Development of the MPEG-H TV Audio System for ATSC 3.0

Robert L. Bleidt, *Senior Member, IEEE*, Deep Sen, *Senior Member, IEEE*, Andreas Niedermeier, Bernd Czelhan, Simone Füg, Sascha Disch, *Member, IEEE*, Jürgen Herre, *Senior Member, IEEE*, Johannes Hilpert, Max Neuendorf, Harald Fuchs, *Member, IEEE*, Jochen Issing, Adrian Murtaza, Achim Kuntz, Michael Kratschmer, Fabian Küch, Richard Füg, *Member, IEEE*, Benjamin Schubert, Sascha Dick, Guillaume Fuchs, Florian Schuh, Elena Burdiel, Nils Peters, *Member, IEEE*, and Moo-Young Kim, *Senior Member, IEEE*

(Invited Paper)

***Abstract*—A new TV audio system based on the MPEG-H 3D audio standard has been designed, tested, and implemented for ATSC 3.0 broadcasting. The system offers immersive sound to increase the realism and immersion of programming, and offers audio objects that enable interactivity or personalization by viewers. Immersive sound may be broadcast using loudspeaker channel-based signals or scene-based components in combination with static or dynamic audio objects. Interactivity can be enabled through broadcaster-authored preset mixes or through user control of object gains and positions. Improved loudness and dynamic range control allows tailoring the sound for best reproduction on a variety of consumer devices and listening environments. The system includes features to allow operation in HD-SDI broadcast plants, storage, and editing of complex audio programs on existing video editor software or digital audio workstations, frame-accurate switching of programs, and new technologies to adapt current mixing consoles for live broadcast production of immersive and interactive sound. Field tests at live broadcast events were conducted during system design and a live demonstration test bed was constructed to prove the viability of the system design. The system also includes receiver-side components to enable interactivity, binaural rendering for headphone, or tablet computer listening, a "3D soundbar" for immersive playback without overhead speakers, and transport over HDMI 1.4 connections in consumer equipment. The system has been selected as a proposed standard of ATSC 3.0 and is the sole audio system of the UHD ATSC 3.0 broadcasting service currently being deployed in South Korea.**

***Index Terms*—ATSC 3.0, MPEG-H, TV sound, immersive audio, loudness control, interactive audio, audio objects, 3D audio.**

I. Introduction

IN 2014, ATSC issued a call for proposals for a new TV audio system. The envisioned ATSC 3.0 standard was conceived to support new audio capabilities including immersive sound and the ability to offer preset mixes of audio objects sent in the broadcasts. Also, advanced features were part of the requirements, such as the ability for a viewer to create his or her own mix of transmitted objects, the combining of objects sent over the internet with the broadcast program, immersive sound over headphones using binaural rendering, and the potential to correct or adjust the sound image for the consumer's hardware.

Concurrent with the subsequent development of ATSC 3.0, work was being completed on the ISO/IEC MPEG-H 3D Audio standard [1]. A consortium of companies, including those of the authors, proposed the use of MPEG-H for ATSC 3.0 and developed an audio system for ATSC 3.0 based on the MPEG-H audio codec and renderer. This system was successfully evaluated during feature and listening tests conducted by the ATSC audio committee and is now a proposed standard [2] in the ATSC 3.0 suite of standards. Recently, the system was selected by South Korea as the sole audio codec for their deployment of ATSC 3.0.

This paper describes the new features of the MPEG-H 3D Audio standard and its use in the development of the MPEG-H based TV Audio System adopted within ATSC 3.0. In this paper, unless explicit mention is made of the ISO/IEC MPEG-H 3D Audio standard, "MPEG-H" will refer to the MPEG-H based TV Audio System of ATSC 3.0.

This paper is organized into the following major sections:

- New Features of Next-Generation Audio Systems
- Core Codec of the MPEG-H TV Audio System
  - Coding Tools
  - Scene-based Audio

Manuscript received November 21, 2016; accepted December 20, 2016. Date of current version March 1, 2017.

R. L. Bleidt is with Fraunhofer USA Digital Media Technologies, San Jose, CA 95112 USA (e-mail: codecs@dmt.fraunhofer.org).

D. Sen, N. Peters, and M.-Y. Kim are with Qualcomm Technologies, Inc., San Diego, CA 92121 USA (e-mail: dsen@qti.qualcomm.com; npeters@qti.qualcomm.com; kimyoung@qti.qualcomm.com).

A. Niedermeier, B. Czelhan, S. Füg, S. Disch, J. Herre, J. Hilpert, M. Neuendorf, H. Fuchs, J. Issing, A. Murtaza, A. Kuntz, M. Kratschmer, F. Küch, R. Füg, B. Schubert, S. Dick, G. Fuchs, F. Schuh, and E. Burdiel are with Fraunhofer Institute for Integrated Circuits, 91058 Erlangen, Germany (e-mail: andreas.niedermeier@iis.fraunhofer.de; bernd.czelhan@iis.fraunhofer.de; simone.fueg@iis.fraunhofer.de; sascha.disch@iis.fraunhofer.de; juergen.herre@iis.fraunhofer.de; johannes.hilpert@iis.fraunhofer.de; max.neuendorf@iis.fraunhofer.de; harald.fuchs@iis.fraunhofer.de; jochen.issing@iis.fraunhofer.de; adrian.murtaza@iis.fraunhofer.de; achim.kuntz@iis.fraunhofer.de; michael.kratschmer@iis.fraunhofer.de; fabian.kuech@iis.fraunhofer.de; richard.fueg@iis.fraunhofer.de; benjamin.schubert@iis.fraunhofer.de; sascha.dick@iis.fraunhofer.de; guillaume.fuchs@iis.fraunhofer.de; florian.schuh@iis.fraunhofer.de; elena.burdiel@iis.fraunhofer.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2017.2661258

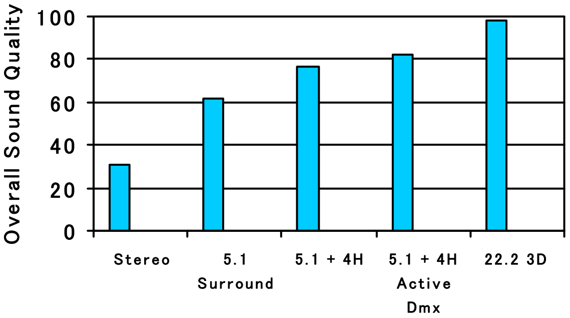


Fig. 1. Overall sound quality improvement with expansion of reproduction system to surround and immersive / 3D formats compared to 22.2 channel reference signal (from [3]).

- Audio Scenes and Related Metadata
- MPEG-H Audio Rendering
- Loudness and Dynamic Range Control
- MPEG-H Transport
- Adapting MPEG-H for TV Broadcasting
  - Field Tests
  - New Features for TV Broadcast Systems
  - System Test Bed for Verification
- Adapting MPEG-H for Consumer Delivery
- MPEG-H and the ATSC 3.0 Development Process

II. New Features of Next-Generation Audio Systems

The TV audio systems under current development have been labeled “next-generation”. They expand prior TV audio systems in three main functional areas.

A. Immersive Sound

Next-generation systems offer immersive sound, distinguished from surround sound by expanding the sound image in the vertical dimension for “3D” reproduction. This offers the user a greater sense of realism and requires less suspension of disbelief to feel he is a part of the scene or program instead of being a remote viewer. Studies such as [3] have shown that perceived improvement in overall sound quality from surround sound to immersive sound can be as large as from stereo to surround sound, as shown in Fig. 1.

Immersive sound can be carried in three primary ways: traditional channel-based sound where each transmission channel is associated with a studio loudspeaker position; sound carried through audio objects, which may be 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 basis functions.

1) Channel-Based Immersive Sound: In channel-based transmission, immersive sound is traditionally carried by building upon the traditional 5.1 and 7.1 surround sound speaker configurations [5] currently used for TV and Blu-ray media. Overhead loudspeaker channels are added to these surround configurations to create sound from above the viewer. Typically, four loudspeakers are used as shown in Fig. 2, though occasionally only two front speakers may be used in limited circumstances. The nomenclature *m.n* + *h*H or *m.n.h* has been introduced in the literature where *m* is the number of loudspeakers in the horizontal plane, *n* is the number of LFE channels, and *h* is the number of overhead or “height” speakers. Some configurations are more complex and do not follow this convention, such as the 22.2 and 10.2(b) configurations previously established [6]. In MPEG standards, ambiguity in loudspeaker configurations is avoided by the use of an index table in the MPEG Coding-Independent Code Points (CICP) standard [4].

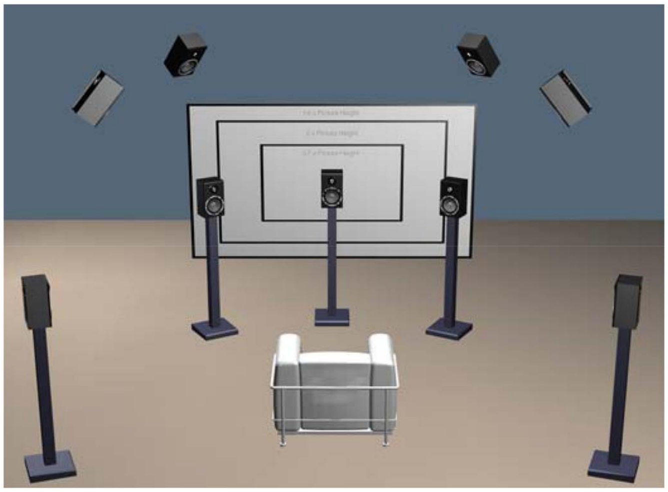


Fig. 2. 5.1 + 4H or 5.1.4 Loudspeaker arrangement typical of channel-based immersive sound. This is CICP index 16 [4].

The MPEG-H TV audio system supports CICP loudspeaker configurations which include stereo, 5.1 surround, 7.1 surround, 5.1+4H, 7.1+4H, 10.2(b), and 22.2 arrangements, among others.

2) Object-based Immersive Sound: Object-based audio has previously been used in gaming and multimedia systems such as VRML and MPEG-4 BIFS to locate sound sources in a three-dimensional universe. In these cases, sound objects may be occluded by structures and a user may move through the universe or world interactively. In contrast, audio objects for video or film use are typically mapped to a fixed coordinate system centered on a translationally static listener without occlusion or room reverberation modeling.

Objects eventually have to be mapped or rendered to physical loudspeakers or earphones to be heard (see Section V.C). To a certain degree, if the listening position is fixed, the use of objects for artificial or panned sources could be argued as merely moving a panning or rendering operation from the mixing studio to the consumer. This, however, ignores that transmitting these sources as objects allows them to be interactively muted or controlled by the listener, as will be discussed below.

The MPEG-H 3D Audio standard offers the capability to support up to 128 channels and 128 objects, mapped to a maximum of 64 loudspeakers. However, it is not practical to provide a decoder capable of simultaneously decoding a bit stream of this complexity in a consumer TV receiver. Thus, profiles have been established to limit the decoder complexity to practical values. The MPEG-H TV Audio System supports the MPEG-H 3D Audio Low Complexity Profile at Level 3, as shown in Table I.

In many implementations such as cinema sound systems, a “bed” of fixed loudspeaker channels in the horizontal plane is combined with overhead objects for production efficiency. The MPEG-H System supports this combination as well.

3) Scene-based Immersive Sound: The philosophy of Scene-based audio is to represent a localized pressure field $p(x, y, z, t)$ as accurately as possible. To do this using Higher Order Ambisonics (HOA), the pressure field is represented as a solution to the Wave equation [7] using Spherical Harmonic basis functions:

$$p(r, \theta, \phi, \omega, t) = \left[\sum_{n=0}^{\infty} j_n\left(\frac{\omega r}{c}\right) \sum_{m=-n}^{n} a_n^m(\omega, t) Y_n^m(\theta, \phi) \right] e^{i\omega t} \tag{1}$$

where, c is the speed of sound, ω is the temporal angular frequency of the wave, $j_n(\cdot)$ is the Spherical Bessel function of degree n and $Y_n^m(\theta, \phi)$ are the Spherical Harmonic functions of order n and degree m for the azimuth ϕ and elevation θ. This can be viewed as a decomposition of the spatial components of the pressure field on the orthonormal basis functions $Y_n^m(\omega, t)$. The decomposed coefficients $a_n^m(\omega, t)$ completely describe the soundfield and are known as the Spherical Harmonic coefficients, HOA coefficients, HOA signals, or just as the “coefficient signals”. For practical purposes, the infinite sum in equation (1) is truncated to n=N, resulting in $(N + 1)^2$ coefficient signals.

At the outset, a few advantages of representing the sound field in this manner should be pointed out:

To rotate the sound field, one needs to just multiply the coefficients, $a_n^m(\omega, t)$, by an appropriate rotation matrix. This makes the format highly conducive to applications such as immersive playback over headphones where tracking head-rotations and movements of the listener is either a necessity (for Virtual Reality type experiences) or at the least makes the binaural listening experience more compelling [8], [9].

It is easy to record a live sound field [10] as these coefficients using a number of off-the-shelf microphones including the Soundfield microphone and the well-known Eigenmike from mhAcoustics. Once in this format, a plethora of new tools become available to a mixer. These include the ability to spatially attenuate sounds emanating from specific regions of the 3D space.

It is also possible to support the offline cinematic or TV episodic workflow where a sound field is created from audio “stems”. This is done by modelling the stems as plane or spherical waves emanating from a certain position in 3D space. They can be further “mixed” to add effects such as “width” and “diffusion”.

Since the $a_n^m(\omega, t)$ representation is oblivious to loudspeaker positions, a renderer [10]–[13] is required to convert the coefficients into loudspeaker feeds. Renderers of this type usually take into account the number and positions of loudspeakers that are available and produce an optimum rendering for that environment. Advanced renderers can also account for local acoustical conditions such as room reverberation. This method of not “tying” the audio to loudspeaker positions, allows the format to be adaptable to practically any loudspeaker layout. It also maintains the spatial resolution required to acoustically focus into spatial regions – allowing the consumer to interact with the sound field – in a way that is not possible with traditional channels-based audio. This approach cannot be compared to “up-mixing” or “downmixing”, with channels-based audio, since those processes are essentially trying to maintain the same audio experience with fewer or larger number of speakers. For example, an up-mixed 7.1.4 signal cannot be expected to provide additional “acoustic” information when played back over 22.2 loudspeakers.

TABLE I
LEVELS FOR THE LOW COMPLEXITY PROFILE OF MPEG-H 3D AUDIO

Profile Level	1	2	3	4	5
Maximum Sampling Rate, kHz	48	48	48	48	96
Maximum core codec channels in bit stream	10	18	32	56	56
Maximum simultaneously decoded core codec channels	5	9	16	28	28
Maximum Loudspeaker outputs	2	8	12	24	24
Example loudspeaker configuration	2.0	7.1	7.1+4H	22.2	22.2
Maximum Decoded Objects	5	9	16	28	28

B. Interactivity/Personalization

The use of audio objects, usually in combination with channel or scene-based audio, enables the viewer to interact with the content to create a personalized listening experience. The MPEG-H system describes a TV program as a graph of objects and related metadata (see Section IV.A) that describe preset sound mixes offered by the broadcaster and set limits on the viewer’s control of the audio.

In the most elaborate case, the system supports sending a program as 16 independent objects which a viewer could change in gain from “muted” to +12 dB and position in three dimensions independently. However, this is likely too complex for the viewer to control or the broadcaster to produce except in special cases.

The simplest case is perhaps the most desired and powerful: where the dialogue or commentary for a program is sent as an object. This allows the viewer to adjust the relative volume or “presence” of the dialogue relative to the rest of the audio elements in the program.

Although the broadcaster will attempt to mix the sound for a good compromise between dialog, natural sound, music, and sound effects, viewer preferences may vary, particularly as the sound mix becomes more complex, such as in sports events or action dramas.

The distribution of viewer preferences for dialogue level in sports, as studied in [14], have been observed as bimodal. This may be due to a desire to experience the event more as a member of the audience and thus a desire for little or no commentary, or a need for more intelligibility due to native dialect or presbycusis.

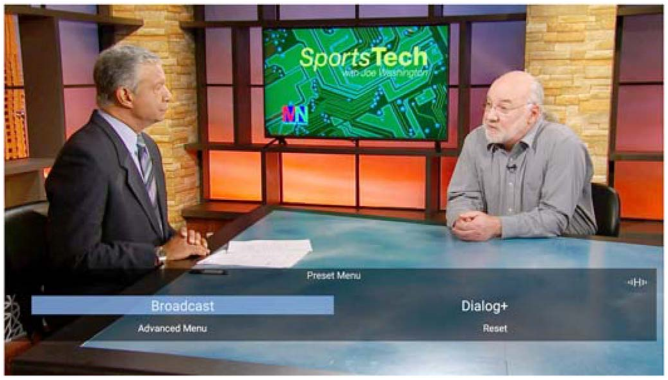


Fig. 3. Typical preset user interface menu for MPEG-H interactivity.

This simple case can be extended to offering two or more dialogue objects for different languages or commentary oriented to each of the teams in a sporting event.

Dialogue is a normal element of most programming, but object-based interactivity also creates the possibility of broadcasting programing with auxiliary or optional objects that could not be included in a broadcast mix in the past. For example, in tests of the MPEG-H system, additional objects have been used to carry radio conversations of teams at auto races.

A feature of the ATSC 3.0 audio requirements is the ability to transmit some objects over Internet channels and combine them in lip-sync with other audio elements in the main over-the-air broadcast. This feature could be used, for example, to support dialogue or voice-over in less popular languages where it is desirable to avoid consuming bandwidth in the broadcast payload. Related to this concept is the need to support video descriptive services, which provides an audio description of the important visual content of a program so it can be enjoyed by visually impaired users. In the MPEG-H system, this “VDS” service is simply implemented as another audio object sent over the air or on broadband networks.

Television is an activity that may involve passive viewing or intense engagement depending on the content, viewer, and environment. Fine and unlimited control of the audio elements in a program is possible and may be desired by enthusiast viewers. However, casual viewers, particularly during the initial deployment of interactive audio, will likely benefit from simple interactivity that is limited to a single button-push on their remote control.

The MPEG-H system has been designed with presets to accomplish this. A broadcaster 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 presented on a simple menu to the user, as shown in Fig. 3.

C. Universal Delivery

The traditional TV broadcast environment uses a well-defined end-to-end solution to deliver audio content to the end user. Accordingly, it has been a good compromise to define a particular target loudness and dynamic range for this specific delivery channel and well-known type of sound reproduction system of the receiving device.

However, new types of delivery platforms and infrastructures have become significant and are constantly evolving. In a multi-platform environment, the same MPEG-H content is delivered through different distribution networks (e.g., broadcast, broadband and mobile networks) and is consumed on a variety of devices (e.g., AVR, TV set, mobile device) in different environments (e.g., silent living room, noisy public transport).

From the consumer’s point of view, the characteristics of the audio content should fit the individual listening condition and preference irrespective of the origin and distribution channel of the content. As a consequence of the wide variability of listening environments, flexible adaptation of the audio content is required to avoid user annoyance in many cases. To be more specific, here are a few examples that illustrate common problems that may result from the scenario described above:

- The user needs to adjust the playback volume when switching between different distribution channels because the loudness is not consistent.
- Consecutive program items usually do not share a common loudness level, e.g., a movie is followed by an annoyingly loud commercial.
- The intelligibility of movie dialog is adversely affected in soft parts due to a noisy listening environment.
- The dynamic range of the audio content is too large for the desired playback level: The level of loud parts of a movie is annoyingly high when soft parts are just loud enough; or soft parts are inaudible when loud parts are at a reasonable level.
- The dynamic range of the audio content is too large for the employed playback device, e.g., low-quality loudspeakers in portable devices.
- The audio signal clips after downmixing the original content to a lower number of playback loudspeakers.

