MPEG-H Audio System for SBTVD TV 3.0 Call for Proposals

Adrian Murtaza
Stefan Meltzer
Yannik Grewe
Nicolas Faecks
Mickael Raulet
Lucas Gregory

CITE THIS ARTICLE
Murtaza, Adrian; Meltzer, Stefan; Grewe, Yannik; Faecks, Nicolas; Raulet, Mickael; Gregory, Lucas; 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

SET INTERNATIONAL JOURNAL OF BROADCAST ENGINNERING - SET IJBE V.7, 2021, Article 3, 16p.
 / ISSN Print: 2446-9246 | ISSN Online: 2446-9432

MPEG-H Audio System for SBTVD TV 3.0 Call for Proposals

Adrian Murtaza[(1)], Stefan Meltzer[(1)], Yannik Grewe[(1)], Nicolas Faecks[(1)],
Mickael Raulet[(2)], Lucas Gregory [(2)]
[(1)] *Fraunhofer Institute for Integrated Circuits (IIS),* [(2)]*ATEME*

***Abstract*— Under the name "TV 3.0 Project", the Brazilian Terrestrial Television System Forum (SBTVD) has issued the Call for Proposals (CfP) for a next generation Brazilian digital TV system, in July 2020. The MPEG-H Audio system, based on the open international standard ISO/IEC 23008-3, has been proposed by Fraunhofer IIS, ATEME, the Digital Broadcasting Experts Group (DiBEG) and the Advanced Television Systems Committee (ATSC). This paper provides an overview of the MPEG-H Audio system and the TV 3.0 Project requirements for the audio component. The TV 3.0 Project specifies a detailed test and evaluation procedure for verifying the fulfillment of the requirements. With wide industry support, the MPEG-H Audio system brings immersive sound, advanced interactivity, and accessibility options, as well as advanced features like hybrid delivery, consistent loudness after user interaction, connectivity options for external sound devices and seamless configuration changes. The MPEG-H Audio proponents have submitted a complete production and broadcast real-time chain to the SBTVD Forum which demonstrates the most advanced features.**

***Index Terms* — 3D and Immersive Audio, Accessibility, ATSC 3.0, Audio Coding, Broadcast, Broadband, Emergency warning system, Hybrid, Immersive Sound, MPEG-H Audio, Next Generation Audio, Object-based broadcasting, Personalized Sound, SBTVD TV 3.0, Streaming, Virtual Reality, Augmented Reality**

I. Introduction

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" [1]. Without the constraints of a backward compatibility requirement, the TV 3.0 Project is paving the way for an advanced and modern next-generation television system in Brazil. The SBTVD Forum has established a set of TV 3.0 requirements and use cases, covering six system components (Over-the-air Physical Layer, Transport Layer, Video Coding, Audio Coding, Captions, and Application Coding). The CfP was divided into two phases: Phase 1 required an initial submission from proponents identifying the candidate technology and providing basic information, while during Phase 2, the proponents were expected to submit a full specification of the candidate technology as well as hardware and software solutions for the feature evaluation.

In response to the SBTVD TV 3.0 Call for Proposals, Fraunhofer IIS, ATEME, DiBEG, and ATSC have proposed the MPEG-H Audio system for the audio component [1] and have provided a complete production and broadcast chain to the SBTVD Forum for Phase 2 feature evaluation. The MPEG-H Audio system is fulfilling all TV 3.0 requirements listed in the CfP and provides the most advanced feature set and use cases as detailed in this document.

This paper describes a snapshot of the MPEG-H Audio proposal to SBTVD TV 3.0 Project. It is structured as follows: given that MPEG-H Audio is an open international ISO/IEC standard, the MPEG standardization process and adoption in various worldwide application standards is briefly introduced. Then, existing production workflows using MPEG-H Audio for live broadcast and post-production are outlined. Finally, we describe how the MPEG-H Audio system fulfils the most challenging TV 3.0 requirements and ensure an easy transition from the existing ISDB-Tb broadcast system to the future based TV 3.0 system.

II. MPEG-H Audio System Introduction

MPEG-H Audio is the most advanced Next Generation Audio (NGA) system and based on an open international standard: ISO/IEC 23008-3, MPEG-H 3D Audio [2].

The MPEG-H Audio system provides more realism through sound from above and below as well as around the listener and an unprecedented degree of freedom to consumers for personalizing the audio experience. With its unique interactivity features, MPEG-H Audio offers viewers flexibility to actively engage with the content and adapt it to their own preferences. The easiest way to interact with the content is to select one of several predefined audio presentations, called Presets. Those are complete audio mixes with a descriptive label attached to them, for example "Default TV mix", "Dialog enhanced audio" or "Venue sound".

Furthermore, simple adjustments are possible, such as increasing the dialogue prominence in relation to other audio elements. Interested viewers can dive deeply into advanced scenarios, select certain audio elements of the audio mix and adjust these elements in level and/or position.

A menu displaying all personalization options is available on MPEG-H Audio enabled devices or applications for viewers to personalize their content using, for example, the remote control of the TV or the touch screen on a mobile device. With its innovative system design, the MPEG-H menu will automatically adapt to the content creator's intentions and only display the interactivity options currently

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

available.

MPEG-H Audio opens an entire new level of sound going beyond stereo and surround. With sound coming from above or below, a third dimension is added to the audio experience and lets listeners experience sound in a more realistic and natural way. Depending on the playback system, the soundscape can be extended with sounds from below like footsteps down on the floor completing the immersive experience.

Another unique feature of MPEG-H is its capability to adapt the playback of content to the capabilities of the playback device. With the built-in renderer and advanced dynamic range and loudness management, the content will always be reproduced in the highest quality and with the best user experience achievable on the device in use.

This feature eases the content creation process, as a single MPEG-H stream can deliver the content to all kinds of receiver and playback devices, from headphones to sound bars and discrete loudspeaker systems providing the best quality possible.

A. MPEG Standardization

Finalized in 2015, MPEG-H 3D Audio (ISO/IEC 23008-3 [2]) is the latest audio compression standard developed by Moving Pictures Experts Group (MPEG), following a defined, competitive and collaborative process with the participation of the world's leading experts in the field of audio coding technology. MPEG is a joint working group of the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC).

From the start of the MPEG-H Audio development, the goal was set to deliver the best possible experience with immersive sound as well as enabling advanced accessibility, interaction and personalization features in one solution, taking audio to the next level. The combination of highly efficient encoding technologies, the ability to represent audio content using three different formats (see below), as well as advanced loudness and Dynamic Range Control (DRC) management is the base of the MPEG-H Audio system. The core codec represents the latest evolution of the AAC codec family with the highest coding efficiency and flexibility in computational complexity. The ability to represent audio content in three formats — channel-based, object-based, and scene-based (HOA) – or even as a simultaneous mix of all three, enables maximum flexibility for content producers and allows the usage for all kind of program material and applications including VR/AR.

The use of audio objects introduces the option of interactivity and personalization for the end user into the creative's toolbox and enhances the user experience in an unprecedented way. At the same time, these new elements do not necessarily require large upgrades to the broadcasting infrastructure. The flexibility, the rich set of features, and the coding efficiency are also the reasons why MPEG has selected MPEG-H as the core audio codec for the next generation of audio standards currently under development in the MPEG-I Immersive Audio Call for Proposals [3].

MPEG standards are published by the International Organization for Standardization (ISO) and are publicly available. MPEG is also maintaining the standards and provides updates which bring corrections and clarifications of the standard's document as well as new technologies and adaptations to new market developments. These facts are key for broad adoption of the standards in the market and ensure that multiple implementations provide a competitive environment and bring forward the best products to serve market needs. To support the development of such products, MPEG provides compliance bit streams, and MPEG members help to create test suites for certain applications, e.g., the MPEG-4 AAC/HE-AAC test suites for ISDB-T receivers in Brazil. A detailed description of the newly introduced coding tools in MPEG-H Audio can be found in [6].

B. Profiles and Levels

MPEG standards are defined as a toolbox, including a wide range of tools. Different profiles are defined based on use cases and applications by selecting only the tools fitting best for the targeted application space. This is also the case for the MPEG-H 3D Audio standard and its profiles.

While profiles establish a subset of the tools, levels put additional constraints on the parameters of these tools, allowing a finer adaptation on certain use cases. The combination of profile and level finally determines the necessary processing power and memory requirements for a specific application.

The ISO/IEC 23008-3 3D Audio standard, defines the following audio profiles:

1) The High Profile – includes all tools of the standard and provides a complete set of features. The High Profile is a theoretical profile and a superset of the Low Complexity and Baseline Profiles.

2) The Low Complexity Profile – provides features for broadcasting, VR/AR and streaming applications (ISO/IEC 23008-3, subclause 4.8.2.1[2]).

3) The Baseline Profile – provides features for broadcasting and streaming applications (ISO/IEC 23008-3, subclause 4.8.2.5 [2]). The Baseline Profile is a subset of the Low Complexity Profile.

TABLE I
LEVELS AND THEIR CORRESPONDING RESTRICTION FOR BASELINE AND LOW COMPLEXITY PROFILES

Level	1	2	3	4	5
Max. sampling rate [kHz]	48	48	48	48	96
Max. number of core channels in compressed data stream	10	18	32	56	56
Max. number of decoder-processed core channels	5	9	16(*)	28	28
Max. number of channels in referenceLayout	5	9	16(*)	24	24

(*) The Baseline profile supports in Level 3 up to 24 objects if the additional complexity restrictions given in ISO/IEC 23008-3, subclause 4.8.2.5.2 [2]) are applied on the encoding process.

Table 1 provides an overview of the levels and their corresponding characteristics for Low Complexity and Baseline Profiles. A complete description can be found in ISO/IEC 23008-3, subclause 4.8.2 [2].

1) Low Complexity Profile

The Low Complexity Profile is a superset of the Baseline

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

Profile. It includes two additional coding tools for Higher-Order Ambisonics (HOA) and Linear Prediction Domain (LPD), which is irrelevant for the majority of Next Generation Audio broadcast and streaming applications.

The HOA path of the Low Complexity Profile uses tools for decoding and rendering the HOA signals, in addition to the channel and object signals. Implementation of these tools can be quite complex and can lead to doubling the implementation effort in comparison to the Baseline Profile. Similarly, the testing effort also increases as higher number of conformance tests streams and test cases need to be verified for compliance. The increased implementation and testing effort for the Low Complexity Profile is justified for applications that benefit from having the HOA format available, such as VR/AR use cases.

2) Baseline Profile

The Baseline Profile is tailored especially for today's broadcast and streaming use cases. Without the support for HOA and the Linear Prediction Domain (LPD) tools, it offers significantly reduced implementation and testing effort without limiting the capabilities for all significant broadcast and streaming applications. This makes the Baseline Profile the natural choice for Consumer Electronics (CE) manufacturers.

As the Low Complexity and Baseline Profile share the majority of the tools, the MPEG-H 3D Audio standard specifies a compatibility signaling mechanism (see ISO/IEC 23008-3, subclause 4.8.2.7 and Annex P [2]), which ensures that Baseline Profile bit streams can also be decoded by Low Complexity Profile decoders and vice versa. The signaling maximizes the amount of content, which could be handled by a decoder to the benefit of the users.

