Vive Pro," https://www.vive.com/us/product/vive-pro/ (accessed March 30, 2023).

[2] Microsoft, "Microsoft HoloLens 2," https://www.microsoft.com/en-us/hololens (accessed March 30, 2023).

[3] Magic Leap, Inc., "Home," https://www.magicleap.com (accessed March 30, 2023).

[4] Apple, "Listen with Personalized Spatial Audio for AirPods and Beats," https://support.apple.com/en-us/HT213318 (accessed January 2023).

[5] Sony, "How to analyze your ear shape (360 Reality Audio)," https://www.sony.com/electronics/support/articles/00233341 (accessed September 2022).

[6] Genelec, "Aural ID," https://www.genelec.com/aural-id (accessed January 2023).

[7] Unity Technologies , "Unity Real-Time Development Platform," https://unity.com (accessed March 30, 2023).

[8] Epic Games, Inc., "The Most Powerful Real-Time 3D Creation Tool - Unreal Engine," https://unrealengine.com (accessed March 30, 2023).

[9] M. Bosi, K. Brandenburg, S. Quackenbush, et al., "ISO/IEC MPEG-2 Advanced Audio Coding," *J. Audio Eng. Soc.*, vol. 45, no. 10, pp. 789–814 (1997 Oct.).

[10] ISO/IEC, "Information Technology - Generic Coding of Moving Pictures and Associated Audio Information – Part 7: Advanced Audio Coding (AAC)," *Standard 13818-7:1997* (1997 Dec.).

[11] J. Herre and M. Dietz, "MPEG-4 High-Efficiency AAC Coding [Standards in a Nutshell]," *IEEE Signal Process. Mag.*, vol. 25, no. 3, pp. 137–142 (2008 Apr.). https://doi.org/10.1109/MSP.2008.918684.

[12] EBU, "EBU Evaluations of Multichannel Audio Codecs," Tech. Rep. 3324 (2007 Sep.). https://tech.ebu.ch/docs/tech/tech3324.pdf.

[13] J. Hilpert and S. Disch, "The MPEG Surround Audio Coding Standard [Standards in a Nutshell]," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 148–152 (2009 Jan.). https://doi.org/10.1109/MSP.2008.930433.

[14] ISO/IEC, "Information Technology — MPEG Audio Technologies — Part 1: MPEG Surround," *Standard 23003-1:2007* (2007 Feb.).

[15] J. Herre., H. Purnhagen, J. Koppens, et al., "MPEG Spatial Audio Object Coding – The ISO/MPEG Standard for Efficient Coding of Interactive Audio Scenes," *J. Audio Eng. Soc.*, vol. 60, no. 9, pp. 655–673 (2012 Sep.).

[16] ISO/IEC, "Information Technology — MPEG Audio Technologies — Part 2: Spatial Audio Object Coding," *Standard 23003-2:2010* (2010 Oct.).

[17] M. Neuendorf, M. Multrus, N. Rettelbach, 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 (2013 Dec.).

[18] ISO/IEC, "Information Technology — MPEG Audio Technologies — Part 3: Unified Speech and Audio Coding," *International Standard 23003-3:2012* (2012 Apr.).

[19] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties "MPEG-H Audio—The New Standard for Coding of Immersive Spatial Audio," *IEEE J. Sel. Top. Signal Process.*, vol. 9, no. 5, pp. 770–779 (2015 Aug.). https://doi.org/10.1109/JSTSP.2015.2411578.

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

[21] ISO/IEC, "Information Technology — Coded Representation of Immersive Media — Part 2: Omnidirectional Media Format," *Standard 23090-2* (2019 Jan.).

[22] ISO, "MPEG-I Audio Architecture and Requirements," ISO/IEC JTC 1/SC 29/WG 6 Document N18158 (2021 Jan.).

[23] ISO, "MPEG-I Immersive Audio Encoder Input Format," ISO/IEC JTC 1/SC 29/WG 6 Document N0054 (2021 Apr.).

[24] T. Robotham, O. Rummukainen, J. Herre, and E. A. P. Habets "Evaluation of Binaural Renderers in Virtual Reality Environments: Platform and Examples," presented at the *145th Conventionof the Audio Engineering Society*, (2018 Oct.), paper 424.

[25] ISO, "MPEG-I Immersive Audio Documentation for the Audio Evaluation Platform (Max External Interface)," ISO/IEC JTC 1/SC 29/WG 6 Document N0086 (2021 Jul.).