To meet the requirements of multi-platform environments, the MPEG-H system offers a flexible tool set for dynamic range and loudness control, which is based on the MPEG-D Dynamic Range Control (DRC) standard [15]. MPEG-D DRC defines a comprehensive metadata format – including program loudness information and time-varying gain values – and how it is applied. The metadata is typically generated by the content provider and attached to the content. The audio content is delivered unmodified and the metadata can be applied at the receiver if desired. The content provider has full control of the whole process and can ensure that the DRC metadata produces a high-quality result in all scenarios.

For the integration into the MPEG-H system, additional requirements for interactive and immersive audio have been taken into account and corresponding extensions have been added. The dynamic range and loudness control tools of the system provide automatic loudness normalization to a desired target level regardless of the given content format, the loudspeaker configuration for playback, or the user’s selection of a specific preset of the delivered next-generation audio content. It allows reversible adaptation of the dynamic range of the audio as appropriate for the content type, listening environment, and capabilities of the receiving device, as well as user preferences.

III. Core Codec of the MPEG-H TV Audio System

The MPEG-H 3D Audio standard builds upon the previous generations of MPEG audio codecs such as Advanced Audio Coding (AAC), High Efficiency Advanced Audio Coding (HE-AAC) and Extended High Efficiency Advanced Audio Coding (xHE-AAC) of Unified Speech and Audio Coding (USAC). The MPEG-H TV Audio System uses a powerful subset of efficient audio coding tools from the Low Complexity (LC) profile of the MPEG-H 3D Audio standard.

Like its predecessor, USAC, MPEG-H includes a perceptual audio transform coder similar to AAC [16] and a dedicated speech coder, both of which can be alternatively used dependent on the type of signal content to be coded. Also, MPEG-H includes enhanced noise filling as well as audio bandwidth extension functionality. In combination with the multi-channel joint coding tool, said techniques extend the codec's operational range at satisfactory perceptual quality towards lower bit rates.

Table II provides a list of the most important coding modules and tools that are contained in USAC, MPEG-H Audio Low Complexity Profile and the MPEG-H Audio High Profile, respectively. The MPEG-H Audio High Profile incorporates all tools inherited from USAC plus all the newly introduced tools of MPEG-H. The MPEG-H Audio Low Complexity Profile contains a subset of these, which – by design – enable decoding at low computational complexity. Specifically, with respect to USAC, Intelligent Gap Filling (IGF), an improved Linear Predictive Domain (LPD) coding mode, predictors for Frequency Domain (FD) coding mode and Transform Coded Excitation (TCX), and a Multichannel Coding Tool (MCT) have been added to the core codec. As shown in Fig. 4, newly introduced functionality for MPEG-H with respect to USAC is marked in dashed boxes.

This section describes the coding tools that have been newly introduced to MPEG-H Audio with respect to USAC and form the basis of the MPEG-H TV Audio System's audio core codec. The descriptions mainly focus on the related decoder functionality, since MPEG standards are foremost strict definitions of bit stream formats and descriptions of the necessary decoder functionality to arrive at a compliant audio output.

A. Frequency Domain Coding Mode, FD

The Frequency Domain Coding Mode (FD mode) of MPEG-H is derived from USAC's xHE-AAC mode and is preferably applied for high and medium bit rate coding general audio content. The FD mode allows representing audio signals with (almost) transparent audio quality at less than 2 bits per audio sample.

B. Linear Predictive Domain Coding Mode, LPD

The Linear Predictive Domain Coding Mode (LPD) is an extended speech coder derived from USAC's LPD mode that has a transform based extension (TCX) which can handle mixed content and music. It is beneficially applied for low bit rate speech and mixed content coding.

TABLE II
MPEG-H Audio Core Tools

Tool / Module	USAC	MPEG-H Low Complexity Profile	MPEG-H High Profile
Block Switching	X	X	X
Window Shapes	X	X	X
Filterbank	X	X	X
Temporal Noise Shaping	X	X	X
Noise Filling	X	X	X
M/S and Complex Stereo Prediction	X	X	X
Quantization	X	X	X
Context Adaptive Arithmetic Coding	X	X	X
Spectral Band Replication	X		X
MPEG Surround 2-1-2	X		X
ACELP	X	X	X
Frequency Domain Noise Shaping	X	X	X
Intelligent Gap Filling		X	X
Improved LPD Coding		X	X
Predictors for FD and TCX		X	X
Discrete Multichannel Coding Tool		X	X
High Resolution Envelope Processing			X

C. Intelligent Gap Filling, IGF

Intelligent Gap Filling (IGF) is a semi-parametric audio coding technique which enhances the performance of transform-based core coders of MPEG-H in FD and in LPD's TCX mode under low bit rate conditions [17], [18]. Through sharing the MDCT transform domain with the transform coder, IGF is computationally much more efficient than other techniques which require their own analysis and synthesis filter bank pairs for a dedicated processing domain, e.g., the QMF (Quadrature Mirror Filter) domain.

Low bit rates will force a coarse quantization of spectral data, which may lead to large zeroed parts, called *spectral gaps*, in the frequency domain signal, primarily in the high frequency (HF) region. IGF fills these spectral gaps in a synthetic semi-parametric fashion using the de-quantized spectral data from the core coder's low frequency (LF) region together with parametric side information extracted on the encoder side.

Although IGF is not limited to bandwidth extension (BWE) functionality, it can operate in a similar manner as a traditional audio bandwidth extension. The IGF decoder reconstructs missing high frequency (HF) content within the IGF range by copying up low frequency (LF) content to HF content, as depicted in Fig. 5. IGF offers several alternative spectral segmentation possibilities within the copy-up procedure to enable a selection of best-matching LF content.

Nevertheless, a transform coder using IGF is still capable of full-band coding, if required, since selected parts of

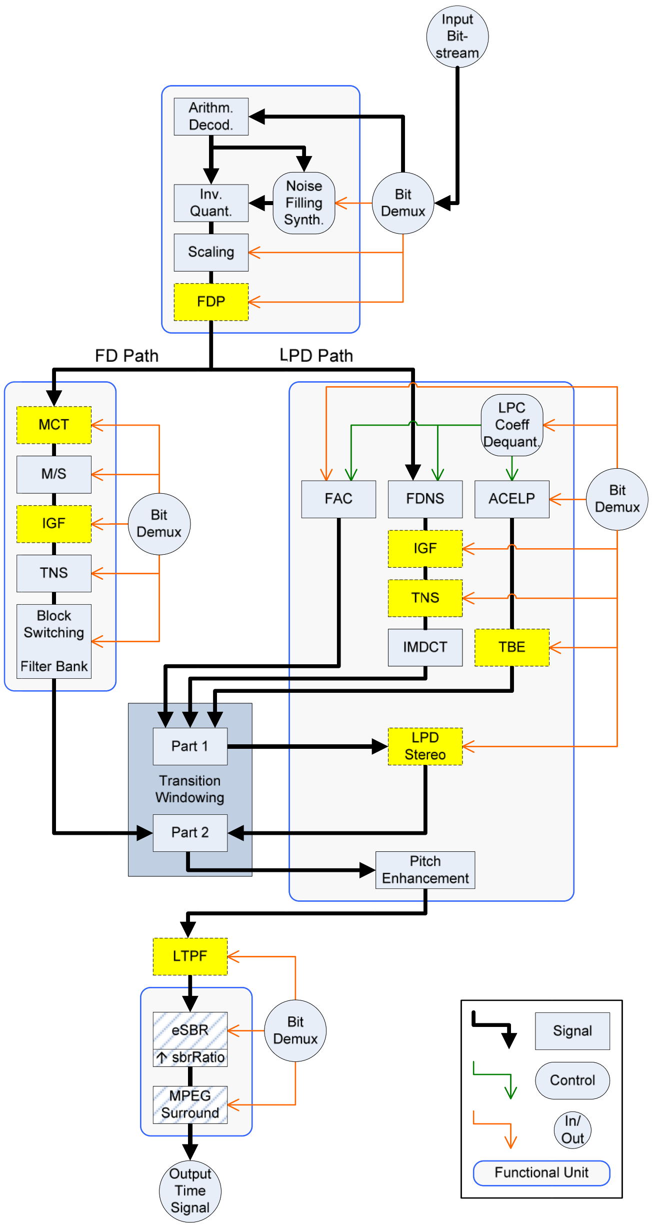


Fig. 4. Block diagram of the MPEG-H audio core decoder. The blocks “eSBR” and “MPEG Surround” are not applicable in the MPEG-H Low Complexity Profile and thus the MPEG-H TV Audio System.

the quantized MDCT signal, the *remaining waveforms*, which may be, e.g., prominent tonal portions of the audio signal, are still encoded with the MDCT transform core coder. These *remaining waveforms* will be decoded in a waveform preserving manner and mixed with the IGF-generated semi-parametric HF content.

Fig. 6 shows the basic signal flow of IGF in the decoder: De-quantized MDCT values are passed together with IGF side-information to the IGF decoder. IGF reconstructed MDCT values and all de-quantized MDCT values are added. The obtained full-band signal is temporally shaped through Temporal Noise Shaping (TNS) and Temporal Tile Shaping (TTS), and finally transformed to time domain PCM audio via inverse MDCT.

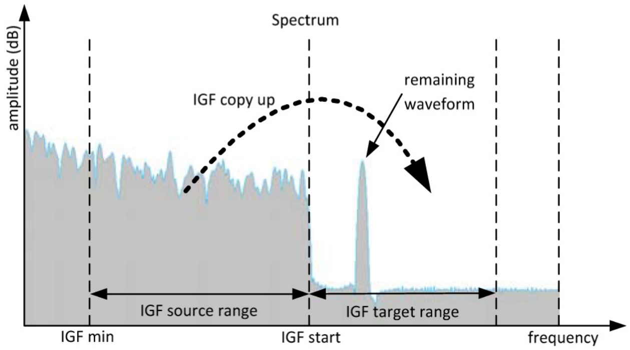


Fig. 5. IGF decoder: copy up from IGF source range to IGF target range.

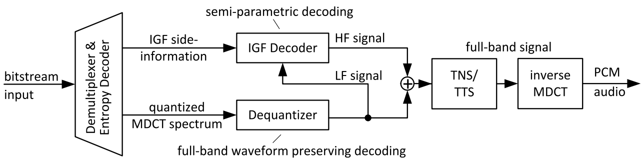


Fig. 6. Signal flow of an IGF decoder.

TTS [19] is an extension to TNS and is functional to control the temporal envelope of IGF-generated HF content to prevent pre- and post-echoes.

In some cases there might be a mismatch between LF and HF content, for example if the LF band has tonal properties whereas the HF band is noisy. With IGF whitening the IGF encoder calculates and transmits parameters to fix this sort of mismatches on the decoder side through flattening the spectral envelope.

IGF also supports joint coding in channel pairs to mitigate spatial unmasking effects.

To save bits, for very low bit rates the IGF low resolution mode is an option to encode the IGF envelope in a coarser scheme. The default mode is IGF high resolution.

D. Long-Term Post-Filter, LTPF

The Long-Term Post-Filter (LTPF) is a prediction-based decoder postprocessor tool that improves perceptual quality for tonal signals at low bit rates.

In transform-coding (especially at low bit rates), tonal components tend to suffer from roughness (or so-called “warbling”) due to unintended temporal modulation caused by coarse quantization in the MDCT domain. The LTPF improves coding gain and mitigates these artifacts.

The LTPF is activated by the encoder for FD- and TCX-mode based on the prediction gain. In this case, the pitch lag and prediction gain are quantized and transmitted using 11 bits per frame.

In the decoder, the LTPF is realized as a pitch adaptive filter, whose filter coefficients are derived from the pitch lag and gain extracted from the bit stream. The filter processes the time-domain output subsequent to the FD- or LPD-core decoder individually for each channel.

E. Multichannel Coding Tool, MCT

The Multichannel Coding Tool (MCT) is specifically designed for a more flexible and efficient joint coding of

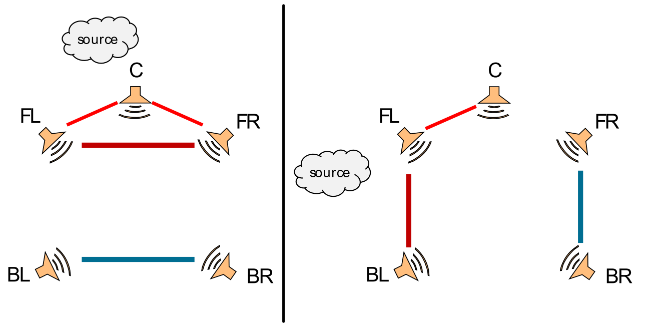


Fig. 7. Placement of stereo coding pairs for different phantom source locations.

multichannel signals in the FD coding path. MCT allows flexible and signal-adaptive placement and cascading of jointly-coded stereo pairs, which significantly improves coding of signals with time-variant channel correlations, e.g., left-right vs. front-back joint coding as depicted in Fig. 7, and signals with high similarities between more than two channels. This advantage can be especially exploited with immersive audio productions. However, subjective listening tests have shown significant improvements of up to 10 MUSHRA points for 5.1 at 144 kbps compared to fixed left/right coding pairs [20].

Especially in terms of immersive audio, pre-defined joint-channel coding topologies cannot cover all possible inter-channel relations as they do not scale linearly with the number of channels. Hence, a more flexible solution is required and provided as part of the MPEG-H core codec. The encoder measures the correlation between all MCT-processed channels. Pairs of two channels with the highest correlation are selected and processed by a stereo operation. The resulting ordered list of channel pairs, the so-called stereo coding tree, highly depends on signal characteristics and is calculated on a frame-by-frame basis. This allows a joint optimization of the stereo coding tree, stereo coding parameters and quantization of spectral information to improve perceptual quality.

In the decoder, the inverse stereo operation is performed consecutively for all transmitted channel pairs.

The channel pairs and corresponding stereo processing side information are transmitted in the bit stream using efficient signaling mechanisms. Accounting for its time-variant property, the stereo coding tree can either be signaled explicitly by transmission of jointly coded stereo pairs or by means of a bit indicating the use of the previous stereo coding tree. As well as supporting full-band, semi-full-band, and band-wise resolution, the pair-wise stereo coefficients are differentially coded in either the frequency- or time-direction, heavily influencing the coding efficiency of the MCT's side information.

For low bit rate operation modes, spectral parts of the "side" or "residual" signal that have been quantized to zero by coarse quantization can be reconstructed by Stereo Filling [20], [21] improving the overall perceptual impression.

F. LPD Stereo

LPD stereo is a joint stereo coding tool dedicated to the Linear Predictive Domain (LPD) core of MPEG-H. It is an advanced Mid-Side stereo (M/S) using prediction or parametric modelling of the Side signal. It can be seen as the counterpart of the joint stereo coding of the Frequency Domain (FD) core. LPD stereo is designed to work in both ACELP and TCX modes. Since ACELP is performed in the time domain, LPD stereo adds a DFT analysis and synthesis stage at the front-end and the back-end of the LPD mono core coder. The stereo tool was specifically designed for improving the processing of stereo speech at low bit rates less than or equal to 32kbps.

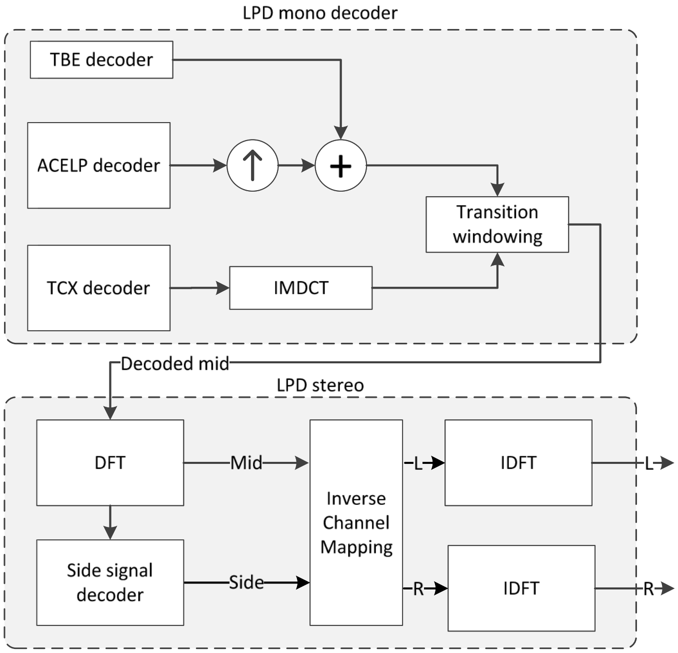


Fig. 8. High level LPD stereo decoder block diagram.

LPD stereo is a joint M/S stereo coding, where the Mid-channel is coded by the mono LPD core coder and the Side-channel coded in the DFT domain. The principle of the decoder is depicted in Fig. 8.

The decoded Mid-channel is output from the LPD mono decoder and then processed by the LPD stereo module. The stereo decoding is done in the DFT domain where the Left and Right channels are derived from the decoded Mid-channel, the decoded stereo parameters and the decoded and predicted Side-channel. The reconstructed Left and Right channels are transformed back in the time domain before being eventually combined with the output channels from the FD core in case of transition from one core to other one. The FD core uses then its own built-in joint stereo tools operating in the MDCT domain.

LPD stereo is a semi-parametric stereo coding. The Mid-channel is a signal-adaptive active downmix of the phase aligned input channels exploiting the Inter-channel Phase Differences (IPDs). It preserves the mid-energy of the two input channels and avoids typical passive down-mixing artifacts like comb-filtering or signal cancellation. The resulting Side-channel is predicted from the Mid-channel using Inter-channel Level Differences (ILDs). The residual of the prediction can either be vector quantized or simply parametrically represented by its energy per frequency band.

G. Frequency Domain Predictor, FDP

The frequency domain prediction (FDP) is a tool for removing temporal redundancy in harmonic signals in the MDCT domain thus improving the subjective perceptual quality for this class of signals. It can be used on individual channels of the FD mode as well as in the TCX part of the LPD mode. The FDP features a design using largely fixed-point calculations as far as possible to ensure consistent operation across different platforms. Furthermore it has low computational complexity and low coding overhead as it only requires a 1 bit On/Off flag and an 8 bit value for signaling the harmonic spacing of the signal.

The working principle of the FDP in the encoder is to remove redundant harmonic components in a frame, using the quantized MDCT spectra of previous frames. By exploiting the special properties of harmonic signals, the harmonic components can be restored by means of the previous frames and the signaled harmonic spacing information in the decoder.

H. Time-domain Bandwidth Extension, TBE

The Time-domain Bandwidth Extension (TBE) is a parametric bandwidth extension method for enhancing the low bit rate perceptual quality of speech and speech-dominant signals. The basic principle is a harmonic extension of the ACELP excitation signal, a by-product of ACELP encoding and decoding, followed by spectral and temporal adaptation to approximate the original characteristics of an input signal. This approach enables the TBE to preserve a realistic harmonic structure while reconstructing even high temporal fluctuations of a signal – both of which are important key features in human speech.