The MPEG-H 3D Audio Baseline Profile fulfils all mandatory requirements for SBTVD TV 3.0 and is also proposed for next-generation DTTB in Japan.

C. Subjective Quality Evaluation

The performance of the MPEG-H Audio system was carefully evaluated by MPEG and documented in two MPEG Verification Test Reports [4][5].

With more than 1 million subjective ratings, from 341 expert listeners at nine prestigious independent test labs around the world (Fraunhofer IIS, Sony, NHK, Gaudio, Nokia, Orange, Qualcomm, Dolby and ETRI), the MPEG-H 3D Audio standard is the most thoroughly tested NGA codec.

The "MPEG-H 3D Audio Baseline Profile Verification Test Report" [4] includes five listening tests assessing the performance of the Baseline Profile. The tests cover a wide range of bit rates as well as an exhaustive range of use cases (i.e., from 22.2 down to 2.0 channel presentations).

The statistical analysis of the test data resulted in the following conclusions:

1) Test 1: Ultra-HD Broadcast

The "Ultra-HD Broadcast" use case was evaluated using highly immersive audio material coded at 768 kb/s and was presented on 22.2 or 7.1+4H channel loudspeaker layouts. The results showed that the Baseline Profile easily achieves "ITU-R High-Quality Emission" quality at the tested bit rate for broadcast applications.

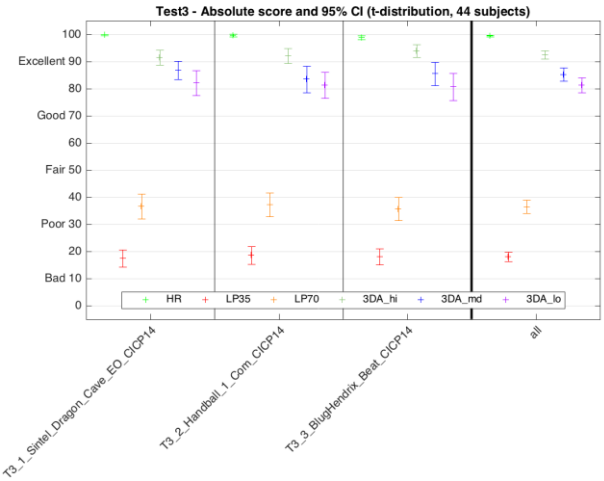

Fig. 1. Test 3 performance for 5.1+2H layout (CICP 14) immersive content at three different bit rate levels: low ("3DA_lo" - 144 kbps), mid ("3DA_md" - 192 kbps) and high ("3DA_hi" - 256 kbps). [4]

2) Test 2: HD Broadcast

The "HD Broadcast" or "A/V Streaming" use case was evaluated using immersive audio material coded at three different bit rates: 512 kb/s, 384 kb/s and 256 kb/s and presented on 7.1+4H or 5.1+2H channel loudspeaker layouts. The test showed that for all bit rates, Baseline Profile achieved a quality in the range of "Excellent" on the MUSHRA subjective quality scale.

3) Test 3: High Efficiency Broadcast

The performance for the "High Efficiency Broadcast" use case was evaluated using audio material coded at three different bit rates, with specific bit rates depending on the number of channels in the material. Bit rates ranged from 256 kb/s (5.1+2H) to 48 kb/s (stereo). The test showed that for all bit rates, the Baseline Profile achieved a quality in the range of "Excellent" on the MUSHRA subjective quality scale.

4) Test 4: Mobile

The performance for the "Mobile" use case was evaluated using immersive audio material coded at 384 kb/s and presented via headphones. The test showed that at 384 kb/s, Baseline Profile with binauralization achieved a quality in the range of "Excellent" on the MUSHRA subjective quality scale.

5) Test 5: High Quality Immersive Music Delivery

The "High Quality Immersive Music Delivery" use case requires delivery of object based immersive music to the receiver with up to 24 objects at high per object bit rates. This test used 11.1 (as 7.1+4H) as a presentation format, with material coded at a rate of 1536 kb/s. The test showed that at the bit rate of 1536 kb/s, Baseline Profile easily achieves "ITU-R High-Quality Emission" quality for high quality music delivery applications.

Fig. 1 shows the performance for 5.1+2H layout (CICP 14) immersive content at three different bit rate levels: low ("3DA_lo" - 144 kbps), mid ("3DA_md" - 192 kbps) and high ("3DA_hi" - 256 kbps).

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

The performance of the Low Complexity Profile of MPEG-H 3D Audio was assessed in the "MPEG-H 3D Audio Verification Test Report" [5]. Since Low Complexity Profile is a superset of the Baseline Profile, the BL verification test results [4] for channel- and object-based described above apply also to Low Complexity Profile. Additionally, the report [5] includes scores for HOA content in Tests 1 – 4.

The efficiency of the MPEG-H Audio codec allows carrying better quality and/or more channels with the same bit budget as currently used to carry only 5.1 channels. Thus, with the current commonly used broadcast audio data rate of 192 kbps, 5.1 surround channels with four additional height channels (i.e., 5.1+4H) can be delivered and that with improved subjective quality.

D. International Standards and Adoption

MPEG-H Audio has been widely adopted and included in various regional and international standards worldwide. Currently, MPEG-H Audio is the only Next Generation Audio system standardized for broadcast, streaming, hybrid and VR/AR/360-degree video streaming applications within 3GPP. The most notable application standards specifying MPEG-H Audio are:

- **ATSC:** The Advanced Television Systems Committee has successfully included MPEG-H Audio in its ATSC 3.0 suite of standards as ATSC Standard A/342-3 [7]. The corresponding transport layer signaling is specified in ATSC A/331 [8].
- **TTA (South Korea)**: The Telecommunications Technology Association (TTA) has selected MPEG-H Audio as the sole audio system for ATSC 3.0 in South Korea, as specified in TTAK-KO-07.0127 [9].
- **SBTVD (Brazil)**: The SBTVD Forum has adopted MPEG-H Audio [10] as part of the ABNT specification for TV 2.5 in Brazil, ABNT NBR 15602-2 [11]. The additional signaling for transport layer is specified in ABNT NBR 15603 [12] and the MPEG-H Audio receiver specification is provided in ABNT NBR 15604 [13].
- **3GPP**: 3GPP has selected MPEG-H Audio as the only audio format for 360° video streaming services over 5G within Release-15 of the specifications, TS 26.118 3GPP Virtual reality profiles for streaming applications [14].
- **DVB**: DVB has also selected and included MPEG-H Audio in the specification ETSI TS 101 154 v2.3.1 [15] defining the usage of audio and video codecs for DVB systems. The proper signaling for MPEG-2 TS DVB systems was specified in ETSI EN 300 468 [16].
- **ITU**: International Telecommunications Union (ITU) issued the recommendation ITU-R BS.1196-7 (01/2019) for Audio coding for digital broadcasting [17].

Additionally, all major OTT Specifications have adopted MPEG-H Audio, including MPEG CMAF, CTA WAVE, HbbTV, DASH-IF or DVB DASH.

The MPEG-H Audio system was the first Next Generation Audio codec worldwide to go on air 24/7 as South Korea launched its 4K UHD TV services using the ATSC 3.0 standard on May 31, 2017 [18]. The Korean standard [9] for this service mandates the MPEG-H Audio system as the only audio codec for delivery of immersive and personalized sound in South Korea. The Korean government timeline requires the percentage of native UHD content to increase in steps from 5% 2017 to 50% in 2025 and finally to 100% in 2027. In 2027, the HD service based on ATSC 1.0 and currently operated in a simulcast, will be completely switched off.

Fig. 2. UHD TV receiving MPEG-H Audio signal over ATSC 3.0 RF input and displaying native user interface for user interactivity

Professional equipment from encoders, metadata authoring and monitoring units, as well as test receivers are available. On the consumer side, TV sets are available on the market supporting the full feature set of the MPEG-H Audio system (see Fig. 2). Additionally, MPEG-H Audio is available for several years in immersive soundbars and AVRs. Thus, the complete end-to-end chain is available allowing broadcasters to make full use of the advanced features. Lessons learned during live broadcasts of major events using MPEG-H Audio are detailed in [19].

In Brazil, MPEG-H Audio was adopted as part of the TV 2.5 Project to enhance the audio experience over ISDB-Tb with immersive and personalized sound and is fully specified in the ABNT standards [10][11][12][13]. Fraunhofer IIS and its technology partners ATEME, Telos Alliance, SSL, EiTV and Sennheiser have showcased for the first time at the 2019 SET Expo trade show, a live ISDB-Tb broadcast local transmission using MPEG-H Audio in Brazil according to the TV 2.5 suite of standards.

Moreover, during one of the world's biggest music festivals, Rock in Rio [20], Rede Globo has successfully used MPEG-H Audio for a live terrestrial broadcast over the ISDB-Tb system. The musical performances on the two main stages, "Mundo" and "Sunset," were delivered over the air with MPEG-H immersive and personalized sound in the Rio de Janeiro area. The same MPEG-H Audio production was simultaneously delivered over multiple distribution platforms. Besides the ISDB-Tb terrestrial broadcast, an HLS streaming service using MPEG-H Audio was offered and with the support of Rohde & Schwarz, MPEG-H Audio was embedded in an experimental broadcasting UHF channel for the first 5G broadcast transmission field test in Brazil. ATEME's TITAN Live encoder created all three services in parallel from the same input.

Following the successful educational program of Fraunhofer IIS in South Korea and China, the first MPEG-H training center in Sao Paulo opened in February 2021 [21]. Fraunhofer has teamed up with Cinecolor Brazil to offer

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

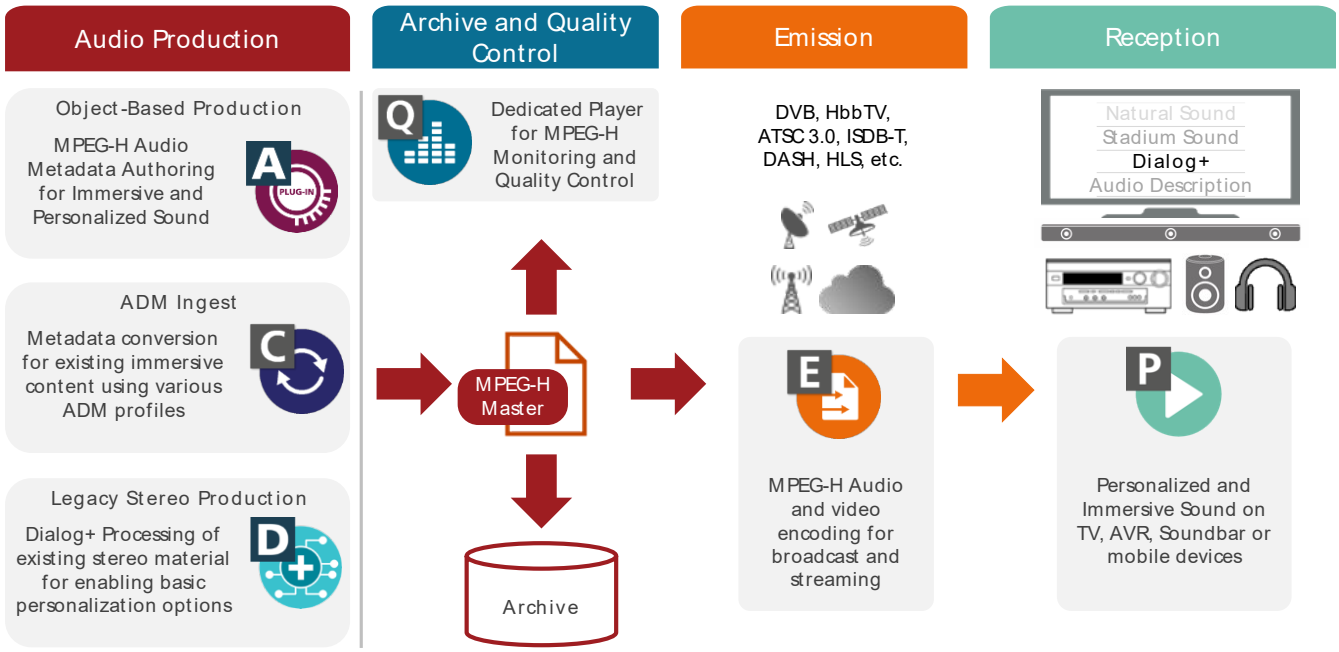

Fig. 3. MPEG-H Master usage in production workflows (simplified)

broadcasters access to the most advanced immersive audio technology.

