Logo

Description automatically generated ISO/IEC JTC 1/SC 29/AG 3 N0090

**ISO/IEC JTC 1/SC 29/AG 3**

**MPEG Liaison and Communications: KR**

**Document type:** General

**Title:** White Paper on MPEG-H 3D Audio

**Status:** Approved

**Date of document:** 2022-10-28

**Source:** ISO/IEC JTC 1/SC 29/AG 3

**Expected action:** None

**Action due date:** None

**No. of pages:** 9 (including cover page)

**Email of Convenor:** kyuheonkim@khu.ac.kr

**Committee URL:** https://isotc.iso.org/livelink/livelink/open/jtc1sc29ag3

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG6, MPEG AUDIO CODING**

**ISO/IEC JTC1/SC29/WG6 N0090**

**October 2022, Mainz, DE/Virtual**

|  |  |
| --- | --- |
| **Title** | **White Paper on MPEG-H 3D Audio** |
| **Source** | **AG 3, MPEG Liaison and Communications** |
| **Status** | **Approved** |
| **Serial Number** | **22135** |

# ABSTRACT

The MPEG-H 3D Audio [1][2] specifies a universal audio coding and rendering environment that is designed to efficiently represent high-quality spatial or immersive audio content for storage and transmission. Since there is no generally accepted “one-size-fits-all” format for immersive 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). MPEG-H 3D Audio supports rendering over a wide range of reproduction conditions (i.e., various loudspeaker setups, headphones, possibly with background noise in the listening environment).

1 COMMERCIAL ADOPTION

The MPEG-H 3D Audio specification was completed in 2017 and a verification test to assess the subjective audio quality provided by the technology was completed the same year [10].

South Korea paved the road to ATSC 3.0, adopting its Ultra High Definition (UHD) TV standard in 2016 and launching 4K ATSC 3.0 broadcasts in May 2017 that now reach more than 70% of the South Korean population (see Fig. 1). MPEG-H 3D Audio is specified in ATSC A/342-3 [3] and is the only audio system selected for ATSC 3.0 UHD services and has been used for terrestrial broadcast since May 2017 [4].

Diagram

Description automatically generated with medium confidence

Fig. 1 ATSC 3.0 deployment in South Korea with MPEG-H 3D Audio [https://www.atsc.org/nextgen-tv/deployments/].

In Brazil, a major technological upgrade of the digital TV infrastructure is underway (see Fig. 2). Part of it is the TV 2.5 project led by the Brazilian Digital Terrestrial TV System (SBTVD) Forum, which was finalized in August 2019 and includes state-of-the art media technologies such as immersive and personalized audio and HDR video. MPEG-H 3D Audio is used in Brazil as an optional audio system over the existing TV 2.5 broadcast system based on ISDB-Tb and in December 2021, Grupo Rede Amazônica became the first broadcaster in Latin America to provide a 24/7 MPEG-H Audio service on one of their regular terrestrial broadcast channels using ISDB-Tb TV 2.5. TV Globo, the largest broadcast group in South America, has enabled MPEG-H 3D Audio a well during major events such as Rio Carnival and Rock in Rio in 2019 and 2022.

The Brazilian Digital Terrestrial Television System Forum (SBTVD) issued, in July of 2020, a Call for Proposals (CfP) seeking input for Brazil's next generation Digital TV system under the name "TV 3.0 Project" [4]. For the audio component, the CfP specified a clear set of requirements and use cases, as well as a very strict feature evaluation phase, conducted by an independent test laboratory, see section 5.5 in [5]. The MPEG-H 3D Audio system [6] has been successfully evaluated by the University of São Paulo researchers and the results have been made publicly available in [7].

A picture containing chart

Description automatically generated

Fig. 2 Terrestrial TV evolution in Brazil [https://forumsbtvd.org.br/tv3\_0/].

Considering the results of the TV 3.0 Project Phase 2 testing and evaluation, as well as market and intellectual property aspects of the candidate technologies, the SBTVD Forum subsequently selected the MPEG-H Audio technology as the mandatory audio component for next generation TV broadcast system in Brazil. The TV 3.0 broadcast is expected to start in 2025 in Brazil.