Fig. 9 depicts the effect of the nonlinear modeling (NL) in the harmonic extension module of the TBE schematically. By applying a nonlinear distortion to the up-sampled ACELP core excitation, the harmonic extension module boosts the fundamental frequency harmonics and spreads them beyond the core coder range while preserving the correct harmonic spacing based on the fundamental frequency F_0. This nonlinear modeled excitation signal provides the basis for both, TBE encoding and decoding.

The TBE in MPEG-H supports three types of non-linear modeling, given by

$$\begin{aligned} exc_{NL,smooth} &= abs(exc_{core}) \\ exc_{NL,harmonic} &= \varepsilon_N \cdot sign(exc_{core}) \cdot (exc_{core})^2 \\ exc_{NL,hybrid} &= H_{LP}(z) \cdot \varepsilon_N \cdot sign(exc_{core}) \cdot (exc_{core})^2 \\ &\quad + H_{HP}(z) \cdot abs(exc_{core}) \end{aligned}$$

where exc_{core} is the ACELP core excitation, ε_N is an energy normalization factor, $H_{LP}(z)$ and $H_{HP}(z)$ correspond to low pass and high pass filter transfer functions and exc_{NL} is the resulting nonlinear modeled excitation. The different harmonic extension options *smooth*, *harmonic* & *hybrid* allow for a specific adaptation of the modeling according to an input signal's characteristics.

As a first encoding step, the original audio input gets split at half of the Nyquist frequency $f_{ny}/2$ and resampled to form a time-domain low band and a time-domain high band signal. The low band signal is processed by the ACELP encoder while the high band signal and the output of the harmonic extension module are fed to the TBE parameter extraction. Subsequently, parameters which are suited for temporally and spectrally adapting the harmonic extension signal to resemble the characteristics of the target high band signal are extracted. Those parameters, i.e., gain values for controlling the overall energy and the temporal envelope, LPC coefficients for modeling the spectral envelope and a factor for adjusting the mixing ratio of the harmonic extension signal and noise, are quantized and written to the bit stream.

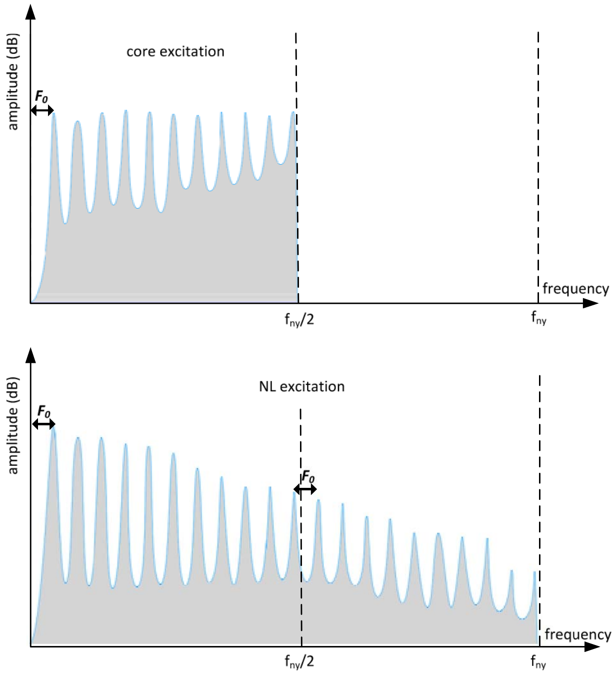


Fig. 9. TBE nonlinear modeling: the up-sampled core excitation spectrum (upper panel) and the bandwidth extended NL excitation spectrum after application of the non-linearity within the harmonic extension module.

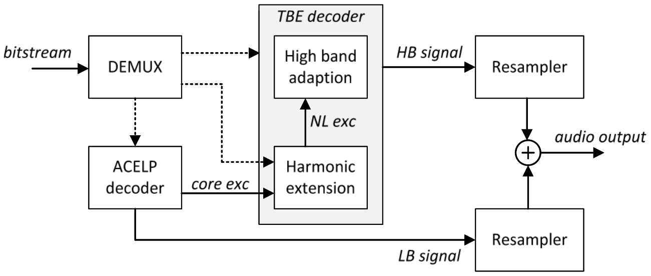


Fig. 10. TBE decoder block diagram.

The basic signal flow and major functional blocks of the TBE decoder are shown in Fig. 10. After dequantization, the parameters from the bit stream are used to adapt the harmonic extension signal to synthesize the TBE output signal. The low band ACELP and the TBE high band decoder signals are each resampled and added subsequently to result in the fullband audio output signal.

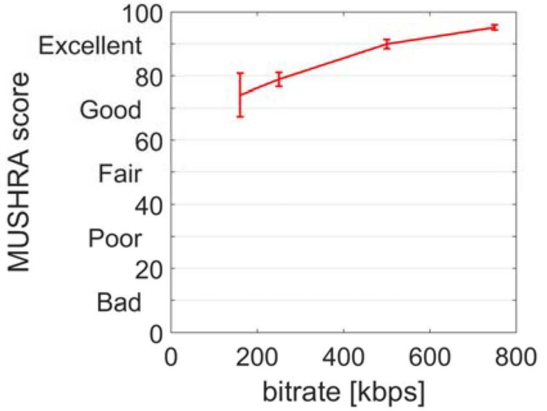


Fig. 11. MUSHRA scores for HOA items coded between 160 kb/s and 768 kb/s. The items were of HOA order 3, 4, and 6.

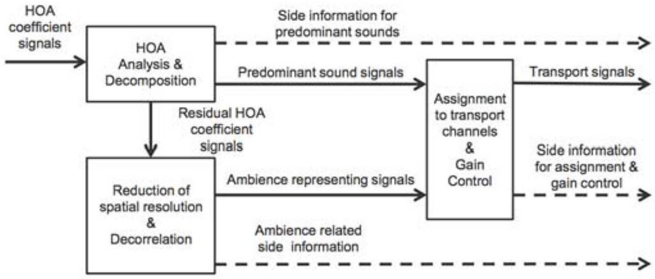

HOA coefficient signals
HOA Analysis & Decomposition
Side information for predominant sounds
Predominant sound signals
Residual HOA coefficient signals
Assignment to transport channels & Gain Control
Transport signals
Side information for assignment & gain control
Reduction of spatial resolution & Decorrelation
Ambience representing signals
Ambience related side information

Fig. 12. HOA Encoding in MPEG-H.

I. HOA Spatial Compression

The MPEG-H standard provides an unprecedented state-of-the-art coding efficiency for scene-based audio content [22]. Broadcast quality transmission (80 MUSHRA points) of HOA content up to sixth order is achieved at bitrates as low as 300 kb/s and transparent quality transmission (90 MUSHRA points) at 500 kb/s independent of the HOA order (see Fig. 11). The following paragraphs describe how the MPEG-H technology is able to achieve this amount of compression.

The first stage of MPEG-H encoding for HOA coefficient signals attempts to decompose the signals into "predominant" (or "foreground") components and "ambient" (or "background") components. This is shown in Fig. 12. The decomposition can be interpreted as a way to decorrelate the HOA signal. The "predominant" or "foreground" components are signals that are perceptually distinguishable from the ambience and are therefore decorrelated from the rest of the sound field. The residual sound field (left-over after the extraction of the predominant sounds) constitute the ambient component. Each predominant component has associated with it a set of data (side-information) that define its spatial characteristics – such as location and width.

One way to conceptualize the decomposition of the predominant signals is using the following equation:

$$H{=}s_i d_i^T, \tag{2}$$

where, s_i and d_i are column vectors depicting the i^{th} predominant component and its associated directional characteristics respectively. The d_i vectors carry the directional characteristics decoupled from the temporal characteristics of the predominant signal. The resulting H matrix is the spatio-temporal representation of the predominant signal. The decomposition can be carried out in a frame-by-frame manner at a resolution of approximately 20 ms per frame. Examples of the d_i vector plotted in 3D space is shown in Fig. 13.

Various techniques [23] can be used to achieve this decomposition such that maximal decorrelation and energy compaction is achieved. The predominant components, using its associated set of information, is used to recombine with the ambient components to recreate the HOA signal at the decoder.

The number of predominant signals ($P(t)$) and ambient signals ($A(t)$) add up to the total number of "Transport" signals ($T = P(t){+}A(t)$). The residual sound field may go through a further decorrelation process to represent the ambient part of the sound field using $A(t)$ number of signals. The numbers $P(t)$ and $A(t)$ can change over time while the total number of transport signals, T, is usually kept constant. This means that channel-assignment information needs to be sent to the decoder. The channel-assignment data, side-information data for the predominant components along with any further information about the ambient components constitute the total side-information – requiring approximately up to 10 kb/s. The value of T is usually a function of the overall target bit rate and is set between 6 and 12. For sixth-order HOA signals (49 HOA coefficients), this represents a dimensionality reduction factor of up to 8. For fourth-order signals the reductions are up to factor 4.

The T transport signals are subsequently processed by the core coding tools described in Section III to achieve further compression. The d_i vectors are quantized using scalar or vector quantization methods – both being allowed in MPEG-H.

At the decoder, the T transport channels are reconstituted from the core decoder before the ambient and predominant components are reconstructed and recombined to reproduce the HOA coefficients. This is shown in Fig. 14. The HOA signal can be subjected to loudness and gain control using Dynamic Range Control (DRC) parameters that are sent within the MPEG-H bit stream. MPEG-H provides an interface to the decoder to allow the local loudspeaker layout to be communicated to the HOA renderer which then produces optimal feeds for that specific layout.

An inherent challenge with deploying Scene-Based-Audio in traditional (non-IP based) Television plants lies in the ability to transport the HOA signals from one part of the plant to the other. Typically, this challenge is due to the use of SD-SDI and HD-SDI routers which limits the number of PCM signals transportable through the TV plant to 8 and 16 respectively. This in turn would limit the HOA order to first or third order respectively. The third order HOA signal would take up the entire embedded 16 HD-SDI channel, meaning that other services such as those for the visually impaired or the ability to have a second mix (e.g., stereo) would be impossible. In order to accommodate the ability to send HOA through both SD-SDI and HD-SDI signals, it is possible to decouple the HOA spatial compression stage of the MPEG-H encoding from the core coding stage [24]. This allows the representation of any

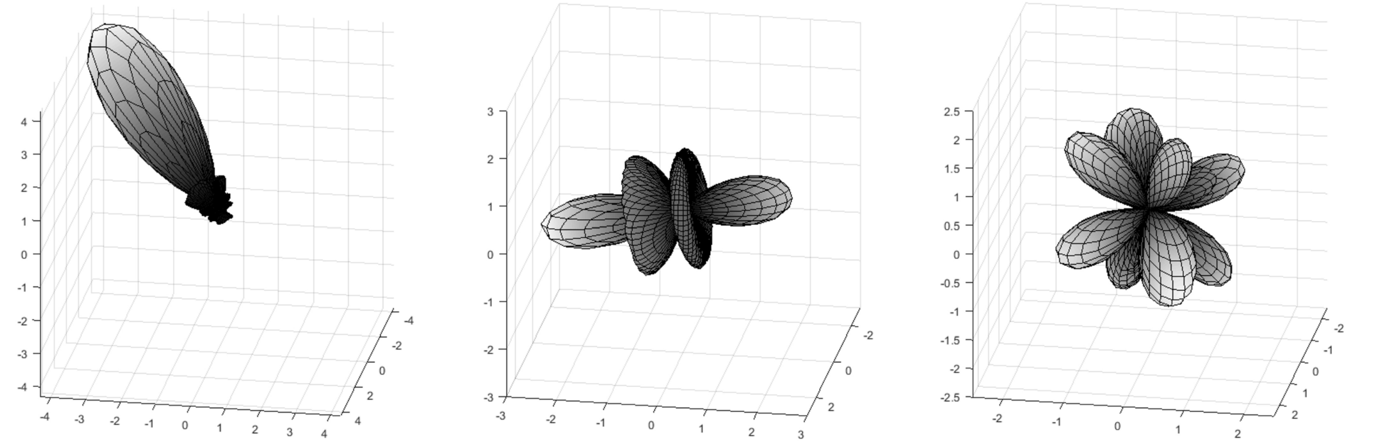


Fig. 13. Spatial plots of various d_i vectors.

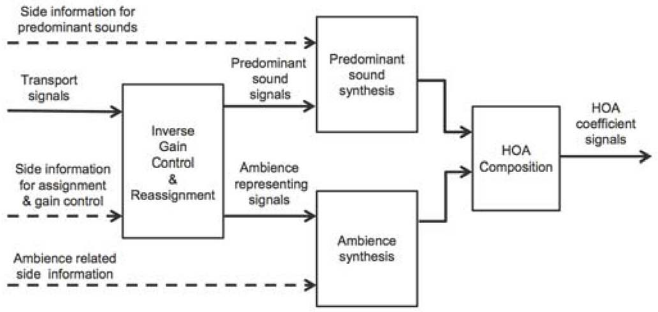

Side information for predominant sounds
Transport signals
Side information for assignment & gain control
Ambience related side information
Inverse Gain Control & Reassignment
Predominant sound signals
Ambience representing signals
Predominant sound synthesis
Ambience synthesis
HOA Composition
HOA coefficient signals

Fig. 14. HOA Decoding in MPEG-H.

orders of HOA content using as low as seven PCM channels (six PCM channels plus one channel for the side-information). This means that a) an HOA signal can be transported through SD-SDI framework and still provide one PCM channel for other services and b) it is possible to deliver fully immersive audio material within existing SD-SDI infrastructures by using one more PCM channel than what would be needed for 5.1.

IV. Audio Scenes and Related Metadata

A. Introduction

Separate audio tracks / sound events and their associated metadata are called "audio elements" in MPEG-H. Audio elements are structured in so-called "audio scenes", using the concepts of groups, switch groups and presets.

Dynamic metadata define object trajectories with high resolution in time. This allows for an accurate reproduction of rapid object movements.

The metadata definition in MPEG-H contains all needed information for reproduction and rendering in arbitrary reproduction layouts.

1) Audio Scene Information: MPEG-H supports several use cases for audio interactivity and object-based audio, such as changing the position or gain of single elements or selecting from different audio scene which offers different languages or semantic contents [1], [25].

The metadata contains information about the hierarchical *structure* of elements inside the bit stream, high-level information of audio elements. It also contains restrictive metadata that defines how *interaction* is possible or enabled by the content creator.

a) Structural metadata: The hierarchical structure facilitates the reference to interactivity options and compilation of Audio Scenes, as shown in Fig. 15. The smallest entity to refer to is called a *Group*, which is a concept for defining arrangements of related elements, e.g., for common interactivity and simultaneous rendering. A use case for groups of elements is the definition of channel-based recordings (stems, sub-mixes) as audio elements (e.g., a stereo recording where the two signals should only be manipulated as a pair).

The concept of a *Switch Group* describes a grouping of mutually exclusive elements, i.e., only one of these elements is played back at a time. This is convenient, for example, for switching between tracks with dialogue in different languages or other audio tracks whose semantic content is not sensible to be played back simultaneously.

By the signaling of high level metadata describing the language and the content type (e.g., language tracks or audio description for visually impaired) the MPEG-H system will be able to adjust the playback. It allows the automatic selection of language or the application of dynamic compression and ducking, e.g., to compensate the gains of a voice over and the main language track.

Using the concept of *Presets* the combination of previously mentioned *Groups* and *Switch Groups* can be combined to present different Audio Scenes to the user from the same bit stream, e.g., a sports event or a match between two teams with different stadium atmospheres or commentaries, one in favor of the home team and one in favor of the guest team.

The combination of *Presets* and high level metadata is able to select an individualized audio scene adjusted to the needs of the viewer by enhancing the dialog volume or selecting a different language track.

b) Interactivity control metadata: 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.

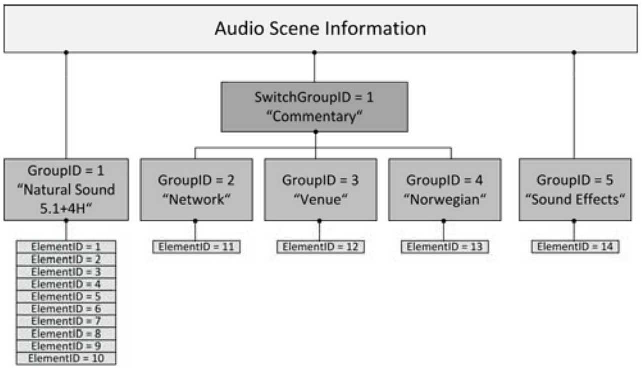


Fig. 15. An audio scene consisting of multiple groups and one switch group.

Gain Interactivity – The overall loudness of the current audio scene will be preserved but the prominence of the referred signal will be increased.

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 fields.

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

B. Dynamic Object Metadata

The merits of object-based representation of a sound scene have been embraced by sound producers, e.g., to convey sound effects like the fly-over of a plane or spaceship. Audio objects are signals that are to be reproduced so as to originate from a specific target location that is specified by associated side information. In contrast to channel signals, the actual placement of audio objects can vary over time and is not necessarily pre-defined during the sound production process but by rendering it to the target loudspeaker setup at the time of reproduction [26].

Each object-based audio track has associated dynamic object metadata that describes the temporal change of the object properties which consist of the following types [1], [27].

Position and Gain - An object's position is given in spherical coordinates (azimuth, elevation, distance, see Fig. 16). The gain of an object, which should be applied to the object-based audio track during rendering, is described by a linear gain value.

Spread – The spread of an object is defined by three parameters, namely width, height and depth. These parameters allow spreading the energy of objects over multiple loudspeakers, thus creating the perception of audio sources with increased extent.

Divergence – If a divergence parameter larger than 0 is specified, two additional objects will be rendered in addition to the original object, such that a phantom object is created in the position of the original object. The value of the divergence parameter controls the gain balance between the three object instances.

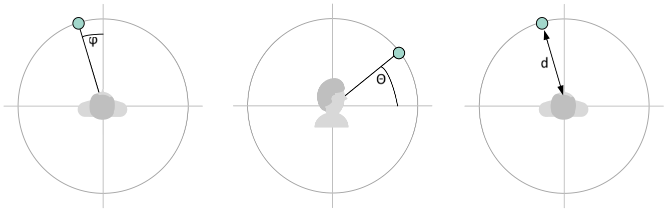


Fig. 16. Spherical coordinates azimuth φ, elevation Θ and distance d.

Diffuseness – The diffuseness parameter describes the diffuseness of an object as a value between 0.0 (not diffuse) and 1.0 (totally diffuse).

Closest Speaker Flag – The closest speaker flag will inform a renderer to send the object's audio to the nearest loudspeaker, rather than generating a phantom image between loudspeakers.

Sector Exclusion – The sector exclusion parameter indicates which loudspeakers/room zones the object should not be rendered through.

isScreenRelative – This parameter indicates whether the object is screen-relative (flag is equal to 1) or not (flag is equal to 0).

Priority – The priority parameter describes the importance of an object (0 to 7). It allows a renderer to discard an object below a certain level of importance if necessary.

Because of the dynamic change of the object characteristics, this metadata would ideally be repeated at a high rate within the bit stream. As this would result in a relatively high data rate if a high number of objects are present, an efficient data compression method for the dynamic object metadata is utilized. For random-access support, a full transmission of the complete set of dynamic object metadata happens on a regular basis, i.e., intra-coded metadata. In-between full transmissions, only differential metadata is transmitted [25].