III. Production Workflows

The production and transmission of MPEG-H Audio introduces new concepts compared to legacy production. The MPEG-H Audio system has been designed specially to explore this new creative options. Besides immersive 3D Audio, content creators can prepare mixes (including the default or main mix of the program) using authoring tools that specify an ensemble of gain and position settings for objects to create preset mix selections that can be presented on a simple menu to the user. All interactivity features offered to the users are strictly defined by the broadcaster during metadata creation. This process of generating metadata is called "authoring" and is the most important difference in production of MPEG-H Audio content compared to a legacy production.

Additionally, a different handling of audio elements is required when using immersive channel-beds and audio objects in production. A 3D-Audio bus structure needs to be available in the production tools, such as Digital Audio Workstation (DAW), broadcast mixing desk and monitoring paths. The additional audio objects must be kept separated from the other components such as *Music & Effects (M&E)* mixes – also called the channel bed – up to the authoring stage, where metadata is generated, including:

- Position information about object reproduction position in 3D space,
- Interactivity limitations for audio objects and presets,
- Loudness information about each component and preset,
- Text labels for presets and audio objects (also in multiple languages),
- Reference and target loudspeaker layouts and
- Many more.

In the following, an introduction to MPEG-H Audio production formats and workflows for live- and post-productions is provided.

A. 3D Audio Studio Recommendations

Jointly with leading industry experts in live production for sports and other major events, Fraunhofer IIS has published 3D Audio Studio Recommendations [22] specifying the main structural requirements and technical specifications for a 3D-Audio production environment. It details best practice for mixing and reproduction in a flexible manner for loudspeaker reproduction systems, including the most common setups such as 5.1+4H and 7.1+4H. The Studio Recommendations support the sound producers with additional guidance for room geometry and room acoustics, loudspeaker positioning and electroacoustic performance, 3D-Audio monitoring and mixing capabilities and provide recommendations for related literature.

B. MPEG-H Master

In the MPEG-H Audio System, all metadata is tightly coupled to the audio essence, ensuring the integrity of the transmitted or stored audio scene. This is achieved by using the "MPEG-H Master" which is a bundle of metadata and the audio content. The MPEG-H Master can be exported as either Broadcast Wave Format File carrying Audio Definition Model (ADM) metadata or MPEG-H Production Format (MPF) file including the metadata as MPEG-H Control Track (see below). Fig. 3 provides a high-level overview of production workflows using the MPEG-H Master format for new object-based productions, ingest of existing ADM metadata and legacy stereo production that can be enhanced with Dialog+ capabilities.

C. MPEG-H Production Format (MPF)

An MPEG-H Production Format (MPF) file is a multi-channel wave file which contains all the audio and metadata of the MPEG-H Audio scene. The metadata is modulated into a regular audio channel called "Control Track" (CT). This is a "time-code like" signal and has been introduced for efficient and robust usage of MPEG-H Audio in SDI-based workflows.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

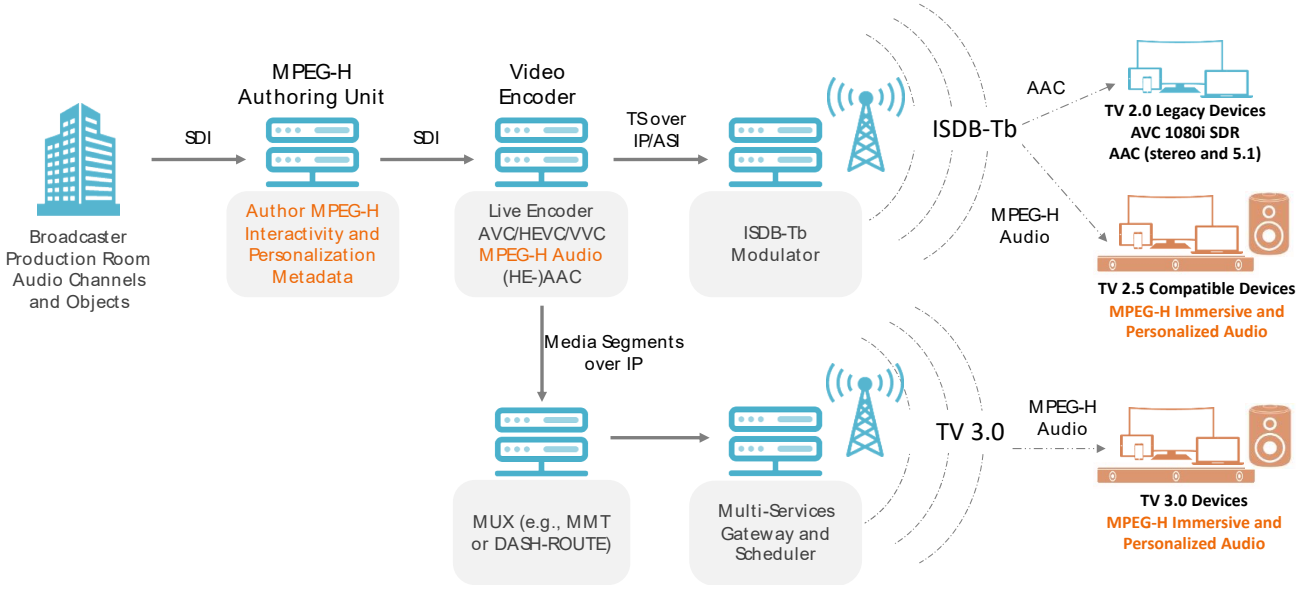

Fig. 4. MPEG-H Live Broadcast workflow (simplified)

The metadata for the audio signal is collected into packets synchronized with the video signal and is modulated with analog channel modem techniques into the Control Track, fitting in the audio channel bandwidth. This signal is unaffected by typical filtering, resampling, or scaling operations in the audio sections of broadcast equipment and ensures the synchronization of metadata with the corresponding audio and video signal. As the metadata contained in the Control Track is aligned to the audio and video data, any configuration change in live or post-production can be applied at every video frame boundary. Typically, the CT is carried on channel 16 within an SDI framework for live broadcast applications or on channel 16 of a multichannel wave-file.

The MPEG-H Control Track does not force audio equipment to be put into data mode or non-audio mode to pass through.

D. Audio Definition Model (ADM)

The Audio Definition Model (ADM) according to ITU-R BS.2076 [23] defines an open metadata format for production, exchange and archiving of Next Generation Audio (NGA) content in file-based workflows. Its comprehensive metadata syntax allows describing many types of audio content including channel-, object-, and scene-based representations for immersive and interactive audio experiences. A serial representation of the Audio Definition Model (S-ADM) specified in ITU-R BS.2125 [24] defines a segmentation of the original ADM for use in linear workflows such as real-time production for broadcasting and streaming applications.

It is acknowledged by ADM experts that application-specific ADM profiles are needed to achieve interoperability in ADM-based content ecosystems. Those ADM profiles incorporate the specific requirements for production, distribution and emission. To ensure interoperability with existing NGA workflows, applications adopting the ADM format should be able to convert native metadata formats to ADM metadata and vice versa such that artistic intent is preserved in a transparent way.

The MPEG-H ADM Profile [27] defines constraints on ITU-R BS.2076 [23] and ITU-R BS.2125 [24]. Those enable interoperability with established NGA content production and distribution systems for MPEG-H 3D Audio as defined in ISO/IEC 23008-3 [2].

E. Post-Production workflows

Typically, DAWs are used for audio post-production. Most common DAWs support the integration of AAX, VST or AU based plugins, which extend the capabilities of the host. Plugins such as the MPEG-H Authoring Plugin – part of the freely available MPEG-H Authoring Suite [25] – or the Spatial Audio Designer (SAD) [26] by New Audio Technology can be integrated seamlessly into the existing post-production workflows for MPEG-H audio content creation.

Different audio tracks from the DAW need to be arranged for the channel-based bed using a 3D panner. Furthermore, selected tracks are configured as audio objects. Compared to real time panning, post-produced content allows for more advanced object movements because automation can be edited and retrieved. The number of objects that move at the same time can be higher for the same reasons. Another benefit of using MPEG-H plugins is that they can overcome the DAWs bus width limitation, e.g., enabling a high number of objects to loudspeaker layouts not natively supported in the DAW, such as 5.1+2H, 5.1+4H or 7.1+4H.

As a next step during post-production, the MPEG-H metadata needs to be created. To monitor the different presets or switch groups that have been defined, a full MPEG-H renderer is included in the authoring tools. When the mix is finished and all metadata entries are authored, the session can be exported to an MPEG-H Master. During this stage, the loudness of all components and presets is measured and embedded into the corresponding metadata fields.

For the case that pre-existing content, e.g., a stereo or 5.1 audio mix, should be prepared for MPEG-H broadcast, the mix does not need to be touched. Only metadata need to be generated. For this, stand-alone tools such as the MPEG-H Authoring Tool (MHAT) – part of the MPEG-H Authoring Suite – are available. It is also possible to monitor the created scene and presets and export the final MPEG-H Master.

