Audio Engineering Society

Convention e-Brief 649

Presented at the 151st Convention
2021 October, Online

MPEG-H Audio Production Workflows for a Next Generation Audio Experience in Broadcast, Streaming, and Music

Yannik Grewe[1], Philipp Eibl[1], Daniela Rieger[1], Matteo Torcoli[1], Christian Simon[1], and Ulli Scuda[1]

[1] *Fraunhofer Institute for Integrated Circuits IIS, Am Wolfsmantel 33, 91058 Erlangen, Germany*

Correspondence should be addressed to Yannik Grewe (yannik.grewe@iis.fraunhofer.de)

ABSTRACT

MPEG-H Audio is a Next Generation Audio (NGA) system offering a new audio experience for various applications: Object-based immersive sound delivers a new degree of realism and artistic freedom for immersive music applications, such as the 360 Reality Audio music service. Advanced interactivity options enable improved personalization and accessibility. Solutions exist, to create object-based features from legacy material, e.g., deep-learning-based dialogue enhancement. 'Universal delivery' allows for optimal rendering of a production over all kinds of devices and various ways of distribution like broadcast or streaming. All these new features are achieved by adding metadata to the audio, which is defined during production and offers content providers flexible control of interaction and rendering options. Thus, new possibilities are introduced, but also new requirements during the production process are imposed. This paper provides an overview of production scenarios using MPEG-H Audio along with examples of state-of-the-art NGA production workflows. Special attention is given to immersive music and broadcast applications as well as accessibility features.

1 Introduction

MPEG-H Audio is a Next Generation Audio (NGA) system based on the open international standard ISO/IEC 23008-3, MPEG-H 3D Audio [1].

It supports sound all around as well as above and below the listeners, enabling an immersive audio experience via broadcast, streaming, and music services.

In MPEG-H Audio, scenes can consist of object-based (OBA), channel-based, and scene-based content, or any combination of them [2, 3].

Even though a three-dimensional auditory scene generally contains a larger amount of audio data than a legacy surround mix, this is mostly compensated for by the high bitrate efficiency of the MPEG-H Audio codec. Typical 3D Audio setups such as 5.1+4H or 7.1+4H can be broadcast using bit rates ranging from 192 kbit/s to 384 kbit/s, which are average values for current 5.1 transmissions [3].

With its unique interactivity features, MPEG-H Audio offers listeners the flexibility to actively engage with the content and adapt it to their own preference by selecting pre-defined presets e.g. “Default Mix”, “Dialogue Enhanced Audio”, and “Venue Sound.”

“Universal Delivery” enables that a single MPEG-H stream can deliver content via streaming or broadcast platforms to all kinds of receivers and playback devices, from headphones to sound bars and discrete loudspeaker systems. The built-in renderer adapts the content to the playback capabilities of the end-consumer device.

The adoption of immersive and interactive audio introduces new possibilities, but also poses new requirements for production, distribution, and reproduction. Efficient NGA production workflows need to be defined without compromising well-established production infrastructure.

Production tools were developed to offer efficient workflows for live and post-production [4]. During NGA production trials, those workflows were evaluated with a focus on immersive music services, interactive broadcasts, as well as solutions to add advanced accessibility features based on legacy stereo mixes. The workflows for these productions are described in this paper.

2 Recording and Mixing

NGA overcomes the limitations of its legacy predecessors. Nevertheless, the main principles of its audio production workflows are similar. Independent of the production scenario, be it broadcast or music, MPEG-H Audio can be produced in a traditional fashion. However, 3D Audio signals and audio objects might be created, processed, or extracted (see 3.6). During major broadcast and streaming events, like the Rock in Rio music festival [5], the European Athletics Championships (EAC) [6], the 2020 Youth Olympic Games or the Eurovision Song Contest (ESC) [7], it has been shown that traditional recording principles remain valid but are enhanced by capturing techniques for height loudspeaker layer reproduction. Several microphone techniques to capture three-dimensional auditory images have already been introduced and compared [8, 9]. During ESC, the height effect was generated by adding an additional 4-channel Hamasaki Square microphone array to the existing crowd microphones (see Fig. 1) [10]. At EAC, capturing height layer information was done using an ORTF 3D-Flat on the arena's stand [6, 11]. Additionally, 3D Audio bus structures within Digital Audio Workstations (DAW) or broadcast mixing and monitoring paths as well as 3D-Panning tools are required. When using additional audio objects, they must be kept separate from the other elements of the mix, such as Music & Effects stems. The process of generating metadata is called "authoring" and it is the most important difference between the production of MPEG-H Audio and legacy audio creation.