C. Relationship to ADM-BW64

One approach for a general description model for audio-related metadata is the so-called Audio Definition Model (ADM), which is an open common metadata model. Its current version is specified in the ITU (ITU-R BS.2076 [28]), but an older version (with a reduced set of metadata) has been published by the EBU in EBU Tech 3364 [29] and as part of the EBU Core metadata set in EBU Tech 3293 [30]. The primary specification language of the ADM is XML. It can describe characteristics of objects, channels and HOA. The ADM is designed especially for use in RIFF/WAV-based environments. ADM metadata can be embedded in specific RIFF/WAVE chunks. The RIFF/WAVE-based format, which is best suited to carry metadata compliant to the current ADM definition (ITU-R BS.2076) is the so-called BW64 (Broadcast Wave 64) format. It has been specified by the ITU as Recommendation ITU-R BS.2088 [31] and is a successor of the "Broadcast Wave File" (BWF), extending the functionality to be able to contain more data in a more flexible way. A detailed overview of the ADM and the carriage of ADM metadata in BW64 files can be found in [27].

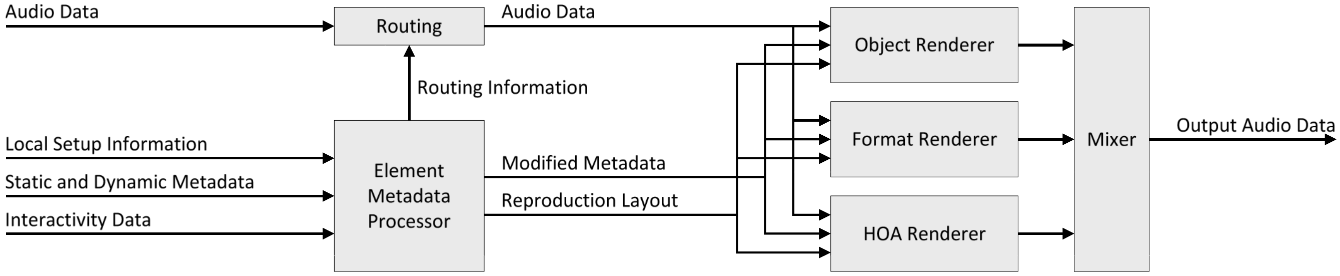


Fig. 17. Block diagram of the MPEG-H rendering stage.

With the usage of ADM metadata in BW64, immersive and interactive audio can be carried and the combination of ADM and BW64 provides an open interchange format for immersive and personalized audio between content creators and broadcasters. As a RIFF/WAVE-based format, BW64 provides the means to store uncompressed audio data with a high fidelity. It supports high sampling rates and bit-depths as required for long-term archiving. With the support of large file sizes, a high number of objects can be stored, such that no removal or combination of single objects has to happen to reduce the file size. As the BW64 format supports the carriage of any XML metadata in a specific chunk, BW64 files can in addition to ADM metadata also contain supplementary metadata required for archiving purposes. Examples are technical metadata from production, post-production and broadcasting stages.

Both the ADM metadata definition and the MPEG-H 3D Audio metadata scheme have similar concepts to describe the content of an audio scene. Both models have the means to define grouping of audio elements, to define exclusive-or relationships (i.e., switch groups in MPEG-H 3D Audio) and to define different "pre-mixes" within an audio scene (called <audioProgrammes> in the ADM and presets in MPEG-H 3D Audio).

The describable audio characteristics are also comparable. Characteristics such as object position, spread (called <width>, <height> and <depth> in the ADM), gain, diffuseness, divergence, the relation of objects to the screen etc. can be defined in both metadata models with similar means and descriptors.

As described above, the main concepts of the ADM and the MPEG-H 3D Audio Metadata are similar. There is a very high compatibility between both models. Therefore, content and audio scenes provided by means of ADM-BW64 can be transferred to MPEG-H and vice versa.

V. MPEG-H Audio Rendering

A. *The MPEG-H Rendering Stage*

MPEG-H offers the possibility for rendering of channels-based content, object-based content and Higher Order Ambisonics (HOA) as a scene-based audio representation. As shown in Fig. 17, channel-based signals are mapped to the target reproduction loudspeaker layout using a "format converter" module. The format converter generates high-quality downmixes to convert the decoded channel signals to numerous output formats, i.e., for playback on different loudspeaker layouts (including non-ideal loudspeaker placement). 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. The object renderer applies Vector Base Amplitude Panning and provides an automatic triangulation algorithm of the 3D surface surrounding the listener for arbitrary target configurations. HOA content is rendered to the target reproduction loudspeaker layout using the associated HOA metadata by a HOA renderer that uses simple matrix operations for manipulation and rendering.

The MPEG-H rendering stage consists of a pre-processing stage followed by the individual renderer modules that render the different audio representations to the target layout. For the pre-processing, a so-called "element metadata preprocessor" module is defined.

The "element metadata pre-processor" module prepares the audio elements for rendering and play-out. It handles user interaction (e.g., choice of a preset) and evaluates the logic implied by the definition of switch groups for mutually exclusive content.

B. *Format Converter*

The Format Converter renders channel-based audio content transmitted in a specific multi-channel format to the target format defined by the actual reproduction loudspeaker layout.

To produce high output signal quality, the format converter in the MPEG-H system provides the following features:

- Automatic generation of optimized downmix matrices, taking into account non-standard loudspeaker positions.
- Support for optionally transmitted downmix matrices to preserve the artistic intent of a producer or broadcaster.
- Application of equalizer filters for timbre preservation.
- Advanced active downmix algorithm to avoid downmixing artifacts.

In addition to the active downmix, passive (i.e., not signal-adaptive) downmix processing can be chosen if required, e.g., to emulate legacy downmix behavior.

The Format Converter features multiple mechanisms for setting the actual downmixing gains to cover all potential scenarios: Downmix configurations can be embedded in the bit stream or decoder generated gains can be applied. All options to set the downmix gains are under full control of the broadcaster or content author. They can be described as follows:

First, the MPEG-H system allows for the transmission of predefined downmix specifications for multiple target channel configurations. Especially in broadcast applications, a producer or content provider may want to retain control over the decoder downmix process, e.g., for artistic reasons. The target channel configuration is defined by nominal loudspeaker positions plus optional displacement information. Transmitted downmix matrices are applied by the Format Converter if the target channel configuration matches the layout assigned to the transmitted matrices. The Format Converter uses a matching scheme to determine if any of the predefined downmix specifications (identified by different "downmixIds") fits to the output setup and selects the most appropriate one. Downmix matrices are transmitted using an efficient compression scheme exploiting symmetries of the channel configurations as well as of the downmix gain matrix. In addition to the downmixing gains, also parametric equalizers can be specified in the bit streams, which are applied to the input signals in the downmix processing.

Second, since any unpredictable combination of transmitted format and target format may occur, the format converter features an algorithm for the automatic generation of an optimized downmix matrix for the given combination of input and output formats: It applies an algorithm that selects for each input loudspeaker the most appropriate mapping rule from a list of rules that has been designed to incorporate psychoacoustic considerations. It takes into account the actual loudspeaker positions, compensating for deviations from nominal layout geometries. Each rule describes the mapping from one input channel to one or several output loudspeaker channels. The optimal mapping for each input channel is selected depending on the list of output loudspeakers that are available in the desired output format. Each mapping defines downmix gains for the input channel under consideration as well as potentially also an equalizer that is applied to the input channel under consideration.

Decoder generated downmix gains are always used if no applicable downmix matrix is found in the bit stream for the desired target loudspeaker configuration. Like for the transmitted downmix gains, also for decoder generated downmix matrices "downmixIds" can be assigned to them. The concept of "downmixIds" allows coupling, e.g., renderings to particular loudspeaker configurations with other MPEG-H metadata like loudness or DRC parameters.

Audio signals that are fed into the format converter are referred to as *input signals* in the following. Audio signals that are the result of the format conversion process are referred to as *output signals*.

Within the format converter module, there are two major building blocks, i.e., a rules-based initialization block and the active downmix algorithm. Both are described in the following.

1) Initialization – Determination of Decoder Generated Downmix Gains: The first sub-module derives optimized downmix coefficients mapping the channel configuration of the format converter input to the output loudspeaker layout. It is applied, if no matching downmix configuration is found in the bit stream.

During the initialization, the system iterates through a set of tuned mapping rules for each input channel. Each rule defines the rendering of one input channel to one or more output channels, potentially complemented by an equalizer curve that is to be applied if the particular mapping rule has been selected. The iteration is terminated at the first rule for which the required output channels are available in the reproduction setup, thus selecting the particular mapping rule. Since the mapping rules have been ordered according to the anticipated mapping quality during the definition of the rules, this process results in selection of the highest-quality mapping to the loudspeaker channels that are available in the reproduction setup.

The rules have been designed individually for each potential input channel incorporating expert knowledge, e.g., to avoid excessive use of phantom sources when rendering to the available target loudspeakers. Thus the rules-based generation of downmix coefficient allows for a flexible system that can adapt to different input/output configurations, while at the same time ensuring a high output signal quality by making use of the expert knowledge contained in the mapping rules. Note that the initialization algorithm compensates for non-standard loudspeaker positions of the reproduction setup, aiming at the best reproduction quality even for asymmetric loudspeaker setups.

In general the rules-based derivation of downmix matrices covers the downmix gain generation for usual target loudspeaker configurations, including those with non-standard loudspeaker positions. In order to always provide a best-effort downmix solution even for arbitrary, very unusual loudspeaker configurations, the MPEG-H Format Converter includes an optimized Vector Base Amplitude Panning (VBAP) [32] matrix generator as a "fallback" solution. Similar to the rules-based matrix derivation, this matrix generator has been tuned to avoid undesired phantom sources that would be present in matrices from plain, non-optimized VBAP solutions.

2) Active Downmix Algorithm: Once the downmix coefficients have been derived, they are applied to the input signals in the actual downmix process. MPEG-H uses an advanced active downmix algorithm to avoid downmix artifacts like signal cancellations or comb-filtering that can occur when combining (partially) correlated input signals in a passive downmix, i.e., when linearly combining the input signals, weighted with static gains. Note that high signal correlations between 3D audio signals are quite common in practice since a large portion of 3D content is typically derived from 2D legacy content (or 3D content with smaller loudspeaker setups), e.g., by filling the additional 3D channels with processed copies of the original signals.

The active downmix in the audio decoder of the MPEG-H system adapts to the input signals to avoid the issues outlined above for passive downmix algorithms: It applies a frequency dependent energy-normalization to the downmix gains that preserves the energy of the input signals that have been weighted by the downmix coefficients. The active downmix algorithm is designed such that it leaves uncorrelated input signals untouched, thus eliminating the artifacts that occur in passive downmixes with only minimum signal adjustments. Operating on a STFT representation of the signals with a hop

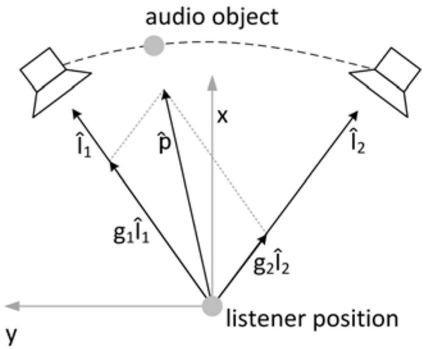


Fig. 18. Audio object described as linear combination of loudspeaker vectors. For simplicity, the 2D VBAP case is shown here with two instead of three base vectors.

size of 256 samples and 50% overlap, the active downmix introduces a short algorithmic delay of only 256 samples, thus helping to allow for short tune-in times of audio and audio-visual devices.

C. Generic Object Rendering in MPEG-H

Generic Object-rendering in MPEG-H consists of two main processing parts:

- Rendering of position
- Rendering of spread in width and height dimension

The individual algorithms are described in detail below.

1) Rendering of Position: The rendering of audio objects on arbitrary trajectories is realized by an object renderer that applies Vector Base Amplitude Panning (VBAP) [32] as shown in Fig. 18.

As input the renderer expects the geometry data of the target rendering setup, one decoded audio stream per transmitted audio object and object metadata associated with the transmitted objects, e.g., time-varying position data and gains.

Two extensions have been included over a generic triangulation and VBAP rendering to improve the perceptual rendering result, especially for arbitrary loudspeaker setups:

Firstly, the object renderer in MPEG-H is complemented by a generic Delaunay triangulation algorithm that provides triangle meshes adapted to the specific geometry of the reproduction loudspeaker setup. It uses an extended Quick Hull algorithm, which was designed to yield left-right and front-back symmetric triangulation meshes, thus avoiding asymmetric rendering of symmetrically placed sound objects. This enables object rendering for arbitrary target configurations.

Secondly, in order to prevent uneven source movements and to avoid the need to restrict object coordinates to the regions supported by the physical loudspeaker setup, imaginary loudspeakers are added to the target setup in regions where the resulting triangle mesh would not cover the full sphere around the listener. This extension ensures the provision of a complete 3D triangle meshes for any setup to the VBAP algorithm. The signal contributions rendered by VBAP to the imaginary loudspeakers are downmixed to the physically existing loudspeakers. The downmixing gains for mapping virtual to available loudspeakers are derived by distributing the virtual loudspeakers' energy equally to the neighboring loudspeakers.

A triplet-wise panning is used for 3D setups and for all other setups, as all setups are extended by imaginary loudspeakers, if necessary. For this, the audio object is applied to a maximum of three loudspeakers. All calculations are performed for each loudspeaker triplet. The triplets are defined by the loudspeaker triangulation as described above.

Metadata is conveyed for every audio object at defined timestamps. To get a smooth transition when metadata changes, the loudspeaker gain factors are interpolated linearly between adjacent timestamps and applied on a per-sample basis.

2) Rendering of Source Extent in Width and Height Dimension: MPEG-H further features a gradual spread parameter that gives the content creator an additional degree of freedom to express artistic intents. It allows spreading the energy of objects over multiple loudspeakers, thus creating the perception of audio sources with increased extent. The spread algorithm in MPEG-H is based on Multiple Direction Amplitude Panning (MDAP) [33].

This method involves the computation of a set of panning gains $\mathrm{g}_{\mathrm{scaled,m}}$ for $\mathrm{M} = 18$ MDAP directions p_{m} around the original object panning direction $\mathrm{p}_0 = \widehat{\mathrm{p}}$.

The determination of the MDAP directions requires the computation of two base vectors, *u* and *v*. These MDAP directions form a specific vector pattern with its centre being the original object direction.

A circular pattern is used to render objects with uniform spread in width and height dimension. An elliptical pattern is used for objects with non-uniform width and height spread. The elliptical form is realized by using the ratio of the values of width spread and height spread to scale the base vector *v*.

D. Enhanced Object Rendering in MPEG-H

In addition to the described generic object rendering algorithms, the MPEG-H 3D Audio Low Complexity Profile also contains the definition of enhanced object rendering features beyond rendering of position and 2D spread. These additional features can be part of future implementations that want to provide a superior user experience by the deployment of additional object-based rendering technologies.

The processing of the enhanced object rendering features is then mainly conducted in the metadata pre-processor module as a pre-processing step to the VBAP- and MDAP-based rendering steps. The processing itself relies on additional enhanced object metadata descriptors as part of the overall MPEG-H 3D Audio metadata scheme.

The following metadata descriptors and corresponding processing steps are defined:

1) Rendering of Object Divergence: The MPEG-H 3D Audio divergence parameter (0.0 ("no divergence") and 1.0 ("full divergence")) indicates the amount an object is split symmetrically into a pair of virtual objects, such that a phantom object is created in the position of the original object, as shown in Fig. 19. The position of the virtual objects can be defined with the use of an additional parameter called

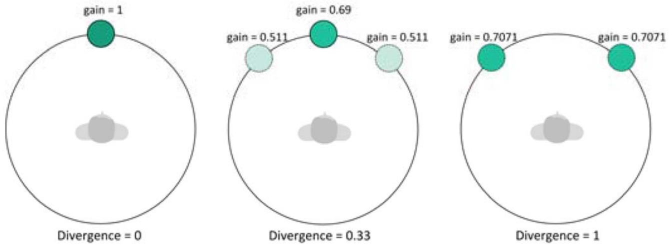


Fig. 19. Relation between the divergence metadata parameter and the object gain in the divergence processing.

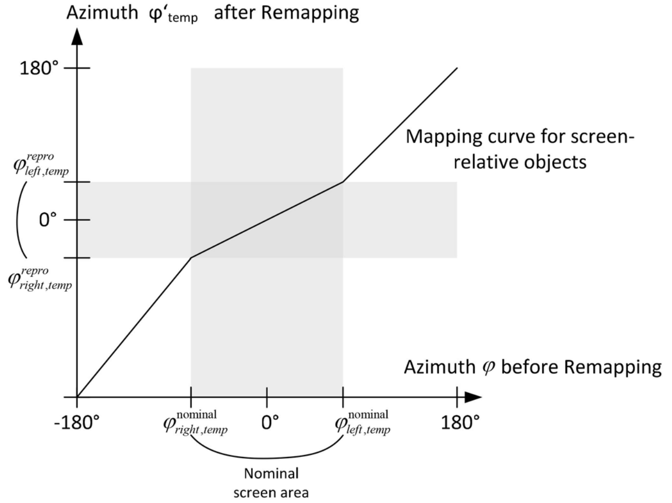


Fig. 20. Exemplary azimuth mapping function for screen-dependent scene scaling/remapping.

"azimuthRange" and the gain of the objects is dependent on the divergence value.

2) Screen-dependent Scene Scaling: The geometric positional data of screen-related objects is mapped to a different range of values by the definition and utilization of a mapping-function (called "screen-related remapping"), as shown in Fig. 20. The remapping changes the geometric positional data as a pre-processing step to the rendering, such that the renderer is agnostic of the remapping and operates unchanged. The screen size of a nominal reference screen (used in the mixing and monitoring process) and local screen size information in the playback room are taken into account for the remapping. If no nominal reference screen size is given, default reference values are used assuming a 4k display and an optimal viewing distance, resulting in a 58° viewing angle.

3) Usage of Exclusion Sectors: In MPEG-H 3D Audio it is possible to define so-called "exclusion sectors". All loudspeakers that are located within the excluded sectors are not used during the rendering of the associated objects. The exclusion of loudspeakers is based on the ideal reproduction loudspeaker position, as the sectors are defined on the encoder side where the real reproduction loudspeaker positions may not be known.

4) Closest Loudspeaker Playout Processing - Snap: It can be defined that an object should not be rendered but directly be played back by the loudspeaker which is closest to the geometric position of the object indicated in the dynamic metadata. A conditioned closest loudspeaker playout can be defined. A magnetic area around an object is then specified by a threshold angle parameter. The object only snaps to a loudspeaker if the object is close enough to a loudspeaker. Therefore, only the speakers within the defined magnetic area are considered.

5) Rendering of Diffuseness: It is possible to define objects with a specific diffuseness (between 0.0 and 1.0). For each object with a diffuseness bigger than zero, two signal versions are created: A "direct sound part" and a "diffuse sound part". The direct sound part is the normal rendered object output. The diffuse sound part is created by replicating the object audio content to the number of total available reproduction speakers (without LFEs) N. LFE speakers are not considered for the reproduction of the diffuse part. Each of these N signals is filtered with a decorrelation filter. The decorrelation filters are defined informatively.