To control created MPEG-H Masters with or without an accompanying picture, the MPEG-H Production Format Player – part of the MPEG-H Authoring Suite – can be used

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

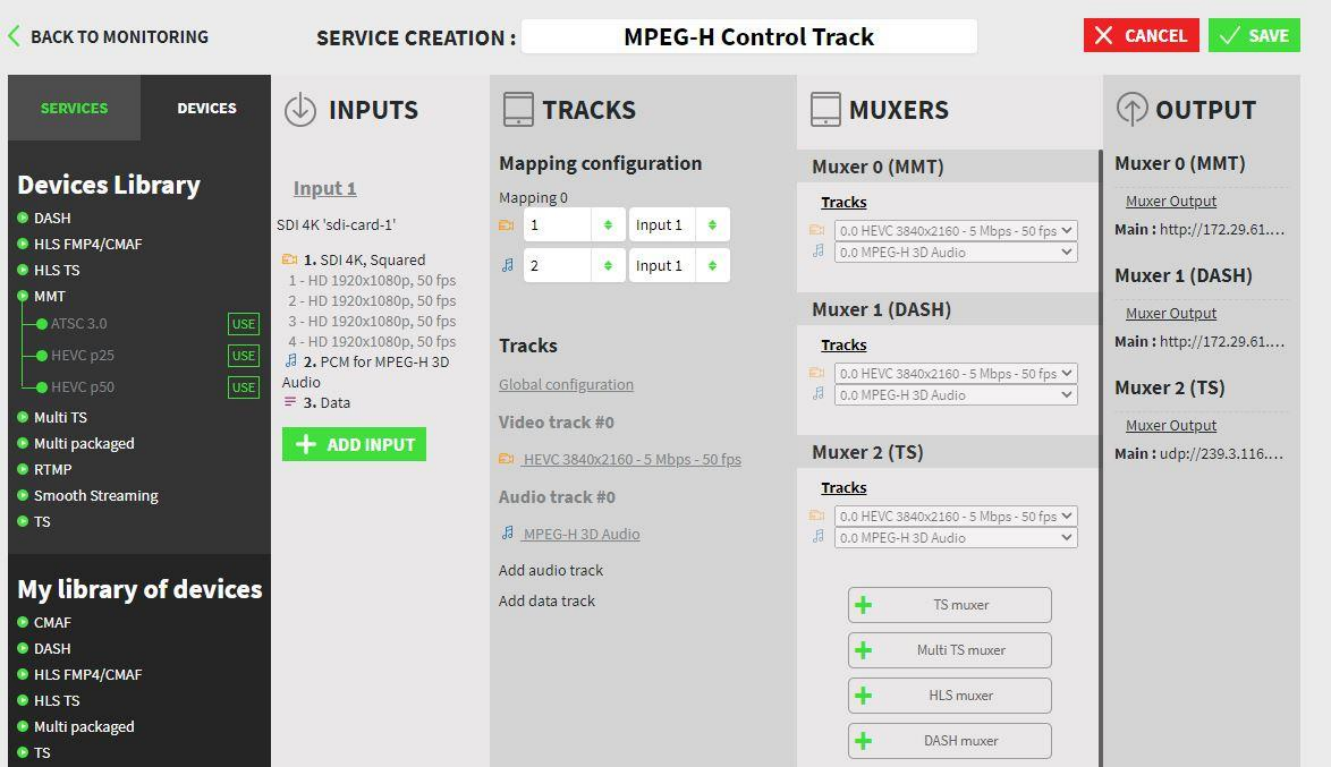

Fig. 5. ATEME Titan Encoder view, MPEG-H Audio single encoding for multiple outputs: MMT/DASH/TS

to ensure that the quality of the mix and the authoring is matching the expectations.

F. Live-Production workflows

The MPEG-H Audio system is designed to work with today's streaming and broadcast equipment using SDI-based workflows as well as with future IP-based infrastructure. In real-time scenarios, the authoring of MPEG-H Audio scenes and the metadata export is handled by a device class called "Authoring and Monitoring Unit" (AMAU).

AMAU systems are easy to integrate into the existing SDI, MADI or future IP based signal flows which are used for broadcast. Fig. 4 shows how AMAU units can be used in live broadcast infrastructure for existing ISDB-Tb TV 2.5 system in Brazil, where legacy receiving devices will decode the AAC signal, while newer receivers are able to provide an immersive and personalized MPEG-H experience. Additionally, the same AMAU can be used for enabling the future TV 3.0 broadcast in Brazil. With a single production and integrated audio and video broadcast encoders, MPEG-H can ensure a smooth transition from TV 2.5 to TV 3.0 with minimum investments for broadcasters.

Audio signals from the mixing console are fed into the AMAU using I/O converters. Within the AMAU, metadata creation and rendering takes place. The AMAU can be controlled using a web interface or hardware controller; typically, the output of the AMAU is an SDI signal including 15 channels of audio and the Control Track on channel 16. In a IP based production facility, the audio and metadata would be transmitted in a container over IP. A video signal can be passed through for visual reference and for synchronization purpose. Finally, the audio and video signals are provided to the emission encoder.

Currently, there are two AMAUs available and used for MPEG-H Audio production in live broadcast: the Multichannel Monitoring and Authoring system from Jünger Audio (MMA) and the Authoring and Monitoring System from Linear Acoustic (AMS). The use of AMAU systems allows productions to make use of all MPEG-H features without changing the entire workflow. Most of today's existing broadcast equipment can still be used.

Due to the advanced capabilities of the MPEG-H system, the monitoring stage during a production is important. Many different speaker layouts from stereo to 7.1+4H can be connected for 3D Audio playback and used for monitoring in an AMAU. Additionally, all interactivity options and the audio quality can be monitored during production using an emulation of end-user receivers with different reproduction configurations.

AMAU systems measure the loudness and true peak values of all channels, objects, output busses and formats, as well as every created Preset in real-time. With the resulting data, correction values are added to the metadata stream compliant with the applicable loudness regulation. The measurement of all generated DRC profiles and real-time loudness correction are also included. Additionally, AMAU production tools support the user with visualizations of all crucial measurement values.

As explained above, AMAU systems interface perfectly with currently deployed broadcast equipment, such as ATEME's Titan Live Encoder (see Fig. 5). This emission encoder includes Fraunhofer's MPEG-H 3D Audio encoder library, which can be either configured by the operator, using presets and encoding profiles up to 7.1+4H, or configured using a Control Track produced upstream by an AMAU and feed to the encoder through the SDI input. In the second scenario, a fallback configuration is also given to the MPEG-H Audio encoder and applied if no consistent Control Track is found in the input stream.

In contrast with the usual operation of Titan Live, or any other emission encoder, the use of the Control Track also permits to address advanced scenarios, such as dynamic

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

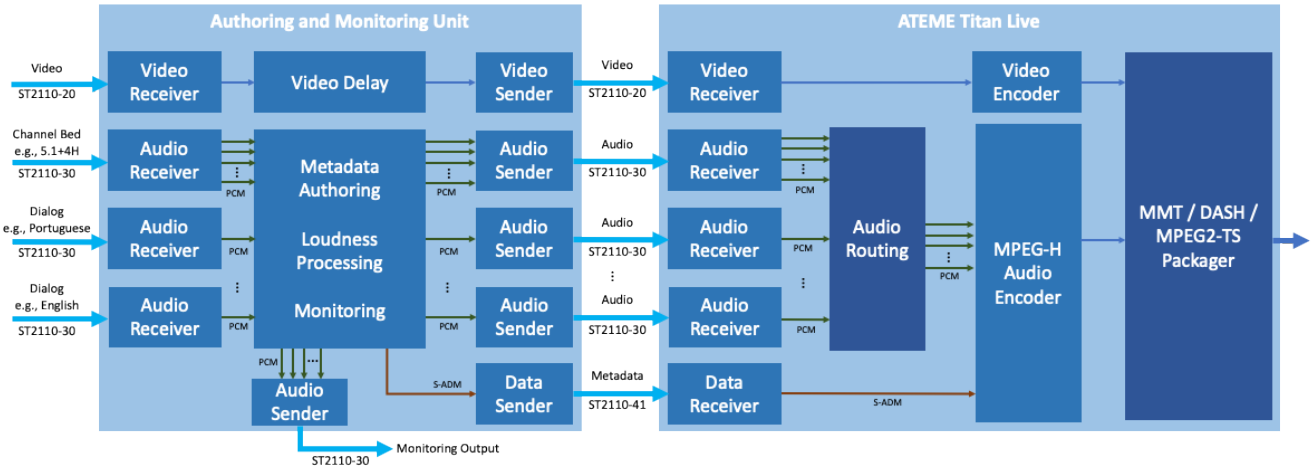

Fig. 6. MPEG-H Audio production workflow based on SMPTE ST 2110 (simplified)

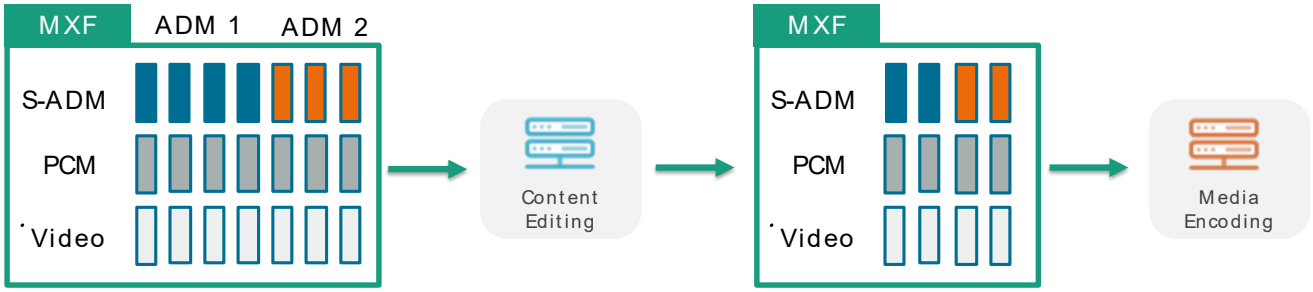

Fig. 7. Example MXF based workflow using ADM metadata

configuration changes for the audio encoder. For instance, the encoder can use a 5.1+4H channel-bed configuration while encoding a live sport event or a movie, and switch to a stereo layout during the commercial breaks. In this case, the configuration changes are triggered by the AMAU, and the MPEG-H Audio encoder seamlessly changes its configuration when a new configuration model is detected in the input Control Track.

Titan Live's muxer and packager unit adapts to these configuration changes by dynamically packaging the encoded audio stream. For DASH and MMT cases, fragmented MP4 (fMP4) segments have variable duration in order to ensure MPEG-H Audio Random Access Points (RAPs) keep aligned with the segments, and that a configuration change does not occur in the middle of a fMP4 segment. High-level signalization is also dynamically updated at configuration changes to reflect the encoded elementary stream's content. This is important for MPEG-H Audio since MPEG2-TS descriptors and fMP4 *SampleEntry* boxes contain part of the information needed to decode and present the content to the end-user. Such information may be used at receiver tune-in or startup and is essential that bitstream data is aligned to the information signaled on the transport layer for a high quality of experience on consumer side.

G. IP-based Production workflows

In legacy SDI and MADI based workflows, PCM audio essence is transmitted via audio channels or embedded audio channels of the SDI video signal and audio metadata is transmitted by means of the MPEG-H Control Track. In IP based workflows according to SMPTE ST 2110, media essences such as video, audio and metadata are transmitted over separate RTP connections and PTP (IEEE 1588, SMPTE ST 2059) is used to synchronize the different essence streams. The transmission of the PCM audio essence (SMPTE ST 2110-30) is already established whereas the standardization for the transport of metadata including serialized ADM metadata (SMPTE ST 2110-41) is still ongoing at the time of writing.

The serial representation of the Audio Definition Model (S-ADM) according to ITU-R BS.2125 defines a segmentation of the original ADM for use in linear workflows such as live production for broadcasting and streaming applications. Like the MPEG-H Control Track, one S-ADM frame contains a set of metadata describing at least the audio frame over the time period associated with that frame. S-ADM has the same structure, attributes and elements as those of ADM, as well as additional attributes to specify the frame format. The S-ADM frames are non-overlapping and contiguous with a specified duration and start time.