In 2019 Sony launched an immersive music audio format called 360 Reality Audio [8] based on MPEG-H 3D Audio. The format uses object-based spatial audio that enables the best possible audio experience independent from the loudspeaker configuration at the receiver side. The 360 Reality Audio format is a complete ecosystem from music creation over distribution to playback. Sony is working with major music labels and distribution services to realize the system. Premium music streaming services already available from: Tidal, Deezer, nugs.net and Amazon Music HD [9]. The major benefit of MPEG-H 3D Audio, in comparison to other formats supported in external sound devices, is that MPEG-H enables delivery over HDMI of an immersive bitstream carrying discretely coded channels and objects. Decoding of the complete bitstream allows the soundbar to reproduce the audio content at the highest quality level utilizing the full potential of its built-in high-end virtualization methods for creating an immersive experience.

# 2 TECHNICAL OVERVIEW

The MPEG-H 3D Audio architecture supports three important production paradigms for spatial sound: channels, objects and higher order ambisonics.

## 2.1 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, 22.2). In this way, each channel signal is associated with a specific spatial position. The degree of spatial realism and immersion generally increases with the number of loudspeaker channels. However, one issue for channel-based content is that it expects a particular loudspeaker layout. MPEG-H 3D Audio overcomes this restriction with a conversion process that adapts the content to any loudspeaker setup.

## 2.2 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 to enable moving sound sources, such as a plane fly-over. These objects are then reproduced on the user loudspeaker setup (or headphones) by a rendering algorithm. This enables the user to create an interactive and personalized sound experience by adjusting the object characteristics of the rendered output [11]. For example, users could increase or decrease the level of the announcer’s commentary or actor’s dialogue 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 permits greater spatial resolution when presented on a setup with more loudspeakers.

## 2.3 Higher Order Ambisonics

A third approach to representing spatial audio content is higher order ambisonics [12], which decomposes the three-dimensional 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).

## 2.4 Decoder Architecture

Fig. 3 shows a high-level view of the MPEG-H 3D Audio decoder. The incoming MPEG-H 3D Audio bitstream is decoded by the Core Decoder which reproduces the encoded waveforms that represent either channel signals, object signals or HOA coefficient signals. These waveforms are then processed in dedicated processing chains for each of these signal types. A Format Converter processes the channel signals to convert them to the target rendering loudspeaker layout. 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 together in a Mixer to generate the final loudspeaker feed signals. Optionally, binaural output can be produced. The most important blocks of the system are discussed in more detail in the following sections.

Diagram

Description automatically generated

Fig. 3 Top-level architecture of MPEG-H 3D Audio decoder.

## 2.5 Audio Signal Compression

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) [13][14] system which is a state-of-the-art MPEG codec for compression of mono, stereo and multi-channel audio signals.

The codec bitstream has been designed to enable seamless transitions between encoded streams, for example with different coding bitrates, which is particularly advantageous for adaptive streaming over IP networks. This means that MPEG-H 3D Audio is fully compatible with MPEG Dynamic Adaptive Streaming (DASH) [15],

# 3 RENDERING

## 3.1 Channel rendering: Format Conversion

In order to allow reproduction of encoded channel-based content on any available loudspeaker setup connected to the MPEG-H 3D 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 downmix must be performed in order to enable the best possible listening experience, given the available speaker layout.

## 3.2 Object Rendering

In MPEGH 3D Audio, objects consist of an encoded monophonic waveform and associated metadata describing how to render these objects to a specific spatial location. An associated object gain also can be transmitted and both position and gain can be defined as dynamic trajectories (see MPEG-H 3D Audio metadata overview [16]).

The MPEG-H 3D Audio object renderer is based on the Vector Base Amplitude Panning (VBAP [17]) algorithm and renders the transmitted audio objects to any given output (target) loudspeaker setup. It processes one decoded audio stream per 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. MPEGH 3D Audio provides an automatic triangulation algorithm that supports both standard and 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, for any loudspeaker setup, complete 3D triangle meshes to the VBAP algorithm [2].

In accordance with the VBAP approach, the MPEG-H 3D Audio renderer searches for the loudspeaker triangle that encloses each object and builds an associated vector base. For this loudspeaker triangle, panning gain values are computed that preserve overall signal energy. The computed gains can be linearly interpolated between the panning values computed at sequential time stamps. Finally, the contributions of each object are summed to form the final renderer output signals.