Both the direct sound part and the diffuse sound part for each object are in addition weighted with a gain factor, which is dependent of the diffuseness value of the corresponding object. The direct sound part is sent to the object renderer module, which renders position and spread, the diffuse sound part is directly sent to the mixer that combines the different signal paths.

6) Rendering of Distance and Depth Spread: For distance rendering an informative description is included in MPEG-H 3D Audio. The approach used employs methods of signal processing to render objects with a defined distance that can be nearer or further away than the original loudspeaker distance.

The object radius parameter is mapped to a distance factor, describing the distance relationship of objects and speakers. Dependent on this distance factor, the original object is copied to additional specific positions. The multiple object instances are then adjusted in level with different gain factors, a specific gain ratio implies the perception of "nearness" or "distance."

If an object should be rendered behind the reproduction speakers, the object instances are additionally decorrelated with individual decorrelation filters, e.g., according to the filters defined for diffuseness rendering. This results in a lowering of phase coherence and inter-aural coherence.

Rendering of object depth (depth spread) can also be realized by the means of the described informative distance rendering. Rendering of object depth is also only defined informatively in MPEG-H 3D Audio.

To make use of the depth spread value, which is transmitted in case of non-uniform spread, the corresponding object is rendered at multiple distances:

- Once at the "front" of the depth expansion
- Once at the "back" of the depth expansion
- Once at the original distance

The rendering of width and height spread by the use of 18 additional MDAP directions is then only applied to the object at the original distance.

E. Scene-based Audio Rendering

To listen to the HOA content over loudspeakers, HOA content is rendered into loudspeaker feeds. This process is efficiently carried out within the MPEG-H decoder by multiplying the decoded HOA content with a rendering matrix. This matrix can be generated based on the number of decoded HOA coefficients and the number and locations of the loudspeakers in the reproduction environment.

In contrast to channel-based audio, the upmixing and then downmixing to accommodate loudspeaker configurations that differ from the configuration for which content was intended is not necessary in scene-based audio.

The HOA order N (and therefore the number of $(N+1)^2$ HOA coefficient signals) does not dictate the number of loudspeakers or a certain loudspeaker arrangement. Even HOA content of small orders (e.g., second order = 9 reconstructed HOA coefficient signals) can be rendered to a large loudspeaker array; and reconstructed HOA content with a higher order (e.g., sixth order, 49 HOA coefficient signals) can be rendered to stereo. This all just depends on the rendering matrix, which adapts to the conditions in the listening environment.

Although the MPEG-H decoder is equipped with state-of-the-art HOA rendering capabilities, content creators may want to use a different HOA renderer perhaps for artistic reasons or simply as a means to differentiate. To maintain artistic intent, the same HOA renderer has to be used in consumer devices. This can be accommodated in MPEG-H which allows the transmission of HOA rendering matrices within the bit stream.

By exploiting left/right symmetries within the loudspeaker configurations as well as rotation symmetries between the spherical harmonics, HOA rendering matrices can be efficiently compressed and stored into a dedicated MPEG-H bit stream extension payload.

When rendering the HOA content according to the reproduction environment the MPEG-H decoder will prefer the transmitted rendering matrices over the default rendering process if appropriate.

F. Binaural Rendering in MPEG-H

The MPEG-H 3D Audio standard decoder contains a module for binaural headphone rendering. This binaural rendering module includes two tools for binaural rendering: A frequency-domain binaural rendering is included, as well as a time-domain binaural rendering, which has an additional mode to directly create binaural headphone signals from HOA content.

The frequency-domain binaural renderer shown in Fig. 21 can be used for generating binaural headphone signals for all types of input content (channel-based and/or object-based and/or scene-based). It takes loudspeaker feeds as input signals. The frequency-domain binaural processing is carried out as a decoder process converting the decoded signal into a binaural downmix signal that provides an immersive sound experience when listened to over headphones.

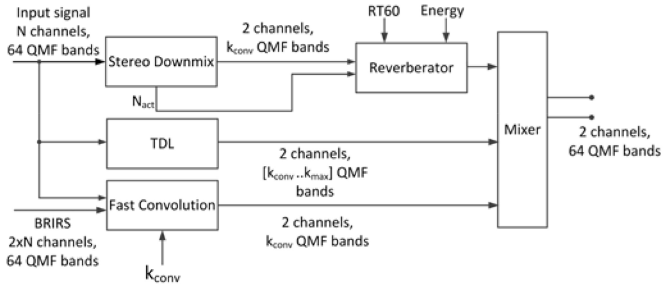


Fig. 21. Frequency-domain binaural rendering.

The processing is conducted in the QMF-domain (Quadrature Mirror Filter domain, see [16] for a detailed explanation). Parameterized binaural room impulse responses (BRIRs) are used to generate the binaural headphone signals. The BRIRs are split into a so-called "D&E" part, containing direct sound and early reflections, and a late reverberation part.

The QMF-domain audio signals are processed with the QMF-domain D&E parts of the BRIRs by means of a convolution.

For efficient processing an FFT-based partitioned fast convolution is implemented on-top of the QMF-domain bands. A partitioned FFT transform is calculated for each QMF band of the audio input signal and each QMF band of the D&E BRIRs and the convolution is conducted in this "pseudo-FFT domain". The FFT size and the partitioning are frequency-dependent and based on the RT20 reverberation time in each band of the used BRIRs. This QMF-domain fast convolution is also referred to as Variable Order Filtering in Frequency domain (VOFF).

A sparse QMF-domain reverberator (SFR) is used to generate 2-channel QMF-domain late reverberation. The reverberator uses a set of frequency-dependent RT60 reverberation times and energy parameters, calculated from the late reverberation parts of the BRIR set to adapt the characteristics of the reverberation. The waveform of the reverberation is based on a stereo downmix of the QMF-domain audio input signal. The late reverberation is adaptively scaled in amplitude dependent on the number of currently active channels.

The convolutional result and the reverberator output are combined by a mixing process that band-wise adds up the two signals forming a combined output signal.

In the upper QMF bands (usually band 48 and higher) a QMF-domain tapped delay line (QTDL) is used instead of convolution and late reverb generation.

The filters used for the convolution, as well as the reverberation times and energies for the reverb generator and the parameters of the QTDL block are provided by an adaptive filter parametrization technique of the BRIRs, which is also defined in the MPEG-H 3D Audio standard.

VI. Dynamic Range and Loudness Control

A. Overview

The dynamic range and loudness control tools of MPEG-H support audio processing in the following main categories:

- Loudness control of single program items and the entire program.
- Dynamic range control of single program items.
- Peak and clipping control.
- Ducking for voice-over applications.

Typically, a combination of these different processing categories is required to achieve a desired adaptation of the audio characteristics. For example, to normalize the audio content to a high target loudness level, dynamic range compression is needed to provide the required headroom and additional peak limiting should be performed to avoid clipping distortions.

The specific features provided by MPEG-H for each of these functional categories mentioned above are outlined in this section. Whenever necessary, all functions work together to produce the most suitable output for a given playback scenario.

B. Loudness Control

The MPEG-H system can automatically provide consistent loudness of the reproduced audio content at the decoder. This is accomplished in two steps: Loudness Normalization aligns the loudness between program items under different playback conditions and Loudness Compensation additionally compensates for loudness changes due to user interaction.

1) Loudness Normalization: MPEG-H supports mandatory loudness information that is included in the metadata of the MPEG-H stream. Various loudness measurement systems (e.g., ITU-R BS.1770-4 [34], EBU R-128 [35], ATSC A/85 [36]) are supported in order to fulfill applicable broadcast regulations and recommendations. It is possible to specify whether a loudness descriptor relates to the loudness of the full program or whether it refers to a specific anchor element of the program such as the dialog or commentary.

The general concept of loudness normalization is illustrated in Fig. 23, where the meaning of the different bars is shown in Fig. 22. The thick horizontal line in Fig. 22 corresponds to the measured loudness of the considered audio item. The light bar represents the loudness range [37] and the top of the dark bar illustrates the peak level [34] of the audio item. It is important to note that the loudness range of an audio item is closely related to its dynamic range, i.e., a large loudness range indicates high dynamic range audio, and consequently, a small loudness range implies a reduced dynamic range of the audio item.

Fig. 23 shows three example items of a broadcast program. The top of Fig. 23 illustrates the case of unprocessed program items. None of the three different program items matches the target loudness level at the decoder, where the commercial and the sports program are too loud and the movie's loudness is too low compared to the desired target level. After loudness normalization has been applied, as illustrated at the bottom of Fig. 23, the playback loudness of all three program items is the same and matches the desired target level at the decoder. Note that the loudness range and the signal peak relative to the program loudness remain unchanged after loudness normalization, since loudness normalization is performed based on a time-invariant normalization gain derived from the difference of the loudness of the original program item and the target loudness level for playback at the decoder.

More advanced loudness related parameters including loudness range, maximum short-term and maximum momentary loudness, production mixing level and room type (see ITU-R BS.1771-1 [38], EBU R-128 [35] [37], ATSC A/85 [36]), as well as signal peak information can be optionally transmitted in the MPEG-H metadata if desired. A detailed overview of supported descriptors is provided in [25].

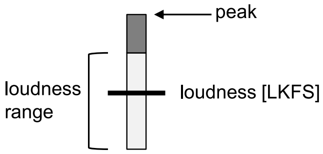


Fig. 22. Legend for loudness and DRC related illustrations. Loudness is measured in LKFS according to [34].

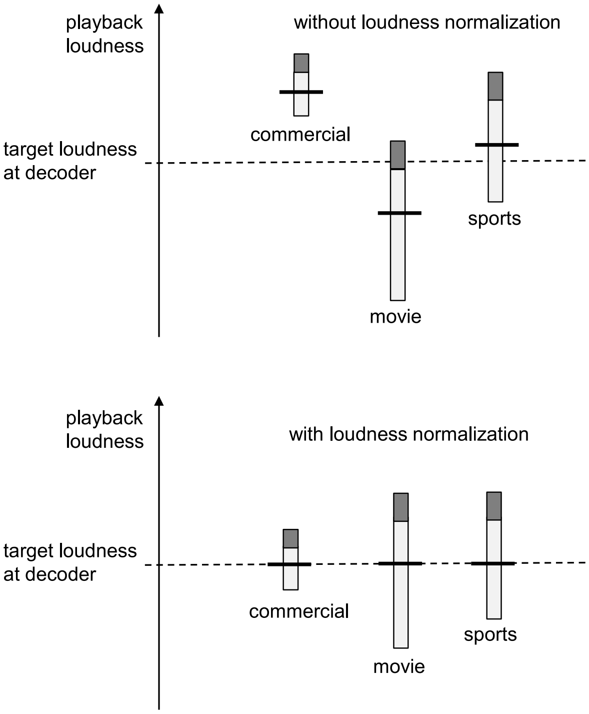

playback loudness
without loudness normalization
commercial
target loudness at decoder
sports
movie
playback loudness
with loudness normalization
target loudness at decoder
commercial
sports
movie

Fig. 23. Example of loudness normalization for three program items having different loudness.

DRC processing or downmixing for format conversion may potentially change the loudness of the original audio content. Loudness information of the program after applying a specific DRC or performing format conversion can be additionally included in the metadata to compensate for any related loudness variation.

In case the MPEG-H stream contains multiple presets of the same program, loudness information can be input for each of them separately at the encoder. This enables immediate and automated loudness control for interactive and personalized audio reproduction. For example, when the user switches between different presets the loudness normalization gain is instantaneously adjusted.

2) Loudness Compensation: MPEG-H allows users to dynamically interact and control the rendering of individual audio elements as described in Section II.A.3). For example, the level of a dialog or commentary object within the program mix can be changed through a corresponding control interface provided to the user in the receiving device. However, when increasing the level of the dialog object, the overall loudness

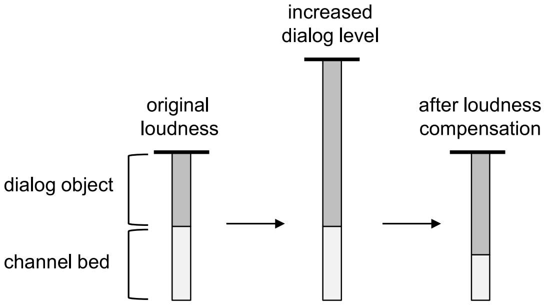


Fig. 24. Illustration of the loudness compensation concept for user interaction with the dialog object gain. The height of each bar corresponds to the loudness portion of the dialog object and the channel bed, respectively.

of the resulting mix will also be increased compared to the original preset. This behavior interferes with the requirement of consistent loudness and preservation of signal headroom. Therefore, MPEG-H includes a tool to compensate for loudness variations due to user interaction with rendering gains. The loudness compensation tool is based on metadata included in the audio stream that provides the measured 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 by the user, which is then applied together with the loudness normalization gain. The loudness compensation concept is illustrated in Fig. 24 for the example of a program consisting of a dialog object and a channel bed.

On the left hand side of Fig. 24, the loudness of the original preset is shown, where the different bars correspond to the loudness portion of the dialog object and the channel bed. In the center, the loudness distribution is shown if the level of the dialog is increased and no loudness compensation is applied. Obviously, increasing the level of a dialog object also results in an increase of the overall loudness of the full program mix. The right-hand side depicts the loudness distribution within the mix after applying the loudness compensation gain to restore the desired decoder target loudness. As can be seen, the relative level of the dialog object to the channel bed in the mix is increased as desired, where this effect is partly achieved by boosting the dialog, but also by attenuating the channel bed signals when applying the loudness compensation gain.

C. Dynamic Range Control

Traditionally, Dynamic Range Control (DRC) is associated with compression, i.e., reduction of the dynamic range by means of a dynamic range compressor. This results in softer segments of an audio item being louder and/or loud segments being softer. A simple dynamic range compressor generates time-varying gain values that are applied to the audio signal to achieve a desired compression effect.

Instead of applying the gain values immediately, MPEG-H employs a reversible approach: The DRC gain values are provided as metadata accompanying the audio content such that they can be applied at the receiver if desired. If no DRC processing is required, the audio remains unchanged and is played back without any modifications. The DRC configuration data is fully controlled at the encoder and consists of static information describing the different DRC configurations included in the DRC metadata and dynamic data representing encoded DRC gain sequences. The temporal resolution of the encoded DRC gains can be chosen to be as low as 1 ms and thus allows for transmitted DRC sequences meeting requirements of professional dynamic range control. The efficient gain coding schemes offered by the DRC tool avoid undesirable high bit rate overhead for the dynamic DRC metadata.

There are various parameters specified in [1] and [15] that can be used to control the DRC processing, where important examples include:

- Compressor characteristics that are used to compute the DRC gain sequences.
- The target level range for which the DRC configuration is optimized. This relates to the appropriate playback target levels commonly used for different reproduction devices and scenarios (e.g., AVR, TV, mobile device).
- The DRC effect types provided by a certain DRC configuration. They serve a wide range of scenarios such as Late Night Listening, Noisy Environment, Dialog Enhancement, or General Compression.
- The frequency band configuration for flexible multiband DRC processing.
- The target loudspeaker configuration: Unique identifiers define whether a DRC configuration is designed for processing the original content format or for playback with a specific loudspeaker setup (e.g., optimized for stereo playback).
- Preset identifiers to indicate that a DRC configuration is assigned to a specific preset of the original audio program.

The configuration and gains of several independent compressors can be included in the DRC metadata to achieve optimized compression effects for various playback scenarios. The DRC system allows specifying separate DRC gain sequences with different compressor characteristics for single channels or groups of channels, individual objects or groups of objects, HOA content or different pre-defined presets of the program. Several DRC configurations can be defined, each consisting of a well-defined set of DRC gain sequences. At the decoder, the appropriate DRC configuration is automatically selected considering the given target loudness level for playback, the desired DRC effect type, the reproduction setup, the selected preset of a program or other user input.

It is common practice to use different compressors for different playback target levels. While for low playback target levels only moderate or even no compression is desired, high target levels imply the need for a considerable amount of signal compression. Typical relations of target level and dynamic range for different receiver types are illustrated in Fig. 25. It shows practical examples for three different target levels, representing typical choices as used for playback over AVRs, TVs, and mobile devices.

Consequently, a specific target level range is defined for each DRC configuration in the MPEG-H bit stream to identify the most suitable one for a given playback system. Fig. 26 illustrates an example of three different sets of DRC configurations that are designed for specific receiver types

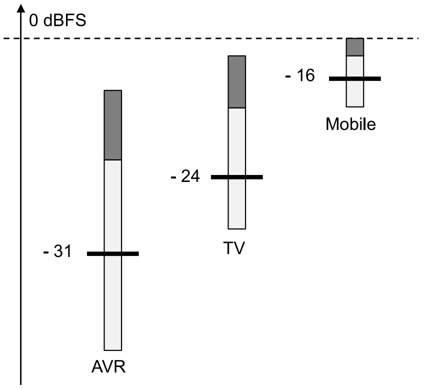


Fig. 25. Typical relations of target level and dynamic range for different receiver types.

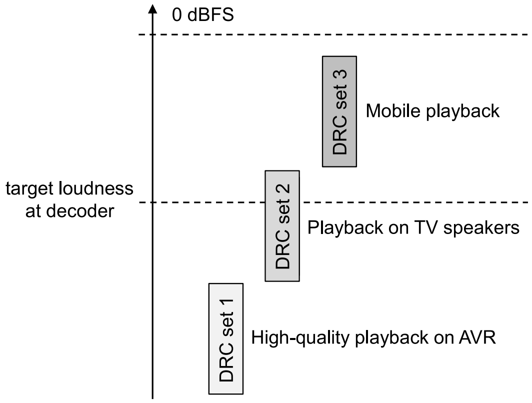


Fig. 26. Illustration of the target level range information included for different DRC configurations (so-called DRC sets). The dashed line indicates the desired target loudness for this example.

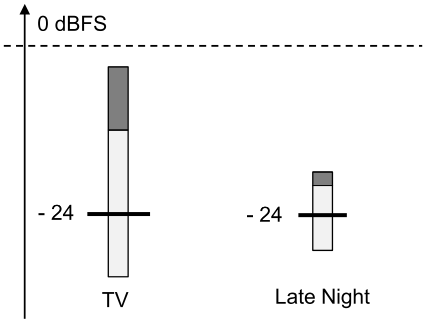


Fig. 27. Illustration of dynamic range characteristics after processing with the default DRC configuration for TV and a DRC configuration with effect type "Late Night".

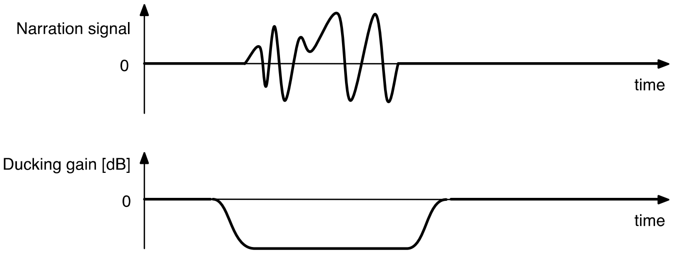


Fig. 28. Narration object signal (top) and corresponding ducking gain sequence (bottom) applied to the main program.

(AVR, TV, mobile device) and that declare corresponding target level ranges.

The DRC effect types supported by MPEG-H cover a wide range of use cases and include: Late Night Listening, Noisy Environment, Dialog Enhancement, and General Compression. Fig. 27 illustrates a comparison of regular playback on TV with watching a movie late at night. For the latter case, loud parts of the movie should be attenuated, e.g., to avoid disturbance of family members or neighbors, while soft parts of the dialog should still be intelligible at low listening levels. This results in a reduced dynamic range compared to the audio signal after processing with the default DRC configuration for a high-quality receiver type.