Fraunhofer is actively participating in the standardization of live production workflows based on S-ADM and SMPTE ST 2110 in following international standardization organizations: ITU-R, EBU, AES and SMPTE. As soon as the standardization is complete, Fraunhofer and its technology partners will provide and deploy complete solutions for authoring and monitoring of MPEG-H Audio for workflows based on S-ADM and SMPTE ST 2110 in IP-based production environments. For illustration purposes, a simplified ST 2110 Audio and Metadata over IP workflow for MPEG-H Audio is depicted in Fig. 6.

For storage and playout of S-ADM based content, Fraunhofer IIS is actively participating in the standardization of the transport of PCM audio essence and S-ADM metadata inside the Material Exchange Format (MXF, SMPTE ST 377-1), which is underway at the time of writing. The MXF Format is optimized for content interchange or archiving by creators and/or distributors and provides a complete

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

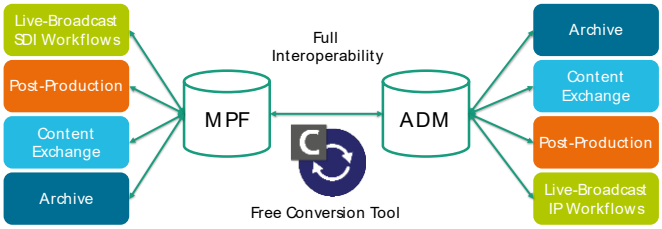

Fig. 8. Full Interoperability between MPEG-H Masters and ADM-based workflows

framework for the transport of NGA. An example application of MXF in practice is shown in Fig. 7.

H. Interoperability

The MPEG-H Conversion Tool [25] provides a lossless conversion between MPF and ADM based production formats. As shown in Fig. 8, the MPF format is a robust and reliable solution for existing SDI-based live workflows and it may be used for all other purposes such as post-production, content exchange and archiving. On the other hand, the ADM format is ideal for content storage, content exchange using MXF based workflows, and with its serialized option for IP-based workflows according to SMPTE ST 2110 suite of standards.

In the future, Fraunhofer IIS and its technology partners will provide tools for the lossless conversion between SDI and ST 2110 interfaces including MPEG-H Control Track and/or S-ADM metadata.

IV. SBTVD TV 3.0 Requirements

The SBTVD TV 3.0 Call for Proposals [1] specifies detailed requirements for each component or sub-component as well as detailed test procedures for verification and validation of the features for each candidate technology. The TV 3.0 audio coding requires – amongst others – support for:

- immersive and interactive audio,
- state-of-the-art coding efficiency,
- live (real-time) encoding with minimum end-to-end latency,
- audio description delivery in the same stream as the main audio,
- emergency warning information audio description,
- seamless and frame-accurate configuration changes or ad-insertion at any time instance.

Additionally, the audio coding shall enable a single delivery format for multiple audio playback configurations, consistent loudness across programs and inside the same program, frame-accurate audio/video synchronization, new immersive audio services, such as VR / AR / XR / 3DoF / 6DoF and extensibility. In the following sub-sections, few of these features are explained in context of the MPEG-H Audio system.

A. Immersive Sound

In MPEG-H, immersive sound can be carried in three primary ways: traditional channel-based sound where each transmission channel is associated with a loudspeaker position; sound carried through audio objects, which are positioned in three dimensions independently of loudspeaker positions; and scene-based (or Ambisonics), where a sound scene is represented by a set of coefficient signals that are the linear weights of spatial orthogonal spherical harmonics base functions.

As the only audio format natively supporting HOA, MPEG-H Audio can provide immersive sound using any combination of these three well-established audio formats.

B. Interactivity and Personalization

The MPEG-H Audio metadata, defined during authoring, carries all the information needed to allow viewers to change the properties of audio objects by attenuating or increasing their level, disabling them, or changing their position in the three-dimensional space. Additionally, the MPEG-H Audio metadata structures empower broadcasters to enable or disable interactivity options and to strictly set the limits to which extent a user can interact with the content.

The simplest use case is probably the most desired and powerful one, for which the dialog or commentary for a program is sent as an object. This allows the viewer to adjust the relative volume or "prominence" of the dialog relative to the rest of the audio elements in the program. In general, broadcasters attempt to mix the sound as a good compromise between dialog, natural sound, music, and sound effects. Viewer's preferences may vary, particularly as the (immersive) sound mix becomes more complex, such as in sports events or action dramas. Additionally, the reproduction setups at home are ranging from tiny building speaker to elaborated AVR controlled multi-speaker setups resulting in a huge influence on the user's experience. This simple case can be extended in offering two or more dialog objects for different languages or commentary oriented to each of the teams in a sporting event.

Moreover, the MPEG-H Audio metadata enables broadcasters to provide several versions of the content, as so-called "presets/preselections", which describe how all channels and objects signals are mixed together and presented to the viewer. Choosing between different presets is the simplest way to interact with the content. Advanced interactivity settings can be offered to more experienced users for manipulating objects individually.

C. Metadata and Broadcaster Control

The MPEG-H Audio system standardizes a rich metadata set to define an audio scene, the "Metadata Audio Elements" (MAE) as specified in ISO/IEC 23008-3, Clause 15 [2]. Each audio track with accompanying metadata is called an "audio element". This enables the most advanced and flexible end-user interactivity and personalization experience, while still offering full control over these features to the broadcasters. This is achieved with standardized metadata for controlling the personalization options such as setting the limits in which the user can interact with the content. The set of MAE metadata consists of:

1) Descriptive metadata: Information about the existence of objects inside the bit stream and high-level properties of audio elements, e.g., textual descriptions by labels, content kind and content language.

2) Control metadata: Information of how interaction is possible or enabled by the content creator.

3) Playback-related metadata: Information about special playback options.

4) Structural metadata: Grouping and combination of elements.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

Audio objects are associated with metadata that contains all information necessary for personalization, interactive reproduction, and rendering in flexible reproduction layouts. The metadata (MAE) is structured in several hierarchy levels. The top-level element of MAE is the "AudioSceneInfo". Sub-structures of the Audio Scene Info contain: "Groups", "Switch Groups" and "Presets".

1) Groups of Elements

The concept of an element group enables arranging related element signals that are to be treated together as a unit, e.g., for interactivity in common or for simultaneous rendering. A use case for groups of elements is the definition of channel-based recordings as audio elements (e.g., a stereo recording in which the two signals should only be manipulated as a pair). Grouping of elements allows for signaling of stems and sub-mixes by collecting the included element signals into groups that then can be treated as a single component.

2) Switch Groups of Elements

The concept of a switch group describes a grouping of components that are mutually exclusive with one another. It can be used to ensure that exactly one of the switch group members is enabled at a time. This allows for switching between, e.g., different language tracks or different commentators, when it is not desired to simultaneously enable multiple language tracks.

3) Presets

Presets can be used to offer combinations of groups and objects for more convenient user selection. Properties of the groups, like default gain or position can be set differently for each preset. It is not necessary to include all groups and objects in a preset.

4) Personalization and Interactive Control

Using the information in the MAE, the MPEG-H Audio system offers listeners the ability to interactively control and adjust various elements of an audio scene within limits set by broadcasters (e.g., to adjust the relative level of Dialog only in a range specified within the AudioSceneInfo structure).

The metadata allows for the definition of different categories of user interactivity as listed below:

- ***On-Off Interactivity***: The content of the referred group is either played back or discarded.
- ***Gain Interactivity****:* The overall loudness of the current audio scene will be preserved but the prominence of the referred signal will be increased or decreased.
- ***Positional Interactivity***: The position of a group of objects can interactively be changed. The ranges for azimuth and elevation offset, as well as a distance change factor can be restricted by metadata.

In order to reflect the content creator's intention to what extent their artistic intent may be modified, the interactivity definitions include minimum and maximum ranges for each parameter (e.g., the position can only be changed in a range between an offset of -30° and 30° azimuth).

D. Advanced Accessibility Options

Using object-based audio, MPEG-H Audio offers advanced and improved accessibility services (i.e., Dialog Enhancement and Audio Description) allowing hearing and visual impaired audience to experience at a new quality level.

1) Dialog Enhancement

MPEG-H Audio includes Dialog Enhancement (DE) for automatic device selection (prioritization) as well as for user manipulation. As an additional feature, MPEG-H Audio supports the personalization of Dialog Enhancement through a user interface offering the direct adjustment of the enhancement level, inside the range defined by the broadcaster. This range (e.g., minimum and maximum values) can be set differently for each audio object of the content.

Furthermore, MPEG-H Audio can also enable DE functionality for legacy stereo content (e.g., archive material stored in stereo without the original stems). This ensures a consistent user experience since the same functionality can be offered not only for object-based productions but also for existing stereo archive material.

The first and probably most important step is to obtain a "clean Dialog" version from the stereo content. This is achieved using a so-called “Dialog Separation” (DS) pre-processing technology. Several DS solutions are available on the market and can be used together with MPEG-H Audio. Relying on an open format such as BWF/ADM ensures interoperability of the MPEG-H system with existing and future solutions for Dialog Separation.

2) Audio Description

Similarly, the MPEG-H Audio system allows the delivery of Audio Description (AD) in multiple languages. AD services can be enabled by automatic device selection (prioritization) as well as by manual user selection. For each AD object, all advanced interactivity options are available and can be enabled by the broadcaster. The AD level can be adjusted independently and moreover, MPEG-H Audio is the only standardized audio system that allows the user to spatially move the Audio Description to a user selected position (e.g., to the left or right). This enables a spatial separation of main dialog and Audio Description, leading to a better intelligibility of the main dialog as well as of the Audio Description.

The system supports advanced connectivity use cases, where the Audio Description can be also provided via a Bluetooth channel for example to the headset of the person requiring the Audio Description. This person can now get same enhanced experience with the AD, while the rest of the family is experiencing the content without the AD. The standardized interfaces in MPEG-H Audio, ISO/IEC 23008-3 [2] can enable receivers and application layers to implement such advanced used cases.

3) Multi-language services

With existing audio codecs, multi-language programs are broadcasted as separate complete mixes in each language. Using one stream for each mix requires a high bit rate, directly proportional to the number of additional languages offered. Moreover, if Audio Description services have to be provided as additional complete mixes, the required bandwidth would significantly increase.

MPEG-H Audio enables a much more efficient way of offering accessibility and multi-language services by making use of object-based audio, similar to the DE feature, as described in previous sections. With a common channel bed and individual audio objects for each language dialog and audio description tracks, MPEG-H Audio requires a

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

Fig. 9. Example of multi-language labels (English – Upper side, French – Lower side).

significantly lower bit rate than legacy systems. For example, a 5.1 program is delivered in 5 different languages in a single stream using one audio object for each language. A legacy system would require transport of six complete 5.1 mixes in five different streams.

4) *Presentation of* services

The MPEG-H Audio metadata uses textual labels for describing the presets and the audio objects in multiple languages. The content creator can decide based on the regions where its content is distributed to author all labels in one or more languages. Based on the receiver's preferred language setting the correct labels will be displayed to the viewer. Fig. 9 shows the labels authored during a live broadcast trial in two languages: English and French.

Using the MPEG-H metadata, the content creators can ensure that their artistic intent and the various features they want to enable are correctly displayed to the user. In this way content creators are always in control of their content and the users will experience the content in the same way on all devices.