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

[27] F. Brinkmann, A. Lindau, S. Weinzierl, et al., "The FABIAN Head-Related Transfer Function Data Base," (DepositOnce: Technische Universität Berlin) (2017 Feb.). https://doi.org/10.14279/depositonce-5718.5.

[28] F. Brinkmann, A. Lindau, S. Weinzierl, et al., "A High Resolution and Full-Spherical Head-Related Transfer Function Database for Different Head-Above-Torso Orientations," *J. Audio Eng. Soc.*, vol. 65, no. 10, pp. 841–848 (2017 Oct.). https://doi.org/10.17743/jaes.2017.0033.

[29] ITU-R, "Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems," *Recommendation ITU-R BS.1534-3* (2015 Oct.).

[30] ITU-R, "Method for the Subjective Quality Assessment of Audible Differences of Sound Systems Using Multiple Stimuli Without a Given Reference," *Recommendation ITU-R BS. BS.2132* (2019 Oct.).

[31] T. Robotham, O. Rummukainen, J. Herre, and E. A. P. Habets, "Online vs. Offline Multiple Stimulus Audio Quality Evaluation for Virtual Reality," *145th Convention of the Audio Engineering Society* (2018 Oct.), paper 10131.

[32] A. Silzle, S. George, E. Habets, and T. Bachmann "Investigation on the Quality of 3D Sound Reproduction," *Proceedings of the International Conference on Spatial Audio (ICSA)* (Detmold, Germany) (2011 Jan.).

[33] ITU-R, "General Methods for the Subjective Assessment of Sound Quality," *Recommendation ITU-R BS.1284* (2019 Jan.).

[34] M. Perez-Ortiz and R. K. Mantiuk, "A Practical Guide and Software for Analysing Pairwise Comparison Experiments," *arXiv preprint arXiv:1712.03686* (2017).

[35] ISO, "MPEG-I Immersive Audio Call for Proposals," ISO/IEC JTC 1/SC 29/WG 6 Document N0056 (2021 Apr.).

[36] ISO, "MPEG-I Immersive Audio Test and Evaluation Procedures," ISO/IEC JTC 1/SC 29/WG 6 Document N0056 (2021 Jul.).

[37] ISO, "Report on MPEG-I Immersive Audio Call for Proposals," ISO/IEC JTC 1/SC 29/WG 6 Document N0056 (2022 Jan.).

[38] C. Anemüller, A. Adami, and J. Herre, "Efficient Binaural Rendering of Spatially Extended Sound Sources," *J. Audio Eng. Soc.* vol. 71, no. 5, pp. xx–xx (2023 May).

[39] V. Pulkki "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466 (1997 Jun.).

[40] L. Terentiv, C. Fersch, D. Fischer, and P. Setiawan, "Voxel-Based Occlusion and Diffraction Modelling for the Upcoming ISO/MPEG Standard for VR and AR," in *Proceedings of the AES International Audio for Virtual and Augmented Reality Conference* (2022 Aug.), paper 10.

THE AUTHORS

Jürgen Herre

Sascha Disch

Jürgen Herre 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 Post-Doc term working on MPEG-2 Advanced Audio Coding (AAC). By the end of 1996, he went back to Fraunhofer IIS, working on the development of advanced technology including MPEG-4, MPEG-7, MPEG-D, MPEG-H, and MPEG-I, currently as a Chief Executive Scientist. 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 Audio Coding, and the AES Technical Council and has been an active member of the MPEG audio group for several decades.

•

Sascha Disch received his Dipl.-Ing. degree in electrical engineering from the Technical University Hamburg-Harburg (TUHH) in 1999 and joined the Fraunhofer Institute for Integrated Circuits (IIS) the same year. Ever since, he has been working in research and development of perceptual audio coding and audio processing. From 2007 to 2010, he was a researcher at the Laboratory of Information Technology, Leibniz University Hannover (LUH), receiving his Doctoral Degree (Dr.-Ing.) in 2011. He contributed to the standardization of MPEG Surround, MPEG Unified Speech and Audio Coding (USAC), MPEG-H 3D Audio and the 3GPP Enhanced Voice Services (EVS) codec. Dr. Disch is also part of the development team of the upcoming MPEG-I Immersive Audio standard. His research interests as a Chief Scientist at Fraunhofer IIS and a member of the International Audio Laboratories Erlangen include waveform and parametric audio coding, audio bandwidth extension, and digital audio effects.