The user of an application can request DRC effect types via the control interface of the decoder.

D. Peak and Clipping Control

There are several processing steps within the MPEG-H decoding that can potentially lead to clipping of signal peaks. The most important examples for such processing steps are loudness normalization to high target levels; downmixing to a lower number of playback loudspeakers during format conversion of channel-based content; and user interaction with rendering gains, where the level of certain objects or channel groups is increased. Usually, the dynamic range and loudness control tools described above will largely prevent the signal from exceeding the maximum peak level. To avoid annoying audible distortion due to digital clipping that may still occur, the MPEG-H decoder includes a high quality peak limiter at the very end of the decoder processing chain.

E. Ducking

Ducking of audio content is often employed if an audio signal is overlaid over the main audio signal, such as narration, director's commentary, video description, and other related use cases. In common approaches the main audio signal is attenuated when the narration is active. The attenuation and transitions can be precisely controlled and delivered by the dynamic range control tool of MPEG-H. The realization of ducking via encoder generated gains transmitted within the DRC metadata eliminates the look-ahead needed in traditional playback systems.

The ducking gain sequences for the main audio program are coupled to a corresponding audio object or track. The ducking of the main program is automatically performed if, e.g., the user activates video description or selects a pre-defined preset including it.

Fig. 28 illustrates the signal of a narration object and the corresponding ducking gain sequence.

VII. MPEG-H Transport

A. Encoder Output

The output of the MPEG-H encoder typically consists of several data structures or syntax elements: *mpegh3daFrame*

TABLE III
MPEG-H SAMPLE RATES

decoded samples @48kHz	resampling factor	core coder sample rate
1024	1.0	48.000 kHz
1536	1.5	32.000 kHz
2048	2.0	24.000 kHz
3072	3.0	16.000 kHz

Amount of decoded audio samples at 48kHz sample rate with the resampling factors allowed in MPEG-H 3D Audio LC Profile at Level 3. The core coder uses internally always the audio frame size 1024 but on different sampling rates.

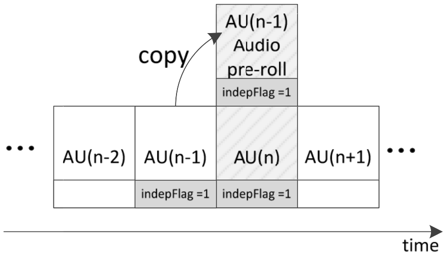


Fig. 29. Construction of an audio pre-roll: Copying and including the AU(*n-1*) to AU(*n*), AU(*n-1*) is stored as an AudioPreRoll-Extension payload in AU(*n*).

carrying the compressed audio bitstream as byte-aligned access units (AUs), *mpegh3daConfig* carrying the corresponding configuration data, and *mae_AudioSceneInfo* carrying the Audio Scene Information. The decoder is unable to properly process an AU unless it first receives the corresponding configuration data.

Note that the number of time domain samples obtained after decoding one AU depends on the core coder sample rate specified in the configuration. See Table III for more details.

AUs may have two properties, which are important on the transport layer:

- *Independency:* An AU with this property is independent from all previous AUs and contains all data needed for the decoder to be able to decode this AU. This property is signaled by setting to 1 the *usacIndependencyFlag* (or *indepFlag* in the following), the very first bit in the *mpegh3daFrame()* syntax element.
- *Audio Pre-Roll:* An AU carrying the *AudioPreRoll* extension contains the information of the previous AUs to allow the decoder compensation of its startup delay, the so-called pre-roll, e.g., for gapless stream switching.

Fig. 29 shows how an encoder might insert the *Audio Pre-Roll* into an AU. An AU containing the *Audio Pre-Roll* with the *indepFlag* set, is called an Immediate Playout Frame (IPF), as a decoder could start decoding this AU to obtain valid samples from the very beginning. Along with a configuration an IPF can be used to generate a Random Access Point (RAP).

B. MPEG-H Audio Stream Format

The MPEG-H Audio Stream (MHAS) format is a self-contained, flexible, and extensible byte stream format to carry MPEG-H Audio data [1] using a packetized approach. It uses different MHAS packets for embedding coded audio data (AUs), configuration data, and additional metadata or control data. This allows for easy access to configuration or other metadata on the MHAS stream level, without the need to parse the complete bit stream. Furthermore, MHAS allows for sample-accurate configuration changes using audio truncations on decoded AUs.

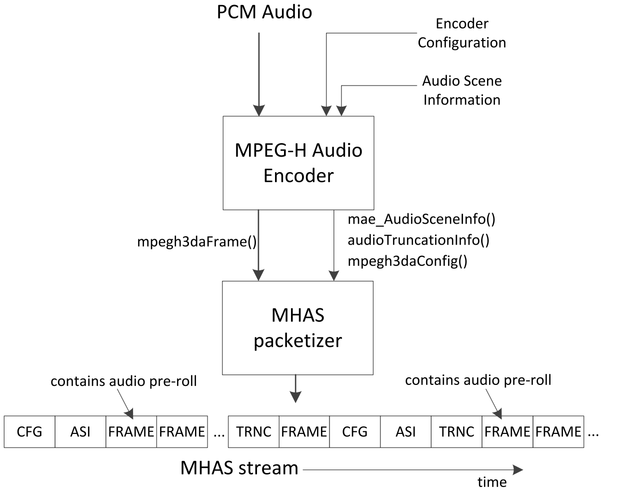


Fig. 30. The output of the MPEG-H encoder is packed into the MHAS bit stream.

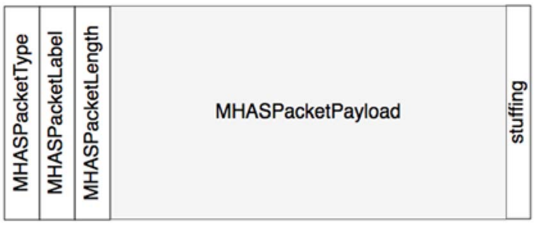


Fig. 31. MHAS packet structure.

Fig. 30 illustrates the workflow of packing the output of a MPEG-H audio encoder into a MHAS stream.

MHAS can either be encapsulated into a MPEG-2 Transport Stream or ISO Base Media File Format (ISOBMFF), as described below.

A MHAS stream is byte aligned and is built from consecutive MHAS packets. An MHAS packet consists of a header with packet type, label, length information and the payload, see Fig. 31.

The most common MHAS packet types are:

- MPEGH3DACFG (CFG) carrying the *mpegh3daConfig()* payload,
- MPEGH3DAFRAME (FRAME) carrying the *mpegh3daframe()* payload,
- AUDIOSCENEINFO (ASI) carrying the *mae_AudioSceneInfo()* payload,
- SYNC for transmission over channels where no frame synchronization is available, and
- AUDIOTRUNCATION (TRNC) carrying the *audioTruncationInfo()* payload, see below.

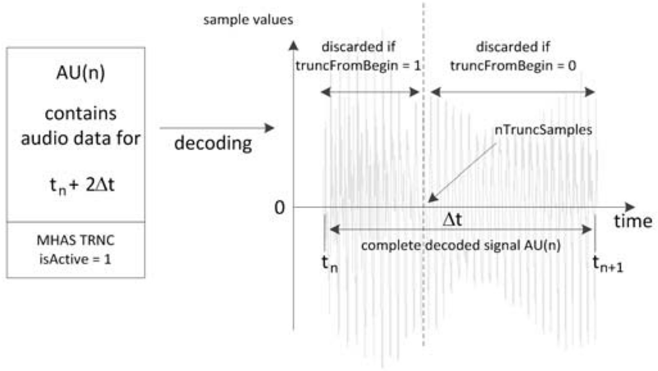


Fig. 32. Applying a truncation to decoded samples of an AU.

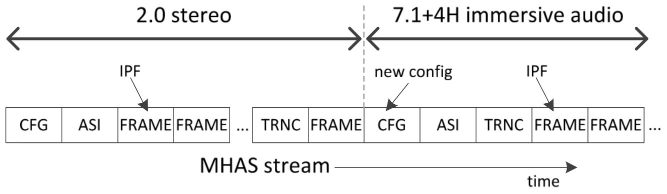


Fig. 33. Example of a configuration change from 2.0 to 7.1+4H in the MHAS stream.

The packet label is used to differentiate between several configurations in one MHAS stream (in case of configuration changes or DASH rate adaptation) and to differentiate between main and auxiliary streams in a multi-stream environment, e.g., if multiple streams are merged into one MHAS stream.

1) Audio Truncation: 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 as shown in Fig. 32. This can be used for alignment of decoded audio data to the video frame boundaries.

The *audioTruncationInfo()* syntax element inside the AUDIOTRUNCATION packet contains a flag to signal if the truncation information should be applied, a flag to signal left or right truncation (*truncFromBegin*) and the number of samples to discard. This syntax element is used whenever decoded time domain samples from an AU should be truncated.

2) Configuration Changes: The MHAS format offers the possibility to perform configuration changes (e.g., from stereo to 7.1+4H) in the stream. Using the audio truncation information as described above, the configuration change is possible with an accuracy of one audio sample. Fig. 33 shows an example of an MHAS packet sequence that implements a sample accurate configuration change from stereo to 7.1+4H immersive configuration.

Note that with the new configuration (in this example for 7.1+4H) the packet label of the stream has to change compared to the previous (in this example 2.0) configuration.

C. Encapsulation of MPEG-H into ISOBMFF

Most of today's adaptive streaming and broadcasting technologies are based on the ISO Base Media File Format (ISOBMFF) [39].

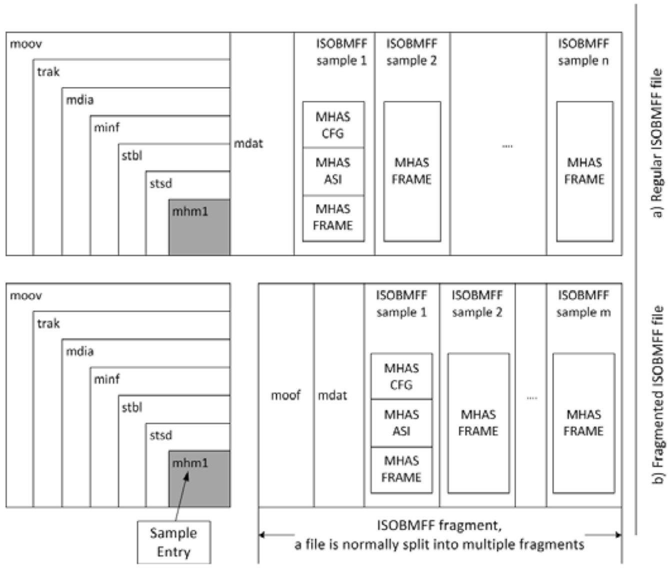


Fig. 34. A regular MPEG-H ISOBMFF file in contrast to a fragmented MPEG-H ISOBMFF file (based on [2]).

A regular ISOBMFF file usually consists of a movie header (moov box) and the encoded media (mdat box) (Fig. 34 top). In a fragmented ISOBMFF file, each fragment is preceded by a fragment header ("moof") followed by that fragment's mdat box (Fig. 34 bottom).

When ISOBMFF is the container format in a streaming environment, usually fragmented ISOBMFF files are used, with the fragment header and its respective mdat box comprising the smallest accessible entity [40], [41].

1) Sample Entries: Each media track in an ISOBMFF file identifies the encapsulated media by a Sample Entry [39] and a respective four-character code (FourCC). The MPEG-H 3D Audio standard defines four Sample Entries for its encapsulation in an ISOBMFF track: "mhm1", "mhm2", "mha1" and "mha2" [42].

For "mha1", a single, static configuration (*mpeg3daConfig*) is stored in the movie header and the mdat box contains just a sequence of AUs. For "mhm1", instead of plain AUs, MHAS packets are stored in the mdat box. While there still can be a static configuration available in the movie header, e.g., to initialize a decoder, MHAS can contain packets of type `MPEGH3DACFG` [1]. This allows configuration changes to be signaled inside the actual media stream.

A similar mechanism [43] is well established for AVC [44] and HEVC [45] and supports the need for dynamic reconfiguration in broadcast and streaming environments.

The samples entries "mha2" and "mhm2" are used if the audio scene is distributed over more than one track or file, e.g., in a scenario where the main audio stream is received by broadcast, and auxiliary streams, such as additional languages, are made available on-demand via broadband. The difference between "mha2" and "mhm2" is the same as the difference between "mha1" and "mhm1", where the former encapsulates plain AUs and the latter encapsulates MHAS packets.

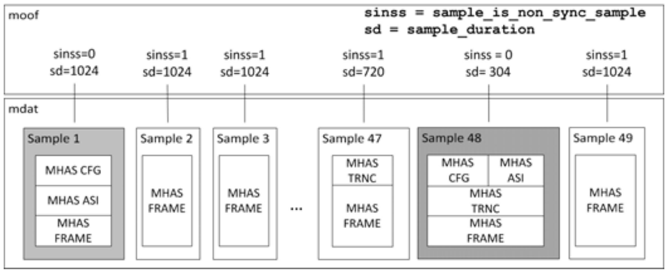


Fig. 35. A fragment containing more than one RAP. Samples with gray background are RAPs.

2) Random Access: The presence and location of a Random Access Point (RAP) (Fig. 35) in the ISOBMFF is signaled by an entry in the sync sample table, which in the case of regular ISOBMFF files is part of the movie header. For subsequent movie fragments, RAPs are signaled by setting the `sample_is_non_sync_sample` [39], [42] flag to 0 for the respective ISOBMFF samples in the fragment header shown in Fig. 35.

In case of mhm1 or mhm2, a RAP ISOBMFF sample consists of a CFG, an ASI and a FRAME MHAS packet. The FRAME packet of a RAP contains an IPF.

Since tune-in into a broadcast stream or bit rate adaptation in a streaming scenario is typically done on movie fragment boundaries, every movie fragment starts with a RAP. For bit rate adaptation, the `applyCrossfade` [46] bit inside the *AudioPreRoll()* extension element should be set to avoid audible artifacts when switching between audio streams.

3) Configuration Changes: In both broadcast and streaming scenarios, it may be necessary to dynamically switch configurations, e.g., when the content changes between programs or during advertisement breaks. MPEG-H typically uses fixed length audio frames with 1024 audio samples per frame at 48 kHz sampling rate, while common video frame rates are 50 or 59.97 frames per second. Thus, audio frame boundaries and video frame boundaries usually do not align. In order to allow for video frame-accurate configuration changes, the audio frame needs to be trimmed to match the respective video frame. This is achieved by embedding MHAS AUDIOTRUNCATION packets with left or right truncation instructions into the MHAS stream.

Fig. 36 shows how this concept works with an example transition from 2.0 stereo to 7.1+4H immersive audio at time t_x. In this example both audio streams are encoded with the same audio frame size.

Please note that in the transition phase two AUs covering the same time span are transmitted. The MHAS truncation mechanism ensures that a decoder will produce the correct amount of audio samples. The first AU of the new configuration will be configured as a RAP.

If an ISOBMFF sample (in case of "mhm1" or "mhm2") contains a right or left truncation, the `sample_duration` [39] (or the `sample_delta` in case of a regular ISOBMFF file) value need to be adjusted accordingly [42].

Streaming systems, especially in case of live streaming, often require movie fragments of a fixed duration. Table IV shows how this can be achieved. (The same example values are also used by Fig. 35.)

TABLE IV
FRAME TRUNCATION AT VIDEO BOUNDARIES

Sample	Sample duration	is_non_sync_sample	MHAS packets	Truncation
1	1024	0	CFG, FRAME	None
2	1024	1	FRAME	None
..	"	"	"	"
47	720	1	FRAME, TRNC	Right
48	304	0	CFG, FRAME, TRNC	Left
49	1024	1	FRAME	None

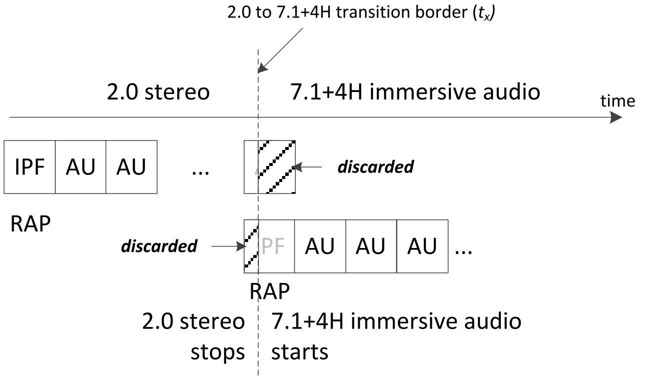


Fig. 36. Configuration change example: With MHAS audio truncation the transition is captured and realized at an audio sample exact position.

D. MMT and ROUTE/DASH

MPEG Media Transport (MMT) [47] and Real-Time Object Delivery over Unidirectional Transport (ROUTE)/Dynamic Adaptive Streaming over HTTP (DASH) [48] are the transport protocols specified in ATSC 3.0. A detailed overview can be found in [49]. Both transport protocols transmit fragmented ISOBMFF files containing Sample Entries "mhm1" or "mhm2".

1) ROUTE/DASH: MPEG-DASH [40] specifies segment formats (ISOBMFF fragments) and the Media Presentation Description (MPD). ATSC 3.0 defines ROUTE [48] to deliver MPEG-DASH content via broadcast.

DASH-IF defines interoperability points for signaling of MPEG-H [50] in the DASH MPD. It supports signaling of MPEG-H stream properties, such as language, accessibility information, role, or "content interactivity", for both single stream and multi-stream use cases. For multi-stream use cases, additional interoperability points are defined, such as the Preselection Element, Selection Priority, or Labels.

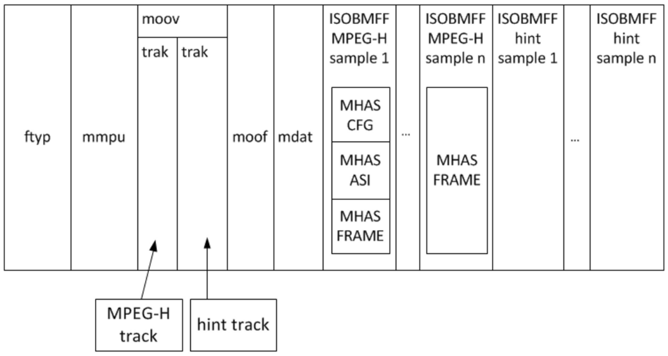


Fig. 37. An MPEG-H MPU based on [47].

The signaling information contained in the MPD enables the receiving device to perform an early selection process before the audio decoder is initialized. This is especially helpful if there are several streams available in a broadcast stream, or if additional optional streams are available on-demand over broadband networks.

2) MMT: One of the building blocks of MMT [47] is the Media Processing Unit (MPU) [51], which contains either timed, as shown in Fig. 37, or non-timed media data. In addition to the usual components of a fragmented ISOBMFF file (movie header, fragment header, media data), an MPU contains an mmpu box, containing unique information about the current MPU, i.e., the `mpu_sequence_number` [47].

To signal MPEG-H in an ATSC 3.0 system, the Audio Stream Properties Descriptor has been defined in [48].