## 3.3 HOA Decoding and Rendering

Higher Order Ambisonics (HOA) describes the audio scene as a three-dimensional acoustic sound field that is represented as a truncated expansion of the wave field into spherical harmonics [12], [18]. 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 wave field that is to be described for transmission or reproduction.

For a complete 3D expansion of order n, (n+1)2 coefficient signals have to be carried and many of those signals exhibit considerable correlations (and thus redundancy) between them. Hence, both 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 tool set 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 bit-stream. Conversely, this process is inverted in the MPEG-H 3D Audio decoder.

# 4 ADDITIONAL TOOLS and FUNCTIONS

## 4.1 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 [19] and allows to control the dynamic range of the playback in consideration of the background noise conditions of the playback environment (e.g., living room, car, airplane, …), effectively providing better subjective user experience in very noisy environments by increasing the level of the most quiet portions of the audio program such that they are more easily heard. Furthermore, a loudness normalization function avoids jumps in loudness when the user switches between different programs.

## 4.2 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.

# 5 PERFORMANCE

As part of the standardization process, MPEG conducts a verification test of all standardized technology. The Verification Test for MPEG-H 3D Audio [10] assessed the performance of a subset of the standardized technology, the Low Complexity Profile. It consisted of four subjective listening tests, that each evaluated audio quality and compression performance at an operating point representing a distinct use case.

The first three tests (Tests 1, 2 and 3) were conducted in high-quality listening rooms that were calibrated to conform to the criteria set forth in BS.1116 [20] and also calibrated to be perceptually similar to each other. The fourth test (Test 4) was conducted in acoustically isolated sound booths. Altogether, the tests results were based on 190 subjects and nearly ten thousand subjective scores.

**Test 1: Ultra HD Broadcast:** This use case assumed an 8k video broadcast and a mix of 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 (Rec. ITU-R BS.1116) [20]. 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.

**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) channel loudspeaker layouts. All audio was coded at three bit rates: 512 kb/s, 384 kb/s 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, the specific bit rates depended on the number of channels in the audio material. Bit rates ranged from 256 kb/s (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.

The test results for tests 2, 3 and 4 are given in Fig. 4, where the vertical axis is MUSHRA [21] 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, lo, indicates the high, medium and low bitrates for audio signals in 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.



Fig. 4 Plot of subjective audio quality of MPEG-H 3D Audio for Test 2, Test 3 and Test 4. Scores of 80 to 100 are “Excellent.” Note that, for greater visibility, only the MUSHRA scale above 75 points is shown here.

# 6 CONCLUSIONS

MPEG-H 3D Audio is an audio coding standard that provides state-of-the-art signal compression with rich metadata. The standard was developed for highly immersive content and supports the coding and transmission of audio as audio channels, audio objects, or Higher Order Ambisonics. It can support up to 64 loudspeaker channels and 128 core codec channels and offers additional tools such as loudness normalization, dynamic range control and can adapt audio content for presentation on any loudspeaker layout or can produce binauralized output for headphone presentation. The rich metadata enables new aspects of user interactivity, such as dialog enhancement. The 3D Audio Low Complexity Profile has excellent performance at bitrates that are compatible with broadcast and streaming delivery service rates.