After the authoring, audio and metadata are exported to an intermediate file, called MPEG-H Audio Master. Within an MPEG-H Audio Master, metadata are always transmitted alongside the produced audio essence. This ensures the integrity of the describing control data in all parts of the production and transmission chain.

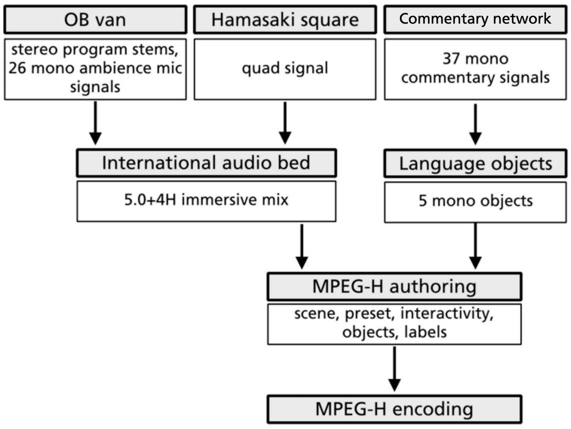

Fig. 1: Schematic audio production and authoring workflow of the ESC MPEG-H production.

3 Authoring of MPEG-H Audio Scenes

Metadata are a key element of NGA systems and serve a variety of purposes, from optimized playback on very different target devices to improved accessibility features.

3.1 Metadata structure

The MPEG-H Audio system standardizes a set of metadata to define an audio scene. The set of metadata consist of descriptive metadata, control metadata, playback-related metadata, and structural metadata [12].

Metadata inform the encoder and decoder about the logical structure of signals and their nature.

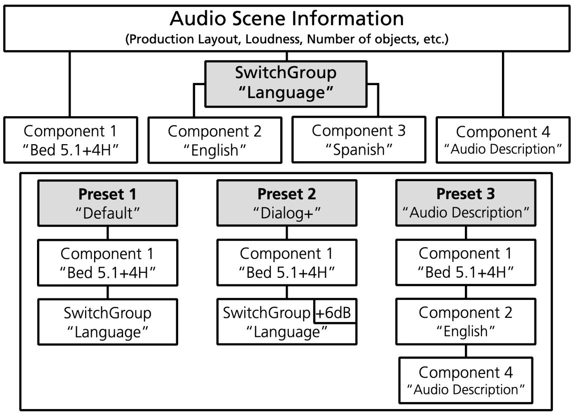

Fig. 2. Metadata structure: components and presets.

Depending on the production scenario, the level of complexity can vary from complex audio scenes with multiple channels, objects, presets and user interactivity (e.g., for broadcast applications), to simple object-only productions for immersive music services, like 360 Reality Audio.

In Fig. 2, 13 signals are used to create an MPEG-H Audio scene. Ten of them are grouped into a channel-based component in a 5.1+4H layout. Signals 11 and 12 are objects combined in a 'Switch Group'-component, which gives them a mutually exclusive characteristic, meaning that only one language can be played back at a time. Signal 13 is an audio description track. These assignments and the way these components are arranged in presets for easy 'single-button-push' user interactivity, are examples for structural metadata. The same elements of the scene also possess descriptive metadata, such as text labels, to identify a Component or Preset on a TV user interface. MPEG-H Audio offers the possibility to transmit multiple sets of labels in different languages.

Other metadata include flags for the language of components, or mark a preset as "Audio Description", "Hearing Impaired", or "Emergency Broadcast." These settings can trigger certain behaviors in the playback device, e.g., selecting the dialogue enhanced presentation by default or always switching to the viewer's preferred language if available.

While these properties are actively assigned during authoring, other datasets are created automatically by the production software. These include loudness information on Components and Presets and Dynamic Range Control (DRC) sets which can be triggered to optimize playback on different types of devices or in various playback environments.

3.2 MPEG-H Master

After the authoring, audio and metadata are exported to an MPEG-H Audio Master. The MPEG-H Audio Master contains linear PCM audio and can be archived, edited, and sent to an encoder. The most common data formats for an MPEG-H Audio Master are MPF (MPEG-H Production Format) or ADM (Audio Definition Model) in either BWF/ADM, according to ITU-R BS.2076-2 [13], or S-ADM, according to ITU-R BS.2125-0 [14].