This descriptor is embedded in the MMT Signaling Message ("mmt_atsc3_message()") and provides information about the MPEG-H streams associated with the asset IDs signaled in the descriptor. The descriptor contains basic information about the MPEG-H stream, such as language, accessibility information, role, "content interactivity" and signaling of available audio/aural representations of the emergency information. This descriptive information can be signaled for each preset defined in the Audio Scene Information.

For multi-stream use cases the main stream is always delivered as an MMT broadcast stream and its descriptor provides signaling information for identifing the auxiliary streams. The auxiliary streams can be delivered within the same MMT broadcast stream using different asset IDs, or over broadband using DASH delivery. In the latter case, the auxiliary streams signaled in this descriptor are linked to audio adaptation sets in the DASH MPD using the "bundle_id" and the "stream_id" identifiers.

The signaling information contained in the Audio Stream Properties Descriptor enables the receiving device to perform an early selection process as described above.

VIII. Adapting MPEG-H for TV Broadcasting

The MPEG-H 3D Audio standard describes a state of the art audio codec and renderer, but does not address how that codec is used in actual end-to-end systems. In this section, the adaption and extension of MPEG-H into a complete and tested TV audio system is explained. This includes practical testing to gain information on system requirements, development of additional technology to integrate into professional and consumer infrastructures, and the construction of a test bed to verify the operation of the system.

A. Field Tests During MPEG-H Development

Experience with next-generation audio in TV broadcasting has been limited so far. There have been experimental broadcasts with immersive sound using the NHK 22.2 system and using the Dialog Enhancement system currently standardized in DVB. Prior to the development of the system described in this paper, the authors are not aware of any tests using immersive sound and interactivity together, or converting immersive feature films from the cinema for broadcast playback.

Thus, field tests [52] were conducted during the development of the system to ensure that it would meet the requirements of TV broadcasting. These tests were conducted at live sports broadcasts by recording the host broadcaster's audio console signals, as well as supplemental microphones, and then mixing and encoding the signals later in the studio. These tests included:

- Winter extreme sports competition (skiing, snowboarding, snowmobile racing) carried on a major cable network
- Summer extreme sports competition (skateboarding, motorcycle racing) carried on a major cable network
- NASCAR race (with pit crew radios) using material from NASCAR
- DTM (European race series) auto race carried on major European sports channels

Production of the interactive objects, such as commentary or sound effects, was easy, as they existed as signals on the sound mixer's console or could be routed from the broadcast infrastructure at the event. In some cases, sound effects from spot microphones were mixed as a separate object to allow testing if viewer adjustment would be useful. The control range of a few objects was also limited in order to experiment with such limits.

Adding immersive sound elements was more challenging. Immersive sound is new to sports and will need experience to be used well in an aesthetically convincing way. Since these tests were conducted through the courtesy of host broadcasters, only limited accommodations for pickup of immersive sound could be made in the time before broadcast. Two techniques were used for capturing immersive sound: channel-based microphone trees or shotgun arrays and spherical sound field microphones for capturing scene-based audio.

For channel-based recording, one technique is to construct microphone trees from eight ordinary cardioid microphones pointing towards each corner of a 60-90 cm cube. For scene-based audio capturing, an array of 16 to 32 microphones mounted in a typically 8 cm large metal sphere may be used to capture a sound image. Of course, broadcasters may find that existing microphone arrangements for surround may be also adapted for 3D sound by supplementing them with additional microphones for the overhead channels.

A sound image capture approach may work well for some events, where a 3D microphone can be placed appropriately to capture crowd ambience and other natural sound. For example, 3D microphones worked well to capture the audience sound in half-pipe snowboarding and skateboard ramp events. In other sports, such as downhill skiing, they were impractical for capturing action sounds due to the length of the course.

TV sound is as much about creating an engaging experience through sound design as it is about realistic capture of the live sound, and thus spot microphones, such as practically arranged pairs or groups of shotgun microphones, have been employed to augment or replace 3D microphones. Existing spot microphones of the host broadcaster were also panned (in one case dynamically) to help create a detailed and immersive sound image.

Similarly, tests of converting immersive feature film sound mixes from cinema formats to MPEG-H were made. This included listening tests by the film's sound mixers to validate that the creative intent of the sound mix was preserved during MPEG-H playback in consumer environments. These tests were more straightforward since the film mixes consist of a channel bed containing mid-plane sounds and objects for overhead sounds. Object sounds and their positions were recovered from the film's final sound mix and converted to channel plus object and HOA representations in MPEG-H for consumer playback.

B. Adapting the System to HD-SDI Broadcast Infrastructure

Although it is expected that TV broadcasters will move to IP-based infrastructure in the future, most TV broadcast signal distribution today uses the HD-SDI standard [53], which provides a minimum of 16 48 kHz/24 bit PCM audio channels embedded in each video signal. 16 channels are adequate to carry audio for most MPEG-H productions, but somehow metadata for loudness control, object names, positions, and ranges, as well as the overall composition of the audio scene must be transmitted or stored. The authors have proposed that simple broadcasts could be conducted with fixed settings or independently stored XML data to program the MPEG-H encoders used [54].

Eventually, productions may evolve to using dynamically positioned objects, or the operational complexity of managing out-of-band metadata may become too large. In this case, a method for carrying metadata within the SDI signal must be employed.

While standards exist for carrying metadata in the vertical or horizontal interval of the SDI format [55], much existing equipment does not support them. An alternate approach is to dedicate an audio pair for carrying a metadata data stream or typically, a complete compressed audio stream [56]. This creates an operational burden as each piece of equipment has to be programmed and maintained to treat the channel pair as a "data mode" signal, where no crossfades, filtering, resampling, or gain changes are performed that would cause the audio channel to not be stored and transmitted in a bit-exact manner.

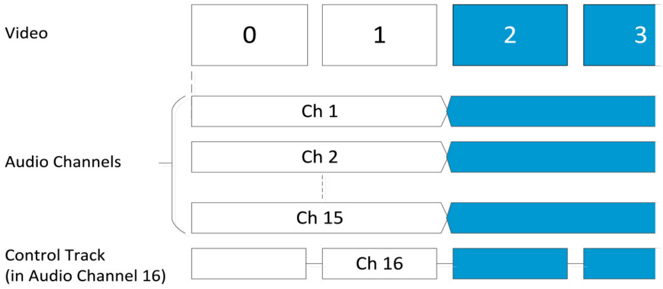


Fig. 38. Switching of HD-SDI signals including the control track signal on audio channel 16 to carry MPEG-H metadata.

In the use of the MPEG-H system over HD-SDI, metadata for the audio signal is collected into packets synchronized with the video signal and is modulated with analog-channel modem techniques into a Control Track signal that fits in the audio channel bandwidth. This signal is unaffected by typical filtering, resampling, or scaling operations in the audio sections of broadcast equipment.

Since the signal includes a guard interval around vertical sync, frame accurate transitions in the audio are automatic as the metadata switches simultaneously with the audio without corruption, as shown in Fig. 38. As the Control Track is just a timecode-like audio signal, it can also be carried as another audio track in audio or video editing systems.

C. Splicing Bit Streams at Video Frame Boundaries

While the Control Track allows SDI signals programming to be cut or switched at any video frame, this creates a concern for the following MPEG-H encoder. Like all perceptual audio codecs, MPEG-H performs time/frequency transforms on a frame basis, with a frame typically being a multiple of 1024 audio samples. The relations between audio and video sampling clocks, raster sizes, video frame rates, and audio codec frame lengths result in there typically being several seconds between instances of exact frame alignment between audio and video frames. This means that a given codec audio frame may contain two different channel formats or scene descriptions if there is a cut at a video frame boundary during the audio frame. Unfortunately, the audio encoder cannot encode two channel formats in one audio frame, and this situation has to be resolved.

One approach is to change the audio frame length to a shorter value that aligns with a given video frame rate, so that an audio frame will only have one configuration. This leads to reduced audio quality for a given bit rate, as the frequency resolution of the time/frequency transforms is reduced.

Instead, in the MPEG-H encoder, a frame consisting of the final audio samples before a video cut is encoded, and another frame consisting of the initial audio samples following a cut is encoded, as shown in Fig. 39. Information is sent in the bit stream indicating the appropriate point to transition between the two frames. This requires encoding and sending an additional audio frame in the bit stream and thus causes a temporary peak in the instantaneous bit rate, but this peak is absorbed by the decoder's input buffer, just as peaks from

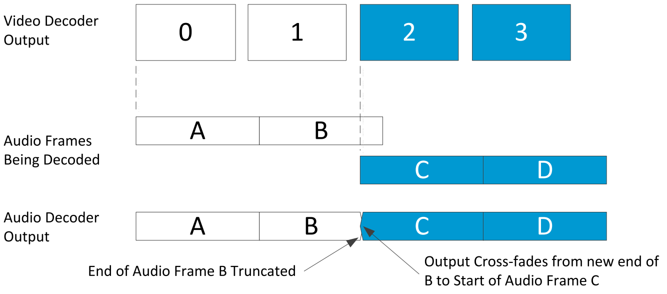

Video Decoder Output
0
1
2
3
Audio Frames Being Decoded
A
B
C
D
Audio Decoder Output
A
B
C
D
End of Audio Frame B Truncated
Output Cross-fades from new end of B to Start of Audio Frame C

Fig. 39. Splicing of codec audio frames to accommodate a new program with different channel configuration beginning at video frame 2. Audio frame B from the previous program is truncated at the decoder output and the playout of audio frame C begins concurrent with the start of video frame 2.

difficult-to-encode audio frames are. This approach offers the ability to cut a program at any video frame boundary while using standard-length audio codec frames for the best coding efficiency.

D. Adapting Existing Broadcast Audio Consoles for MPEG-H Mixing

Today, no TV audio consoles support audio beyond 5.1 channels.[1] Audio consoles for remote production are also surprisingly complex, given the mixing and routing needed in sports productions, and are expensive to replace. Thus, an accessory monitoring unit was developed to transform an existing 5.1 console into one suitable for MPEG-H production. This MPEG-H Audio Monitoring and Authoring Unit (AMAU), shown in Fig. 40 replaced the monitor controller of the console with one that would drive up to 12 loudspeakers and provide downmixes of immersive sound formats.

It is envisioned that broadcasters may prepare some "preset" mixes in addition to the default mix for the casual viewer to quickly choose from. Thus, separate guide metering was included to allow rapid checking of the loudness on each preset, as well as switching to allow monitoring each preset mix.

The parameters of preset mixes, as well as labels describing them and the related audio objects, have to be created. This involves the authoring function of the monitoring unit, which has a web interface to enter this information, as well as the channel configuration, user control range limits, and other parameters into Audio Scene Information for the program. The scene information is constantly encoded in the Control Track along with the loudness and dynamic range metadata for each preset.

Another feature needed is the ability to pan objects in three dimensions in order to track action on screen. The AMAU includes a joystick interface and WebGL panning display for this purpose.

E. MPEG-H in Editing and Post-Production

Ideally, production tools would have support for immersive audio built-in today, with audio being imported and exported in file formats such as ADM-BW64 that support MPEG-H metadata, and the tools allowing the creation and editing of audio scene and interactivity metadata. Of course, such tools would support panning of objects in three dimensions along with conversion between and downmixing of immersive audio formats, as well as previewing consumer interactivity.

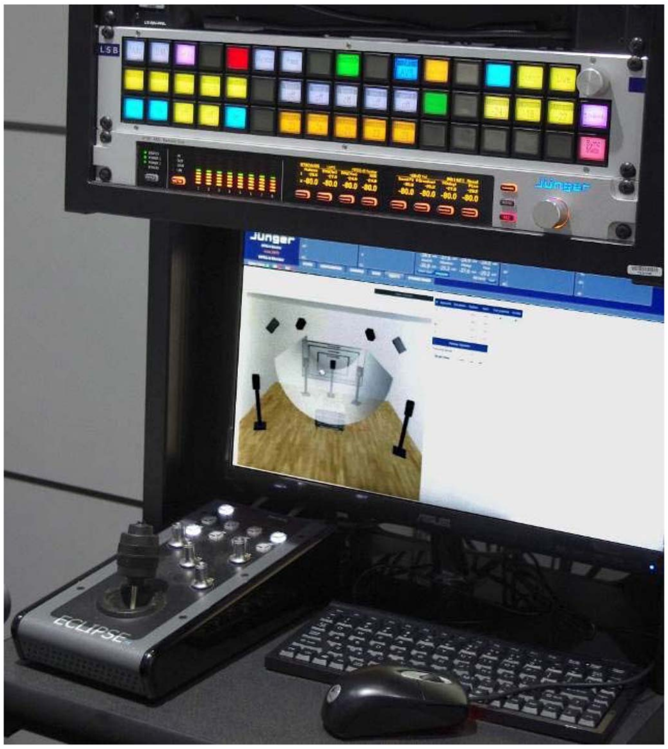


Fig. 40. MPEG-H Audio Monitoring and Authoring Unit (AMAU).

Unfortunately, even tools for the more mature cinema industry used for non-interactive immersive sound mixing have not reached this level of support, relying on plug-ins (auxiliary software integrated at run-time through host tool programming interfaces such as VST or AAX) and external hardware to mix immersive sound.

In the development of the MPEG-H system, work was conducted in parallel on tools for live production and those for off-line or post-production. Live tools were given priority since the live case imposes much more demanding requirements on the tools and underlying MPEG-H technologies.

Concepts used for the live case, including the Control Track signal, may be used as a transitional strategy for post-production editing. Audio for the MPEG-H system with the Control Track can be recorded on existing video servers and ingest into current video editors and digital audio workstations. Editing can be performed at any frame just as with stereo or surround content, even when different MPEG-H audio formats are concatenated, as shown in Fig. 41.

Changing the audio scene, panning, or loudness metadata requires regenerating the Control Track signal with an Audio Monitoring and Authoring Unit. This can be done by playback in the editing software with punch-in recording enabled for the channel containing the Control Track.

[1] *Except for special consoles built for experimental use with the NHK 22.2 system.*

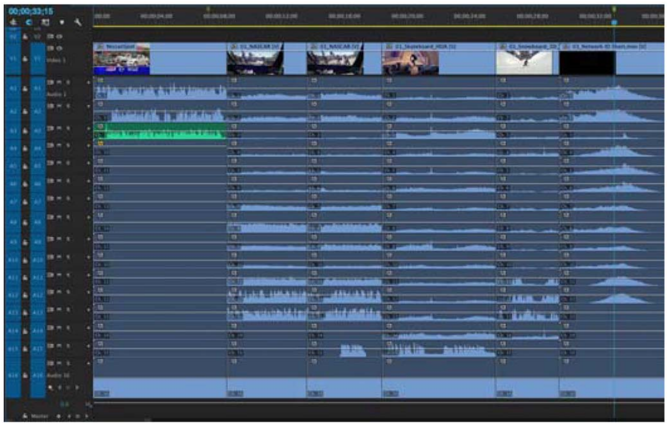


Fig. 41. Editing a highlight reel of test bed content in several audio formats in Adobe Premiere using the control track for carrying audio metadata. Items N150, R 154, F 144, (live), and P 152 were used from Fig. 42.

Using the AMAU allowed editing the majority of the content used for field tests and the test bed described below, as shown in Fig. 41. Simultaneously, work has been conducted to transfer the software libraries developed for the AMAU to plug-ins for software editing applications. These plug-ins can generate or edit the Control Track internally through dialogs presented in the host application. Work is underway on companion applications to translate Control Track based files to and from ADM-BW64 files.

F. Constructing a Test Bed for System Validation

In order to test the MPEG-H system design and discover any missing features or performance limitations, as well as test the use of the system by broadcast creatives, the consortium constructed a test bed representative of a complete TV network signal chain from a remote broadcast to a consumer's receiver. The test bed was organized into four rooms:

- A simulated remote truck where pre-recorded microphone signals from an extreme sports event were mixed and panned live on an unmodified audio console equipped with a MPEG-H Audio Monitoring and Authoring Unit.
- A Network Operations Center (NOC) where the live remote truck signal was switched under automation control with recorded programming from a video server.
- A local affiliate station, WMPG-TV, where local commercials were produced and inserted into the network feed.
- A consumer living room where the content was played back on a set-top box and prototype 3D sound bar. Immersive feature film excerpts could also be received from the MPEG Network's Cable Movie Channel on the set-top box.

The test bed was designed to make use of existing commercially available equipment where possible and to be operated by broadcast industry personnel. Demonstrations were run continuously on a fixed schedule by playout automation in the NOC and WMPG-TV during the NAB 2015 convention and at a special demonstration event organized by the ATSC in August 2015.

In order to test the different modes and possibilities of the MPEG-H system, content was prepared in 13 different formats, as shown in Fig. 42. This allowed testing system features such as the transmission of metadata for dynamic objects and the frame-accurate switching of content in different formats using the Control Track signal.

As shown in the block diagram in Fig. 43, the testbed included three audio/video links between facilities: A contribution link from the remote truck to the NOC, a distribution link from the NOC to WMPG-TV, and an emission link from WMPG-TV to the living room. These links were operated using MPEG-H for audio and AVC for video, multiplexed into an MPEG-2 transport stream and transmitted over an IP network.[2]

Since commercial MPEG-H encoders and decoders were not yet available during the design of the testbed, encoders and decoders were constructed using the GStreamer framework. All encoders and decoders shared the same hardware platform and software framework.

Encoders and decoders for contribution or distribution and those for emission differed only in the configuration of the software framework. Contribution and distribution encoders operated with a fixed channel mapping that encoded 15 audio channels at a higher contribution-quality bit rate. The control track was demodulated and the metadata transmitted as digital side information in the MPEG-H bit stream. The contribution/distribution decoders decoded the 15 audio channels and regenerated the control track from the metadata. The emission encoder encoded the audio to the channel configuration given in the control track and performed configuration changes as required by different content formats. The set-top box rendered the audio to the connected loudspeakers or to the 3D soundbar.