E. *Emergency Warning Information*

Delivery of Emergency Warning Information (EWI) caring emergency alerts, information and instructions to the TV viewers is a key feature of a next generation terrestrial broadcast system. The MPEG-H Audio system was defined including a flexible mechanism for emergency information messages. Multiple audio objects can be signaled as EWI using the MPEG-H Audio content type "mae_contentKind" == 12 ["emergency"], specified in ISO/IEC 23008-3 [2]. The EWI can be signaled as mandatory or optional, enabling the application layer in the receiver device to reproduce optional EWI messages based on the receiver settings or geo-location information. This way, according to the local requirements, the MPEG-H Audio system can be used to offer EWI Audio Description in different ways:

- **Mandatory emergency messages:** One audio object included in all presets and always active. The viewer cannot disable the EWI. Moreover, the EWI message may be delivered in a dedicated preset, replacing the main dialog or even replacing the complete mix
- **Optional emergency messages**: One audio object included in all presets and active based on receiver settings or geo-location of the receivers. These messages can be disabled by the viewer, and they are usually not critical alerts.

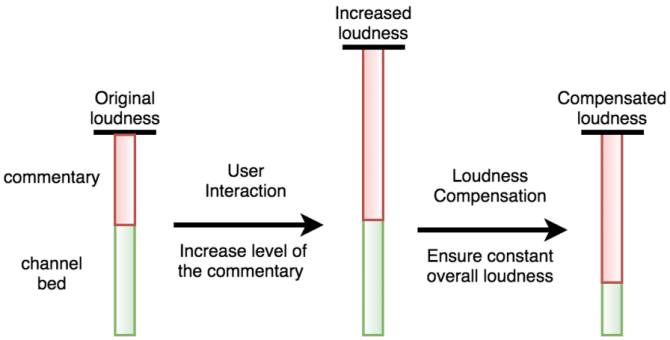

Fig. 10. Illustration of the loudness compensation concept for user interaction with the dialog object gain. The height of each bar corresponds to the loudness portion of the dialog object and the channel bed, respectively.

F. *Consistent Loudness*

The SBTVD TV 3.0 project requires a consistent loudness of the reproduced audio content. The MPEG-H Audio System accomplishes this automatically in two steps: the Loudness Normalization module aligns the loudness between program items to the target loudness of the decoder, while the Loudness Compensation module additionally compensates for loudness changes due to user interaction.

1) *Loudness Normalization*

The MPEG-H 3D Audio standard supports loudness information that is mandatorily included in the metadata of the MPEG-H Audio Stream. Various loudness measurement systems such as ITU-R BS.1770, EBU R128, ATSC A/85 are supported to fulfill applicable broadcast regulations and recommendations. The system allows to specify whether loudness information relates to the loudness of a full program or whether it refers to a specific anchor element of the program, such as the dialog or commentary.

Additionally, the system allows to input at encoding stage loudness information for each available preset separately. This enables immediate and automatic loudness control for interactive and personalized audio. For example, when the user switches between different presets the loudness normalization gain is instantaneously adjusted to ensure consistent playback loudness over all presets.

2) *Loudness Compensation*

The MPEG-H Audio system allows users to interact and control the rendering of individual audio elements which might result in an increase of the overall loudness of the resulting mix compared to the original preset authored in production. This behavior would interfere with the requirement of consistent loudness and preservation of signal headroom. Therefore, the MPEG-H Audio System includes a mandatory tool to compensate for loudness variations due to user interaction with individual audio elements (e.g., increase the dialog level compared to the rest of the mix).

The loudness compensation tool is based on metadata included in the audio stream that provides the loudness for each signal group or object that is part of the program mix. From these individual loudness values, a compensation gain is determined after any gain interaction done by the user, which is then applied together with the loudness normalization gain. The loudness compensation concept is illustrated in Fig. 10, for the example of a program consisting of a dialog object and a channel bed.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

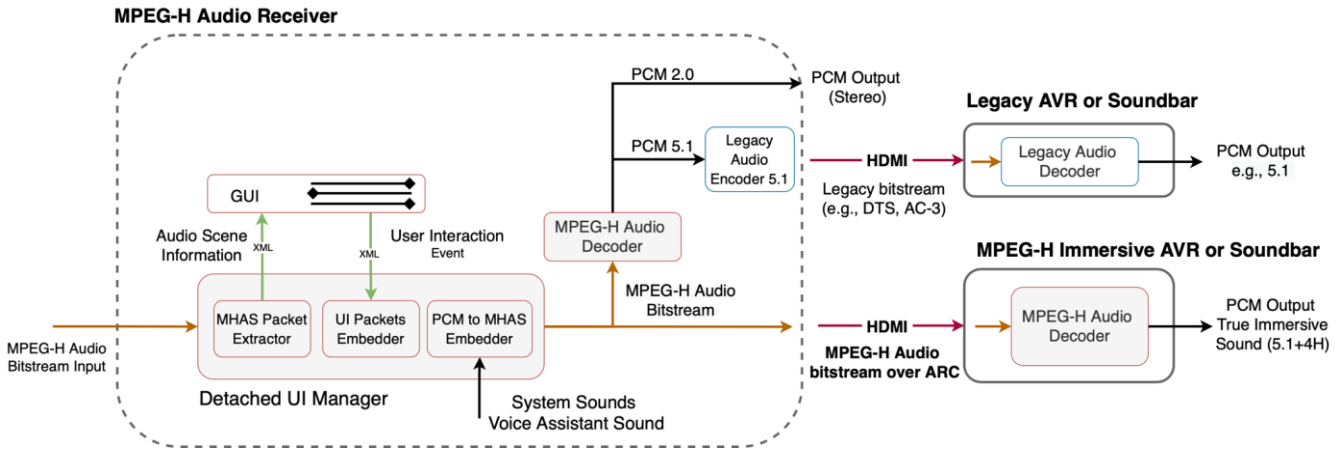

Fig. 11. Top level block diagram of a receiver using MPEG-H Audio

G. *Connectivity to external devices*

The TV 3.0 audio system has to enable the most advanced personalization options and at the same time reproduce the immersive sound, especially when connected to external sound devices. Therefore, the audio component requirements are evaluated by test labs assigned by the SBTVD using various connectivity options. Fig. 10 provides a high-level diagram of a receiver using MPEG-H Audio connected to external playback devices. The signal could be received over RF or IP. Using standardized metadata and interfaces for the systems/application layer, the MPEG-H Audio system enables extremely efficient ways to always provide the best audio experience while offering all its personalization features to the viewers.

Based on the available connections of the receiver (e.g., TV set), different outputs may be used:

- PCM Stereo output,
- HDMI bit stream output for legacy devices, or
- HDMI bit stream passthrough for immersive MPEG-H playback devices (e.g., AVRs and Soundbars).

All connectivity options, requirements and prioritization rules are specified in ABNT NBR 15604:2020, Section C.9 [13].

The MPEG-H Audio system enables a distributed architecture between metadata processing steps and the decoding step. This distributed architecture is enabled by the standardized MPEG-H Audio Stream (MHAS) packetized structure, as specified in ISO/IEC 23008-3, Clause 14 [2]. As seen in Fig. 11 the incoming bit stream is first preprocessed by the so-called "Detached UI Manager" before being sent to the MPEG-H Audio Decoder.

1) *User Interface on systems level*

When a receiver is connected to an external sound system, most of existing audio systems require full decoding, rendering, user interaction and re-encoding to a different audio format to be transmitted over HDMI to the external sound device. This transcoding process is computational complex and introduces additional delay. To avoid such unnecessary transcoding steps, the MPEG-H Audio system is capable to read the metadata required for user interactivity on the MHAS bit stream level without any decoding of the audio data.

As shown in Fig. 11, the Audio Scene Information is extracted from the MPEG-H Audio bit stream at systems level. The "MHAS Packet Extractor" parses the MHAS stream, extracts the MHAS PACTYP_AUDIOSCENEINFO packet and makes it available to the application for usage in a Graphical User Interface (GUI).

In return, the "UI Packets Embedder" accepts the user interactivity information from the application layer. The user interactivity information is encapsulated in the standardized MHAS PACTYP_USERINTERACTION packets, as defined in ISO/IEC 23008-3, Clause 14.4.9 [2]. The user interaction MHAS packets are then inserted "on-the-fly" back into the MHAS packet stream.

If the receiver is using its own loudspeakers or is connected over HDMI to a legacy AVR/Soundbar, the MHAS packet stream is fed into the MPEG-H Audio decoder, which decodes to stereo or 5.1 respectively.

If the receiver is connected over HDMI to an immersive MPEG-H AVR/Soundbar, the MHAS MPEG-H Audio decoder in the receiver is not used and the MHAS stream is provided over HDMI to the external playback device, which will perform the final audio decoding and playout of the audio. All user interactions are embedded into to MHAS stream and will be applied during the decoding in the external sound device.

Using such distributed architecture, the MPEG-H Audio system enables delivery of immersive sound to the end device without any compromise, while at the same time enabling the user interactivity in the receiving device. This unique solution:

- Significantly reduces computational complexity and the audio delay in the receiver,
- Offers the best audio quality for the immersive playback in the external sound system by avoiding any intermediate downmix or rendering in advance of the final playback device,
- Enables the viewer to use a single remote-control for all audio controls and personalization options,
- Provides a consistent UI design independent from the connected external device.

2) *System Sounds and Voice Assistant Sounds*

Providing audible feedback on a user's interaction or system status is important for enhanced accessibility user experience. These system sounds and voice assistant sounds are usually generated by the receiver during playback and guide the user

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

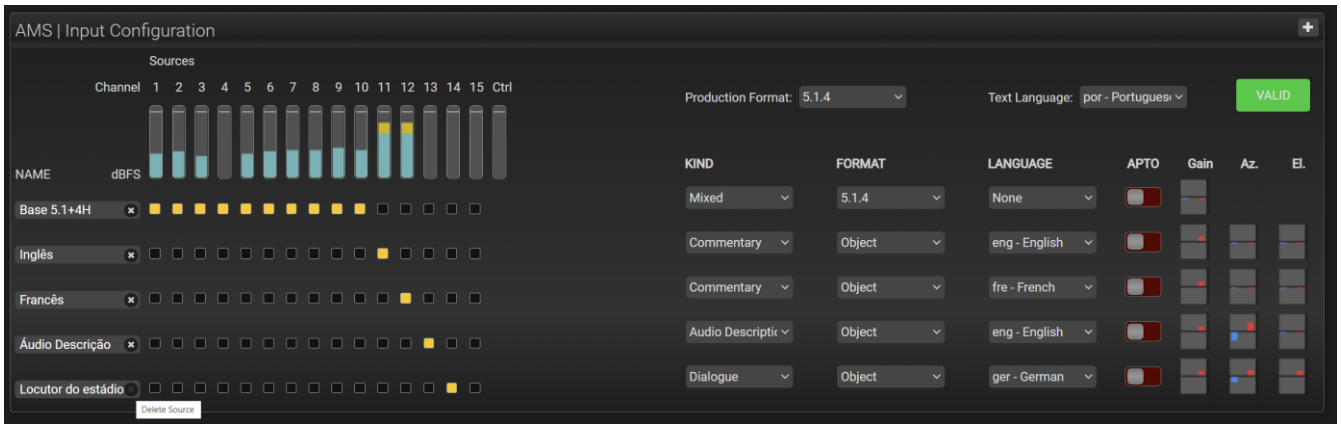

Fig. 12. Linear Acoustic AMS Control Interface – Remove the last audio object example

to navigate through the device options or interactions with the device.