An MPF file is a multi-channel wav in which the metadata is modulated into a timecode-like audio signal, which is typically located on the last channel of the wav file, called "Control Track". It can be created and transmitted in real time and intercut with different MPFs ('Configuration Change' e.g., for seamless ad-insertion), multiplexed with video, and passed through legacy audio transmission chains such as signal routers, AD/DA, and sample rate converters. The Control Track, due to its robust nature, does not require to put audio equipment into data mode or non-audio mode to pass through. The MPF contains all required metadata as well as the current set of dynamic metadata, such as object positions, and can be cut at arbitrary frames. In a future IP-based production facility, the audio and metadata would be transmitted in a container over IP according to e.g., SMPTE ST-2110.

3.3 Live production

Live productions pose very specific requirements for authoring tools, for instance real-time creation and editing of metadata as well as loudness measurement and level control. At the same time, the engineer needs to be able to monitor the application of metadata, to preview different downmixes and interactivity options. These requirements resulted in the development of Audio Monitoring and Authoring Units (AMAU), a device class tailored to this exact application [15].

An AMAU provides monitoring outputs and the audio and metadata stream which can be fed into a contribution or emission encoder, including additional renderings of the immersive mix for legacy distribution (e.g., stereo or 5.1 surround).

3.4 Post-production

In a post-production environment, software tools for authoring offer a higher degree of flexibility than hardware-based solutions. The benefits range from the use of a DAW timeline for automation to offline exports instead of real-time playout.

There are several authoring tools and plug-ins available, which are optimized for different applications, such as immersive music or broadcast production [4].

3.5 Conversion and ingest

An MPF can also be converted to and from a BWF/ADM representation of an MPEG-H Master.

The MPEG-H ADM Profile defines constraints on ITU-R BS.2076 and ITU-R BS.2125 that enable interoperability with established NGA content production and distribution systems, meaning that e.g., Dolby Atmos BWF/ADM files can be converted to MPEG-H BWF/ADM for feature enrichment and further distribution.

3.6 Accessibility and Dialogue Enhancement

Today's broadcasting and streaming audiences are very heterogenous. They span across a broad spectrum of ages, hearing and sight levels, language skills, and cognitive functions [16]. Additionally, there are a number of different playback conditions, which may affect accessibility. Reported difficulties in following speech because of too loud background sounds were documented 30 years ago [17] and are still relevant [18].

Media presentation should therefore be adapted to everyone's needs and tailored to each individual playback situation. This can be realized efficiently with NGA systems.

MPEG-H Audio features the possibility of adjusting the relative audio level of the dialogue, which is particularly well received by the audience [18].

MPEG-H Audio can also deliver multiple language versions, simplified speech, and audio descriptions (AD) in one audio stream, as part of the regular broadcast. This has various advantages, such as easy and efficient delivery of multiple languages and AD with immersive sound quality and the saving of transmission bandwidth.

Taking full advantage of these capabilities requires the use of separated elements as audio objects. New productions can take this into account, but a workflow solution for large archives with existing content (e.g., available only as stereo mix) also must be provided. For the important use case of personalizing the dialogue level, this can be achieved with deep-learning-based source separation as done by the MPEG-H Dialog+ technology [18]. Dialog+ contains a deep neural network trained for extracting dialogue from a legacy mixture. The separation network is combined with automatic remixing so that the dialogue and the background stems are generated along with accompanying metadata, including dynamic gains. At the user-end, personalization is also available for legacy content. Dialog+ was recently employed in a large-scale field test with German public broadcasters [18].

4 Delivery

Until the final encoding and delivery process, the audio is uncompressed PCM. Over an existing SDI-based, future IP-based or file-based infrastructure, the audio and metadata can be sent to an MPEG-H enabled A/V-hardware- or software-encoder. The signals are efficiently compressed into an MPEG-H Audio bitstream and can be distributed over various paths, like broadcast and streaming. Fig. 3 illustrates the production and delivery path of NGA compared to legacy workflows.