# REFERENCES

1. ISO/IEC 23008-3:2022, (Information technology — High efficiency coding and media delivery in heter-ogeneous environments— Part 3: 3D Audio.
2. J. Herre, J. Hilpert, A. Kuntz, J. Plogsties: "MPEG-H Audio - The New Standard for Coding of Immersive Spatial Audio", IEEE Journal of Selected Topics in Signal Processing, vol. 9, no. 5, 2015, pp. 770-779.
3. ATSC A/342-3:2022 "ATSC Standard, A/342 Part 3: MPEG-H System,", Advanced Television Systems Committee, Washington, DC, 3 March 2022. https://prdatsc.wpenginepowered.com/wp-content/uploads/2022/04/A342-3-2022-03-MPEG-System.pdf
4. South Korea launches UHD TV with MPEG-H Audio, <https://www.audioblog.iis.fraunhofer.com/south-korea-uhd-tv-mpeg-h>
5. Grupo Rede Amazonica rolls out MPEG-H Audio on their free-to-air Amazon Sat broadcast channel using ISDB-Tb TV 2.5, https://www.iis.fraunhofer.de/en/pr/2021/20211203\_mpeg\_h\_amazon\_sat.html
6. Brazilian Digital Terrestrial Television System Forum (SBTVD) TV 3.0 Project, Available: <https://forumsbtvd.org.br/tv3_0/>
7. Brazilian Digital Terrestrial Television System Forum (SBTVD) TV 3.0 CfP Phase 2 / Testing and Evaluation: TV 3.0 Project, Available: <https://forumsbtvd.org.br/wp-content/uploads/2021/03/SBTVD-TV_3_0-P2_TE_2021-03-15.pdf>
8. Murtaza, et al (2021): MPEG-H Audio System for SBTVD TV 3.0 Call for Proposals. SET INTERNATIONAL JOURNAL OF BROADCAST ENGINEERING. ISSN Print: 2446-9246 ISSN Online: 2446-9432. doi: 10.18580/setijbe.2021.3. Web Link: <http://dx.doi.org/10.18580/setijbe.2021.3>
9. Brazilian Digital Terrestrial Television System Forum (SBTVD) TV 3.0 Testing and Evaluation Report: TV 3.0 Project - Audio Coding, Available: <https://forumsbtvd.org.br/wp-content/uploads/2021/12/SBTVD-TV_3_0-AC-Report.pdf>
10. About 360 Reality Audio, https://electronics.sony.com/360-reality-audio
11. 360 Reality Audio So immersive. So real. <https://music.com/360RA>
12. N16584, “MPEG-H 3D Audio verification test re-port.” Available https://mpeg.chiariglione.org/standards/mpeg-h/3d-audio/mpeg-h-3d-audio-verification-test-report.
13. R. Bleidt, D. Sen, A. Niedermeier, B. Czelhan, S. Füg, S. Disch, J.Herre, J. Hilpert, M. Neuendorf, H. Fuchs, J. Issing, A. Murtaza, A. Kuntz, M, Kratschmer, F. Küch, R. Füg, B. Schubert, S. Dick, G. Fuchs, F. Schuh, E. Burdiel, N. Peters, M. Kim, “Development of the MPEG-H TV Audio System for ATSC 3.0,” IEEE Transactions on Broadcasting, vol. 63, no. 1, pp. 202 - 236, 2017.
14. F. Zotter and M. Frank, Ambisonics, A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement and Virtual Reality. Springer, 2019.
15. ISO/IEC 23003-3:2012 Information technology — MPEG audio technologies — Part 3: Unified speech and audio coding.
16. M. Neuendorf, M. Multrus, N. Rettelbach, G. Fuchs, J. Robilliard, J. Lecomte, S. Wilde, S. Bayer, S. Disch, C. Helmrich, R. Lefebvre, P. Gournay, B. Bessette, J. Lapierre, K. Kjörling, H. Purnhagen, L. Villemoes, W. Oomen, E. Schuijers, K. Kikuiri, T. Chinen, T. Norimatsu, C. Seng, E. Oh; M. Kim, S. Quackenbush, B. Grill, “The ISO/MPEG unified speech and audio coding standard - consistent high quality for all content types and at all bit rates,” J. Aud. Eng. Soc., vol. 61, no. 12, pp. 956–977, Dec. 2013.
17. ISO/IEC 23009-1:2019, Information technology - Dynamic adaptive streaming over HTTP (DASH) - Part 1: Media presentation description and segment formats.
18. S. Füg, A. Hölzer, C. Borß, C. Ertel, M. Kratschmer, J. Plogsties, “Design, Coding and Processing of Metadata for Object-Based Interactive Audio”, 137th AES convention, 2014, Los Angeles, USA.
19. V. Pulkki: Virtual sound source positioning using vector base amplitude panning. Journal of the Audio Engineering Society, Volume 45, Issue6, June 1997; pp. 456-466.
20. B. Rafaely: “Plane-wave decomposition of the sound field on a sphere by spherical convolution,” Journal of the Acoustical Society of America 116, 2149 (2004), DOI:10.1121/1.1792643.
21. ISO/IEC 23003-4:2015 Information technology — MPEG audio technologies — Part 4: Dynamic Range Control.
22. ITU-R Recommendation BS.1116-3 (02/2015), “Methods for the subjective assessment of small impairments in audio systems.”
23. ITU-R Recommendation BS.1534-3 (10/2015), “Method for the subjective assessment of intermediate quality level of coding systems,” also known as “MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA).”