When the receiver is connected to an external sound device these sounds have to be provided together with the main audio to the external device. This is usually achieved by decoding the audio data, mixing the audio data with the system sounds in the receiver and re-encoding for delivery over HDMI to the sound system. As previously described, such process is highly complex, introduces higher audio delay and may lead to compromising the audio quality by intermediate downmix and rendering process.

For ensuring the best possible audio experience, MPEG has specified a mechanism for embedding PCM samples of system sounds or voice assistant audio "on the fly" into an MPEG-H Audio bit stream without decoding the bit stream. The mechanism is based on the capabilities of the MPEG-H Audio system to handle Earcon sounds, the equivalent of visual icons in computer interfaces. Three MHAS packets have been defined for carriage of the PCM data and associated metadata:

- **PACTYP_EARCON** carrying configuration metadata, such as type, id, status (active/inactive), gain, position etc.
- **PACTYP_PCMCONFIG** carrying PCM configuration data such as sampling rate and frame size.
- **PACTYP_PCMDATA** carrying the uncompressed PCM samples.

Using these three standardized MHAS packets, the receiver is capable to embed the system sounds and voice assistant sounds into the received bit stream and deliver the bit stream to an external audio device (AVR, Soundbar) that supports MPEG-H Audio. The external device will decode and render the MPEG-H Audio bit stream and mix in the system sounds and/or voice assistant audio into the rendered audio scene as described in ISO/IEC 23008-3 clause 28.4 [2].

H. Seamless configuration changes in production

The SBTVD TV 3.0 Project was designed to enable the most advanced audio features in existing and future broadcast workflows. This requires seamless playback during changes in production. The broadcasters must be able to change various aspects of the production during live transmission based on their creative intent and offer the best experience to the viewer. Typical changes in a live broadcast will be tested and evaluated during the TV 3.0 Project, including:

- Change of the audio scene (objects, preselections, etc.),
- Enable/disable dialogs and Audio Description in multiple languages,
- Enable/disable interactivity options for one or more preselections,
- Change the interactivity options (min/max gain and position values) for one or more objects, or
- Change the textual labels for one or more objects or preselections.

The MPEG-H metadata used in production, allows alignment to the video frame rate and therefore any change in production (inside an authoring unit or an SDI-level switch) can be seamlessly applied without disturbing the playback on the consumer end devices. Configuration changes during production are translated seamlessly into configuration changes in the bit stream by the encoder and are seamlessly applied in the decoder.

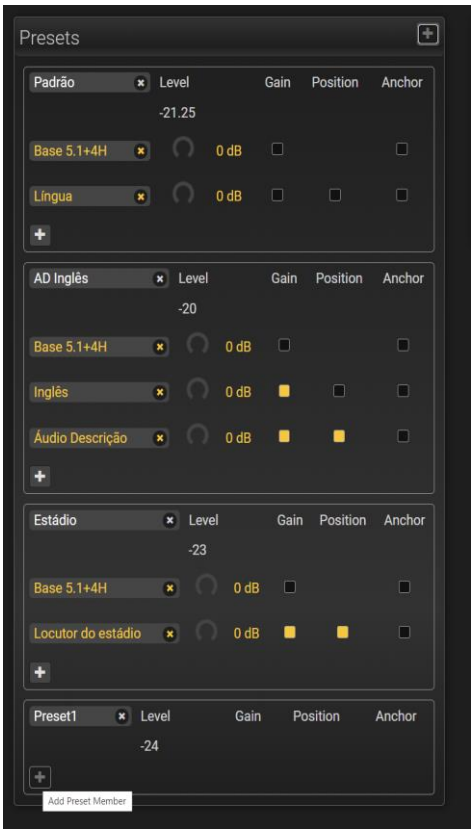

Fig. 13. Linear Acoustic AMS Interface – Add a component to the newly created example.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

Fig. 12 shows an example using the Linear Acoustic AMS authoring unit where the operator is able to remove one audio object in live broadcast. This could be needed in cases where a stadium announcer feed is not available any longer and the broadcaster would prefer to disable the option for the viewers to listen to this audio object. Similarly, the content producer could add an additional audio object, change interactivity options or add a completely new preset as illustrated in Fig. 13. When creating a new preset, the broadcaster can define a new textual label for it, decide which audio components will be part of the preset and what gain and position interactivity options should be allowed.

All these options can be enabled in production using the web control interface of the authoring unit by manually changing each entry or can be configured in advance of the live broadcast and simply uploaded when necessary. Either way, the MPEG-H Audio system will ensure a seamless update of the receiver's user interface.

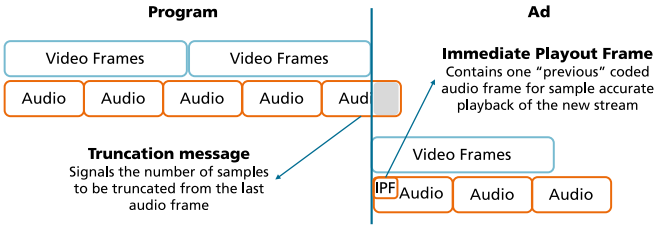

Fig. 14. Sample accurate ad-insertion example.

I. *Seamless ad-insertion and stream splicing*

For legacy systems, various methods have been used for efficiently enabling Ad-insertion. The TV 3.0 Project aims to achieve a seamless behavior during ad-insertion independent of the content used. With Next Generation Audio, it will most likely happen that the live content will use one configuration (e.g., immersive sound with several audio objects and personalization options), while the ad will use a different configuration (e.g., simple stereo without any interactivity options). In such a scenario, four aspects are important for the viewer:

- The transition to and from the ad should be seamless, meaning that no audio dropouts or glitches should occur,
- The ad-insertion can occur at any point in time at a video frame boundary, even in the middle of a coded audio frame,
- The overall perceived loudness level should be preserved before, during and after the ad, and
- The interactivity options displayed to the user should change accordingly, such that they perfectly match the audio content.
- The user shall not be able to change the language or increase the dialog level for example during the ad-break if the ad does not contain such options.
- The user settings should be restored after the ad-break.

For achieving this, the MPEG-H Audio system is using two new concepts: a frame truncation mechanism and an Immediate Playout Frame (IPF) Access Unit (AU).

The MHAS format offers the possibility to transmit truncation information via the AUDIOTRUNCATION packet. The truncation information is used to discard a certain number of audio samples from the beginning or end of a decoded AU. This can be used for alignment of decoded audio data to the video frame boundaries.

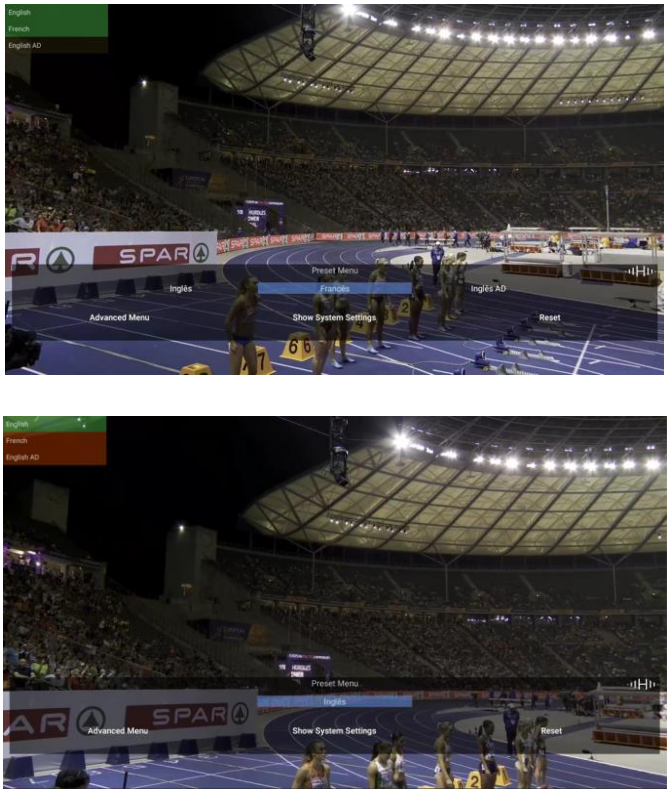

Fig. 15. MPEG-H Receiver Playback of main (broadcast) and two side streams (broadband) [upper picture] and the fallback to broadcast stream when broadband connection is not available [lower picture].

An IPF carries additional information of the "previous" audio frame and thus allows, beyond simple random-access operations, the sample accurate synchronization of audio and video streams. All modern audio codecs require at least two coded AUs for decoding valid audio samples and therefore it is essential to have access to the "previous frame" when changing to a different stream. Defining IPFs at the beginning of the ad or at the beginning of the stream after the ad allows for glitch-free and on-the-fly reconfiguration of the stream.

Fig. 14 illustrates such example, where the last frame of the stream (before the ad-insertion or configuration change) is truncated, and the new stream (after the configuration change) starts with an IPF. This way, the configuration change will be seamless and aligned to the video splicing point in a sample-accurate fashion.

J. *Hybrid delivery and extensibility*

With a forward-looking thinking, the SBTVD Forum has decided from the beginning to select an IP-based transport layer which will allow a smooth interaction between the broadcast and broadband paths. Such approach allows the delivery of a standard program over the broadcast available to all viewers and additional features can be enabled via broadband (e.g., premium commentators, different languages etc.). The TV 3.0 Project requires, for its audio system, such advanced capabilities where the additional features are presented to the user only when available over the broadband connection and fallback to the default settings from the broadcast stream in cases where the broadband connection is suddenly not available. Additionally, all user interactions should not disrupt the audio playback even when the audio components are delivered over the two different transmission paths.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

Fig. 15 illustrates an example of the MPEG-H Audio receiver with hybrid reception using the test content for the TV 3.0 evaluation. In this example, the main stream contains a complete audio scene, meaning that it provides the complete experience (e.g., 5.1+4H immersive sound with English dialog) even without any broadband connection. The additional streams delivered via broadband will contain enhancements of the audio scene (e.g., the French dialog in the second stream and the Audio Description in the third stream).

The upper picture in Fig. 15 shows the available MPEG-H user interaction options when both connections are available, while the lower picture shows the fallback to the options available in the broadcast path only (the side streams are both marked in "red" and not available).

The MPEG-H multi-stream features enable the receiver to easily synchronize the streams and seamlessly switch between the streams. The playback always starts with the main stream (broadcast) but the MPEG-H Audio metadata is used to enable the receiver to display the available streams via broadband without disturbing the main broadcast feed. The user is able to switch between the available interactivity options in a seamless way. Based on the active preset the receiver is requesting the additional streams.