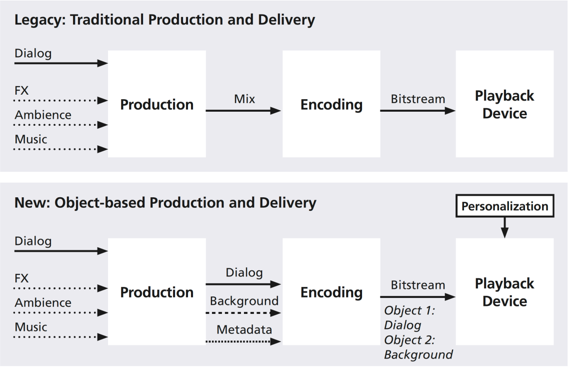

Fig. 3. Legacy and NGA production and delivery.

5 Decoding and Rendering

When delivered to the end-user's device, the MPEG-H Audio bitstream is decoded and the audio signals are rendered based on metadata and user interaction. For channel-based components, a format converter is used to generate high-quality renderings and convert the decoded channel signals to numerous output formats, for example for playback on different loudspeaker layouts, while preserving the original timbre. Object-based signals are rendered to the target reproduction loudspeaker layout by the object renderer, which maps the signals to loudspeaker feeds based on the metadata and the locations of the loudspeakers in the reproduction room.

Additionally, MPEG-H features binaural rendering. During the decoding/rendering process, the content loudness is normalized based on metadata. Advanced loudness handling mechanisms are employed to ensure consistent playback loudness, even in case of user interaction.

6 Conclusions

MPEG-H Audio provides a set of new features, such as immersive sound, personalized audio, and universal delivery, which have already been used in numerous production scenarios, for instance the Eurovision Song Contest 2019, the Youth Olympic Games in 2020, or the 360 Reality Audio music service. These new features are based on accompanying metadata and create the need for new production workflows. Numerous NGA production aspects were described, focusing on recording and mixing practices used in the creation of immersive and interactive content. In addition to that, the authoring step of MPEG-H was introduced, and the delivery and rendering was described.

References

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

[2] Grewe, Y., et al., “Producing Next Generation Audio using the MPEG-H TV Audio System”, BEITC at NAB (2018).

[3] Herre, J., et al. “MPEG-H Audio – The New Standard for Universal Spatial / 3D Audio Coding”, AES 137th Conv. (2014).

[4] Eibl, P., et. al., “Production Tools for the MPEG-H Audio System”, AES 151st Conv. (2021).

[5] Kuwabara, H., et al, “Demonstration on Next-Generation Immersive Audio in a Live Broadcast Workflow”, BEITC at NAB (2020).

[6] De Jong, F., et al, “European Athletics Championships: Lessons from a Live, HDR, HFR, UHD, and NGA Sports Event”, SMPTE Annu. Tech. Conf. Exhibit (2018).

[7] Turnwald, A., et al. “Eurovision Song Contest 2018–immersive and interactive”, 30th Tonmeistertagung (2018).

[8] Grewe, Y., Scuda, U. “Comparison of main microphone systems for 3D-Audio recordings”, 29th Tonmeistertagung (2016).

[9] Lee, H., “Multichannel 3D Microphone Arrays: A Review“, JAES Vol. 69, No. 1/2, pp. 5-26 (2021).

[10] Hamasaki, K., et al. “Reproducing spatial impression with multichannel audio”, AES 24th Conf. (2003).

[11] Wittek, H., Theile, G. “Development and application of a stereophonic multichannel recording technique for 3D Audio and VR”, AES 143rd Conv. (2017).

[12] Bleidt, R., et al., “Development of the MPEG-H TV Audio System for ATSC 3.0”, IEEE Transactions on Broadcasting, Vol. 63, No. 1, pp. 202-236 (2017).

[13] ITU-R BS.2076-2, “Audio Definition Model” (2019).

[14] ITU-R BS.2125-0, “A serial representation of the Audio Definition Model” (2019).

[15] Poers, P., “Monitoring and Authoring of 3D Immersive Next Generation Audio formats”, AES 139th Conv. (2015).

[16] Simon, C., et al. "MPEG-H Audio for Improving Accessibility in Broadcasting and Streaming", arXiv:1909.11549 (2019).

[17] Mathers, C. D., “A Study of Sound Balances for the Hard of Hearing”, BBC White Paper (1991).

[18] Torcoli, M., et al., “Dialog+ in Broadcasting: First Field Tests Using Deep-Learning-Based Dialogue Enhancement”, Int. Broadcasting Conv. (IBC) Technical Papers (2021).