Besides the hybrid delivery which enables scalability of the broadcast system, the TV 3.0 Project requires easy extensibility of the audio system for enabling new applications in the future. The MPEG-H Audio system enables extensibility using well established mechanisms inherited from previous MPEG audio standards as well as through its state-of-the-art packetized bit stream structure: MHAS. Additional MHAS packets may be defined in a backwards compatible way, meaning that existing decoders will simply ignore the unknown newly defined MHAS packets while new decoders will be capable to read and process these newly defined MHAS packets for offering enhanced experiences.

Moreover, MPEG-H Audio was already selected by MPEG as the only audio codec for the future MPEG-I Audio system which will provide support 6 Degrees of Freedom (6DoF) audio playback. The MPEG-I Audio work item, currently under standardization, will define additional metadata which will be embedded in new MHAS packets.

MPEG-H features extension mechanisms on different layers that allow future extensions in well-proven, well-defined, clean, efficient, and backwards-compatible ways.

V. Conclusion

Fraunhofer IIS, ATEME, DiBEG, and ATSC have proposed the MPEG-H Audio system in response to the SBTVD TV 3.0 Call for Proposals. The TV 3.0 Project has been designed to offer the most advanced audio options to the viewers and requires advanced solutions for metadata and audio handling across the entire production and broadcast chain. With requirements for immersive sound, advanced interactivity and accessibility options, hybrid delivery, consistent loudness after user interaction, connectivity options for external sound devices and seamless configuration changes, to be evaluated in a real-time broadcast environment, the TV 3.0 Project has set the ground for the most detailed evaluation of the proposed audio systems.

The MPEG-H Audio system is the only fully standardized audio system fulfilling all TV 3.0 requirements listed in the Call for Proposals and provides the most advanced feature set and use cases as detailed in this document.

With hardware implementations already available and used in 24/7 broadcast, the MPEG-H Audio system can ensure an easy transition from the existing ISDB-Tb broadcast system to the future based TV 3.0 system.

References

[1] Brazilian Digital Terrestrial Television System Forum (SBTVD) TV 3.0 Call for Proposals, Available: https://forumsbtvd.org.br/tv3_0/

[2] ISO/IEC 23008-3:2019: "Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D audio", including ISO/IEC 23008-3:2019/AMD 1:2019 "Audio metadata enhancements" and ISO/IEC 23008-3:2019/AMD 2:2020, "3D Audio baseline profile, corrections and improvements".

[3] MPEG-I Immersive Audio Call for Proposals. Available: https://www.mpegstandards.org/wp-content/uploads/mpeg_meetings/134_OnLine/w20449.zip

[4] N19407, MPEG-H 3D Audio Baseline Profile Verification Test Report. Available: https://www.mpegstandards.org/wp-content/uploads/2020/07/w19407.zip

[5] N16584, MPEG-H 3D Audio Verification Test Report. Available: http://mpeg.chiariglione.org/standards/mpeg-h/3d-audio/mpeg-h-3d-audio-verification-test-report

[6] R. Bleidt et al. "Development of the MPEG-H TV Audio System for ATSC 3.0," in IEEE Transactions on Broadcasting, vol. 63, no. 1, March 2017. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7874294

[7] ATSC A/342-3:2021 "ATSC Standard, A/342 Part 3: MPEG-H System," Advanced Television Systems Committee, Washington, DC, 11 March 2021. Available: https://www.atsc.org/wp-content/uploads/2021/03/A342-2021-Part-3-MPEG-H.pdf

[8] ATSC A/331:2019, "ATSC Standard: Signaling, Delivery, Synchronization, and Error Protection," Advanced Television Systems Committee, Washington, DC, 20 June 2019, https://www.atsc.org/wp-content/uploads/2017/12/A331-2017-Signaling-Deivery-Sync-FEC-1.pdf

[9] TTAK-KO-07.0127R1: TTA - Transmission and Reception for Terrestrial UHDTV Broadcasting Service, Revision 1, December 2016.

[10] MPEG-H Audio selected to enhance Brazilian digital television with immersive and personalized sound, https://www.audioblog.iis.fraunhofer.com/mpegh-brazil-isdbtb

[11] ABNT NBR 15602-2:2020, Televisão digital terrestre - Codificação de vídeo, áudio e multiplexação - Parte 2: Codificação de áudio.

[12] ABNT NBR 15603:2020, Televisão digital terrestre - Multiplexação e serviços de informação (SI), https://forumsbtvd.org.br/legislacao-e-normas-tecnicas/normas-tecnicas-da-tv-digital/english/

[13] ABNT NBR 15604:2020, Televisão digital terrestre – Receptores, https://forumsbtvd.org.br/legislacao-e-normas-tecnicas/normas-tecnicas-da-tv-digital/english/

[14] ETSI TS 126 118 V15.0.0 (2018-10), 5G; 3GPP Virtual reality profiles for streaming applications (3GPP TS 26.118 version 15.0.0 Release 15).Available:https://www.etsi.org/deliver/etsi_TS/126100_126199/126118/15.00.00_60/ts_126118v150000p.pdf

[15] TS 101 154 v2.3.1: Digital Video Broadcasting (DVB) – Specification for the use of Video and Audio Coding in Broadcasting Applications based on the MPEG-2 Transport Stream.

[16] ETSI EN 300 468 V1.16.1 (2019-08), Digital Video Broadcasting (DVB); Specification for Service Information (SI) in DVB systems, https://www.etsi.org/deliver/etsi_en/300400_300499/300468/01.16.01_60/en_300468v011601p.pdf

[17] International Telecommunications Union (ITU) Recommendation ITU-R BS.1196-7 (01/2019), Audio coding for digital broadcasting https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1196-7-201901-S!!PDF-E.pdf

[18] South Korea launches UHD TV with MPEG-H Audio, https://www.audioblog.iis.fraunhofer.com/south-korea-uhd-tv-mpeg-h

[19] A. Murtaza and S. Meltzer, "First Experiences with the MPEG-H TV Audio System in Broadcast," SET INTERNATIONAL JOURNAL OF BROADCAST, 2018 Available: https://www.set.org.br/ijbe/ed4/Artigo%206.pdf

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

[20] Shows do Rock in Rio transmitidos com tecnologia desenvolvida pelo Fraunhofer IIS, https://panoramaaudiovisual.com.br/shows-do-rock-in-rio-transmitidos-com-tecnologia-desenvolvida-pelo-fraunhofer-iis/
[21] Centro de treinamento de áudio MPEG-H no Brasil, https://www.brazil.fraunhofer.com/pt/news/noticias-mais-recentes/novo-centro-de-treinamento-de-audio-mpeg-h-abre-suas-portas-em-s.html
[22] Studio Recommendations for 3D Audio productions with MPEG-H Audio, https://www.iis.fraunhofer.de/content/dam/iis/de/doc/ame/wp/FraunhoferIIS_TechnicalPaper_Studio_Recommendations_3DAudio-MPEG-H.pdf
[23] ITU-R BS.2076-2, Recommendation ITU-R BS.2076-2, Audio definition model, Geneva 10/2019
[24] ITU-R BS.2125, Recommendation ITU-R BS.2125, A serial representation of the Audio Definition Model, Geneva 01/2019.
[25] The MPEG-H Authoring Suite, https://www.iis.fraunhofer.de/en/ff/amm/dl/software/mas.html
[26] The Spatial Audio Designer (SAD) Plugin, https://newaudiotechnology.com/products/spatial-audio-designer/
[27] The MPEG-H ADM Profile. Available: https://www.iis.fraunhofer.de/en/ff/amm/dl/whitepapers/adm-profile.html

Adrian Murtaza received his M.Sc. degree in Communication Systems from the École Polytechnique Fédérale de Lausanne, Switzerland in 2012 with a thesis on "Backward Compatible Smart and Interactive Audio Transmission". Upon graduation he joined Fraunhofer IIS, where he works as a Senior Manager, Technology and Standards.

Adrian joined MPEG in 2013 and since then contributed to the development of various audio technical standards in MPEG-D and MPEG-H. He serves as Fraunhofer's Standards Manager in a number of industry standards bodies, including SBTVD, ATSC, CTA, DVB, HbbTV and SCTE, and is the co-author of multiple specifications in those groups.

More recently he focused on specification of Next Generation Audio delivery and transport in ATSC 3.0 systems and MPEG-2 Transport Stream based systems, as well as on enabling of MPEG-H Audio services in different broadcast and streaming ecosystems. With a strong interest in VR/AR media solutions he is actively involved in MPEG-I efforts targeting future immersive applications.

Stefan Meltzer studied electrical engineering at the Friedrich-Alexander University in Erlangen, Germany. In 1990 he joined the Fraunhofer Institute for Integrated Circuits (IIS) in Erlangen, Germany. After working in the field of IC design for several years, Stefan became the project leader for the development of the WorldSpace Satellite Broadcasting system in 1995 and in 1998 of the XM Satellite Radio broadcasting system. His team was responsible for the system design, chip set design, field trials and development of a reference signal generator.

In 2000 he joined Coding Technologies in Nuremberg as Vice President for business development, Germany. His responsibilities included broadcasting and consumer electronics. During his time at Coding Technologies, HE-AAC was accepted in numerous broadcasting standards and applications. After Coding Technologies was acquired by Dolby Labs, Stefan joined Iosono as CTO in April 2008.

From January 2010, Stefan worked as independent technology consultant with the focus on audio and multimedia. In this role he supported Fraunhofer IIS in the business development and marketing activities within the TV broadcast market. In April 2018, Stefan joined Fraunhofer IIS again and is now in charge of business development for TV broadcast applications.

Yannik Grewe received his M.Eng. degree in 'audiovisual media – sound' with a thesis on 'Perception and reproduction of floor level sound in consumer audio playback'. Yannik joined Fraunhofer IIS in 2013 and serves today as senior engineer for audio production technologies, focusing MPEG-H 3D Audio. He is extensively involved as a sound engineer in producing immersive music applications and MPEG-H immersive and interactive audio. His current role includes a close relation to major broadcasters and streaming service providers in Asia, Europe, and South America to enable MPEG-H Audio in their ecosystems.

Nicolas Faecks received a B.Sc. degree in 'media technologies and a M.A. degree in 'time-based media – Sound – Vision'. Before joining the Fraunhofer Institute for Integrated Circuits IIS in 2014, he was a researcher on lighting technologies at Airbus. At Fraunhofer, he focused on All-IP-Workflows and MPEG-H 3D Audio Systems. As a System Engineer, he was responsible for the MPEG-H 3D Audio broadcast and streaming systems during major events, such as Roland Garros, the European Athletics Championships, the Eurovision Song Contest or the Youth Olympic Games.

Lucas Gregory graduated from Université Polytechnique des Hauts de France (INSA HDF) in 2017 with a Master Degree in Audio and Video System Engineering. He then joined the ATEME Research and Innovation team, where he helped in several French collaborative projects to study and promote innovative technologies such as 360° video, Next Generation Audio codecs and Next-Gen TV (ATSC 3.0). Today, he works on immersive audio technologies as well as the trans-packaging of low-latency content.

Dr. Mickaël RAULET received his Ph.D. degree in 2006. He joined ATEME in 2015, where he is now CTO. He is leading the standardization effort at ATEME and is following the different activities in ATSC, DVB, 3GPP, ISO/IEC, ITU, MPEG, DASH-IF and UHD Forum. He is managing several collaborative R&D projects for ATEME.

Received in 2021-08-25 | Approved in 2021-12-07

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3