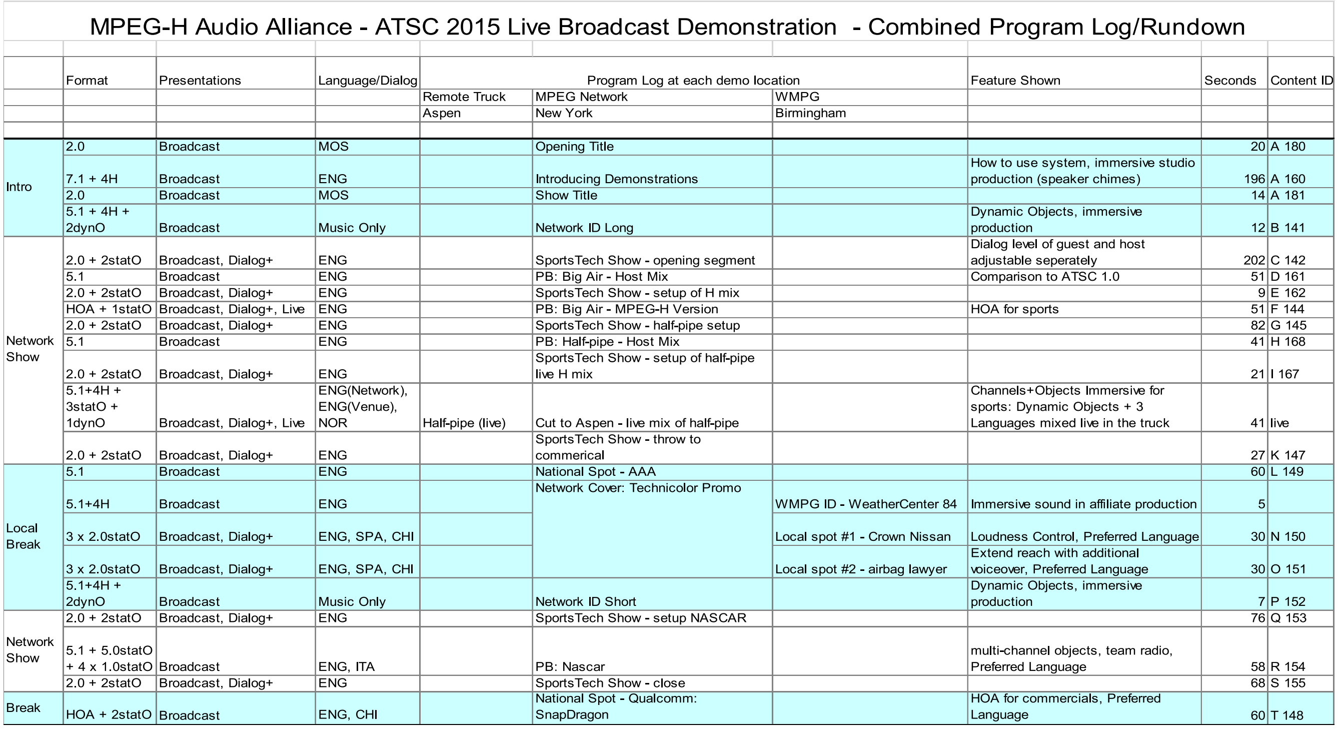
G. Operational Examples

1) Recommended Bit Rates: Subjective listening tests using MPEG-H 3D Audio were conducted during the development of the MPEG-H 3D Audio standard in MPEG, in the ATSC 3.0 standards development process, and on other occasions. Based on the test results, the bit rates shown in Table V for a given configuration can be recommended. Using these bit rates ensures a constantly excellent audio quality for broadcast applications.

It should be noted that these bit rates reflect the current status of the encoder development and may change to lower values over time. They do not include the overhead of transport and metadata side information. Typically, TV broadcast audio is transmitted at "broadcast quality" (4.0 or greater rating according to BS.1116). The MPEG-H system may also be operated at lower bit rates for services with restricted bandwidth or in adaptive streaming to preserve audio during network congestion.

2) Legacy 5.1 Operation: The MPEG-H system may be easily used in situations for legacy content in 5.1 surround or stereo without explicitly considering the advanced next-generation audio system features. For example, 5.1 surround

[2] *The authors chose MPEG-2 TS transport since DASH/Route and MMT standards were not sufficiently mature at the time the test bed were designed. Meanwhile, broadcast encoders and TV sets implementing MPEG-H audio, HEVC video and ATSC 3.0 transport have become commercially available.*



MPEG-H Audio Alliance - ATSC 2015 Live Broadcast Demonstration - Combined Program Log/Rundown

	Format	Presentations	Language/Dialog	Program Log at each demo location			Feature Shown	Seconds	Content ID
				Remote Truck	MPEG Network	WMPG			
				Aspen	New York	Birmingham			
Intro	2.0	Broadcast	MOS		Opening Title			20	A 180
	7.1 + 4H	Broadcast	ENG		Introducing Demonstrations		How to use system, immersive studio production (speaker chimes)	196	A 160
	2.0	Broadcast	MOS		Show Title			14	A 181
	5.1 + 4H + 2dynO	Broadcast	Music Only		Network ID Long		Dynamic Objects, immersive production	12	B 141
Network Show	2.0 + 2statO	Broadcast, Dialog+	ENG		SportsTech Show - opening segment		Dialog level of guest and host adjustable seperately	202	C 142
	5.1	Broadcast	ENG		PB: Big Air - Host Mix		Comparison to ATSC 1.0	51	D 161
	2.0 + 2statO	Broadcast, Dialog+	ENG		SportsTech Show - setup of H mix			9	E 162
	HOA + 1statO	Broadcast, Dialog+, Live	ENG		PB: Big Air - MPEG-H Version		HOA for sports	51	F 144
	2.0 + 2statO	Broadcast, Dialog+	ENG		SportsTech Show - half-pipe setup			82	G 145
	5.1	Broadcast	ENG		PB: Half-pipe - Host Mix			41	H 168
	2.0 + 2statO	Broadcast, Dialog+	ENG		SportsTech Show - setup of half-pipe live H mix			21	I 167
	5.1+4H + 3statO + 1dynO	Broadcast, Dialog+, Live	ENG(Network), ENG(Venue), NOR	Half-pipe (live)	Cut to Aspen - live mix of half-pipe		Channels+Objects Immersive for sports: Dynamic Objects + 3 Languages mixed live in the truck	41	live
	2.0 + 2statO	Broadcast, Dialog+	ENG		SportsTech Show - throw to commerical			27	K 147
Local Break	5.1	Broadcast	ENG		National Spot - AAA			60	L 149
	5.1+4H	Broadcast	ENG		Network Cover: Technicolor Promo	WMPG ID - WeatherCenter 84	Immersive sound in affiliate production	5	
	3 x 2.0statO	Broadcast, Dialog+	ENG, SPA, CHI			Local spot #1 - Crown Nissan	Loudness Control, Preferred Language	30	N 150
	3 x 2.0statO	Broadcast, Dialog+	ENG, SPA, CHI			Local spot #2 - airbag lawyer	Extend reach with additional voiceover, Preferred Language	30	O 151
	5.1+4H + 2dynO	Broadcast	Music Only		Network ID Short		Dynamic Objects, immersive production	7	P 152
Network Show	2.0 + 2statO	Broadcast, Dialog+	ENG		SportsTech Show - setup NASCAR			76	Q 153
	5.1 + 5.0statO + 4 x 1.0statO	Broadcast	ENG, ITA		PB: Nascar		multi-channel objects, team radio, Preferred Language	58	R 154
	2.0 + 2statO	Broadcast, Dialog+	ENG		SportsTech Show - close			68	S 155
Break	HOA + 2statO	Broadcast	ENG, CHI		National Spot - Qualcomm: SnapDragon		HOA for commercials, Preferred Language	60	T 148

Fig. 42. Program log showing audio formats used in the test bed. (H indicates overhead Height channels, statO indicates static objects, and dynO indicates dynamic (moving) objects). The schedule repeated every twenty minutes during demonstrations.

TABLE V
RECOMMENDED CORE BIT RATES FOR EXCELLENT AUDIO QUALITY FOR BROADCAST APPLICATIONS

Channel Configuration	Bit Rate
2.0 Stereo	96 kbps
5.1 Multi-channel Surround	192 kbps
7.1 + 4H Immersive Audio with 4 Height Speakers	384 kbps
22.2 Immersive Audio	768 kbps

may be carried as shown in Table VI. The six channels of audio are supplied in the embedded HD-SDI input (or four HD-SDI inputs for 4K video) of the ASTC 3.0 encoder. Program loudness of the content is pre-normalized to -24 LKFS or other standard value, and standard legacy downmixing gains are used. The MPEG-H encoder configuration is set by encoder menus.

3) Stereo with VDS and Dialogue Objects: A more advanced use case is to provide dialogue and video descriptive services as separate objects to the ATSC 3.0 encoder. This allows receiving devices of visually impaired viewers to automatically enable the video description narration with appropriate ducking of the main audio. Regular viewers hear the normal dialogue of the program. In this case, a mix-minus of natural sound, music and effects is provided as stereo, and the main dialogue and video description narration are provided as mono channels to the ATSC 3.0 encoder. The MPEG-H encoder configuration is set by encoder menus or possibly automation signals.

4) 5.1+4H Immersive Sound with VDS, Multiple Language Dialogue, and Auxiliary Objects: In this case, the example above is expanded to immersive sound in 5.1+4H and by adding additional languages as well as a team radio channel (which is not part of the normal broadcast mix but is enabled by viewer interaction). In this case, the control track is used during production or post production to provide the metadata needed to configure the MPEG-H encoder and to label the audio objects. The object for team radio requires a label that varies from one program to the next, thus the control track needs to be used instead of a fixed encoder configuration. With the control track, programming may be normalized to a standard loudness value, or agile loudness metadata may be carried in the control track.

IX. ADAPTING MPEG-H FOR CONSUMER DELIVERY

A. User Interfaces for Audio Control

Consumer electronics TV products have traditionally included one interactive control for audio – volume. With the MPEG-H system, content can have multiple preset mixes as well as many individual controls for object prominence levels or even spatial positions. A user interface must be provided in consumer devices for selecting and controlling these options.

Given that the user will naturally operate the remote control (or similar user interface) for the source device he is watching, it is most convenient to place the MPEG-H user interface software on the source device, so that he only has to use a single remote control.

To achieve this, the MPEG-H system allows the functions of MPEG-H audio decoding and MPEG-H user interactivity handling to be separated, as shown in Fig. 44. The MPEG-H system combines all user interface related tasks into the

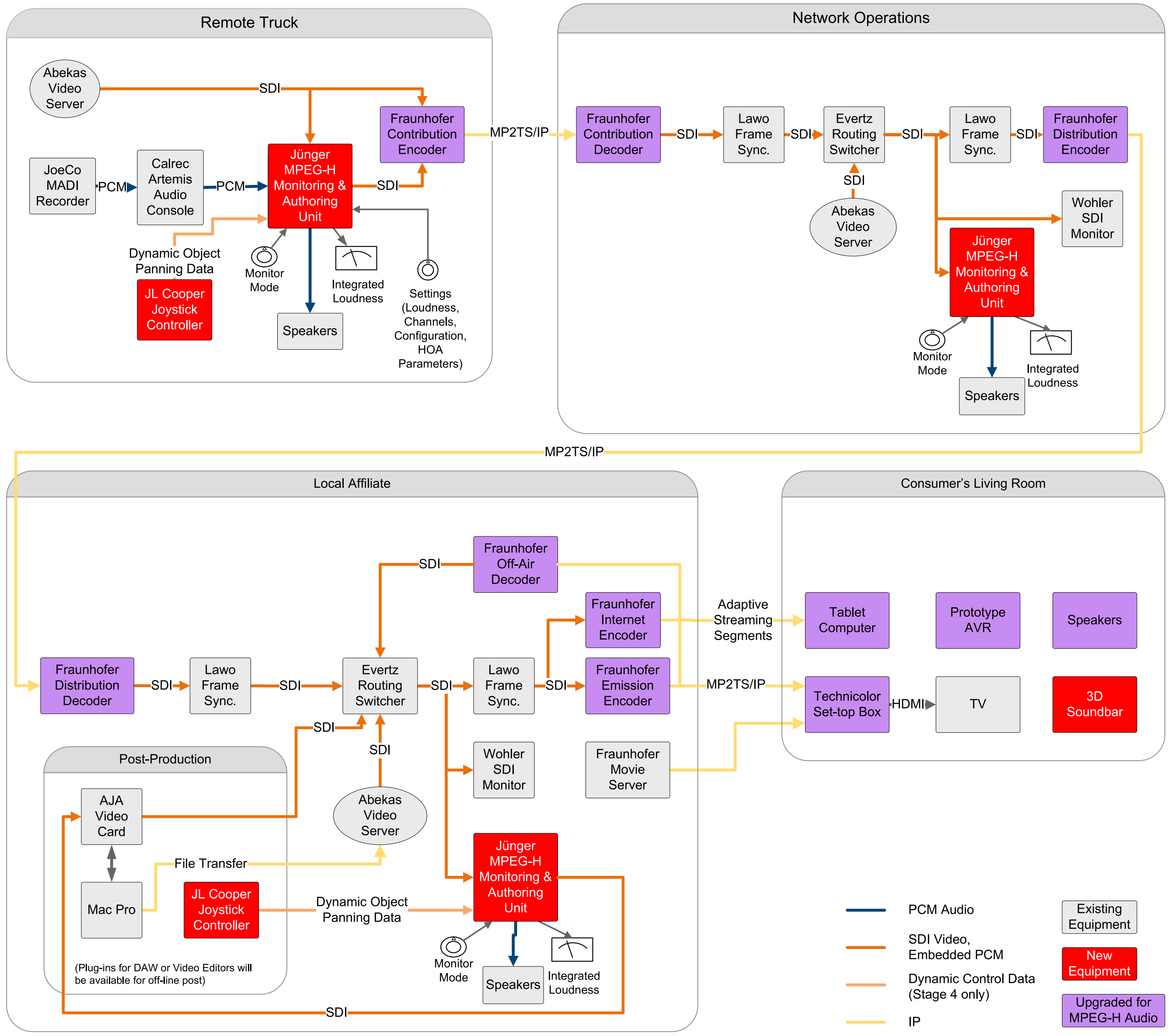


Fig. 43. Block diagram of the test bed.

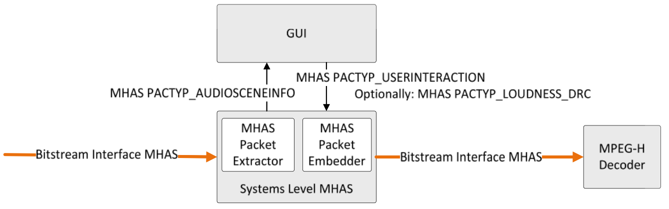


Fig. 44. Processing GUI interactivity commands in a source device and including MHAS USERINTERACTION Packets in the MHAS stream for processing by an MPEG-H decoder in a sink device.

MPEG-H UI manager component. The MPEG-H UI manager component is a separate module that can be integrated in the device that is responsible for the user interface (e.g., by On-Screen-Display) rendering. The MPEG-H UI manager component inserts additional information into a processed MPEG-H bit stream, as shown in Fig. 45, so that the user inputs are considered in the subsequent MPEG-H audio decoding step. This technique allows the user interface to reside in the source device, such as a TV or set-top box, while the MPEG-H audio decoding is done in an audio/video receiver or soundbar. This technique also allows easy partitioning of the MPEG-H software between processors, even if all MPEG-H functions are in the same device.

B. Control Persistence

Programming that includes interactivity requires that user settings can be maintained for recent programs. For example, if a user selects a certain mix preset for a program or adjusts an audio object's presence, his or her expectation is that the setting will be maintained for the duration of the program. The program may be interrupted by commercials or news bulletins or the user may momentarily switch to another channel. Thus, the user control settings for each recent content item are stored

TABLE VI
HD-SDI EMBEDDED AUDIO CHANNEL ASSIGNMENTS FOR OPERATIONAL EXAMPLES

HD-SDI Embedded Audio Channel Number	Legacy 5.1	Stereo with VDS	5.1+4H with VDS, Multi-language, Aux Objects
1	L	L (M&E)	L
2	R	R (M&E)	R
3	C	Main Dialogue	C
4	LFE	Video Description	LFE
5	Ls		Ls
6	Rs		Rs
7			Top Front Left
8			Top Front Right
9			Top Back Left
10			Top Back Right
11			Main Dialogue - English
12			Main Dialogue - Korean
13			Team Radio #1
14			Team Radio #2
15			Video Description
16			Control Track

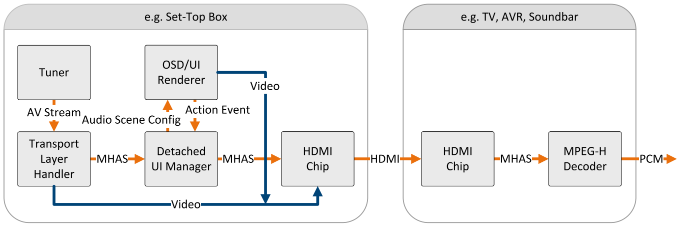


Fig. 45. Separation of MPEG-H UI manager from MPEG-H decoder component.

in the playback device (i.e., TV or STB) under the control of the MPEG-H user interface manager.

C. Interconnectivity in the Home Environment

While provisions have been made in the HDMI 2.0 and 2.1 standards for carrying up to 32 channels of immersive audio, these portions of the standards have not been implemented in any consumer devices. For the near future, MPEG-H must rely on the features of HDMI 1.4 for transport between connected devices in the consumer's living room. A related issue is that consumers often prefer to connect their audio/video receiver or soundbar to the TV or display instead of inserting it before the TV in the signal chain. This requires use of the Audio Return Channel (ARC) feature of HDMI 1.4 to carry the audio from the TV to the audio device in the backwards direction on an HDMI cable, or the use of the legacy S/PDIF cable.

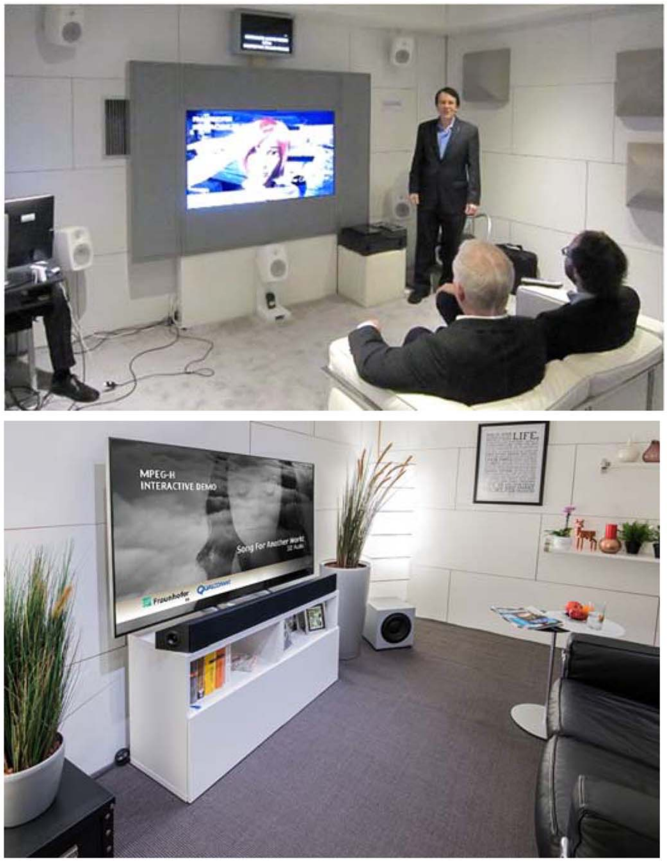


Fig. 46. 3D soundbar designs: (above) First-generation design. (below) Second-generation design. Wall-mounted loudspeakers in the left photo were used for discrete playback to simulate professional monitoring and enthusiast listening.

The MPEG-H system has several modes to operate within these constraints. In all cases, compressed MPEG-H bit streams using MHAS format can be carried in forward or ARC paths using the standard IEC 61937 [57] packing. MPEG-H has also been included in the latest version of the CTA-861 [58] standard for EDID negotiation to discover MPEG-H device capabilities over HDMI.

In some cases, it may be desirable to carry uncompressed rendered immersive audio over HDMI. For this purpose, the above standards support FIAS, a format for packing PCM data into the HBR Audio Stream Packet introduced in HDMI 1.3. FIAS supports uncompressed audio in loudspeaker configurations up to 9.1 + 4H.

D. Immersive Sound Playback with 3D Soundbars

Although individual loudspeakers provide the highest spatial accuracy and were used in the professional rooms of the test bed, the expense of installing ceiling-mounted loudspeakers for immersive sound at home likely limits this approach to enthusiasts. In order to enable mainstream consumers to experience immersive sound, a "3D Soundbar" as shown in Fig. 46 was developed. The initial concept prototype used in the test bed was a frame of loudspeakers surrounding the TV.

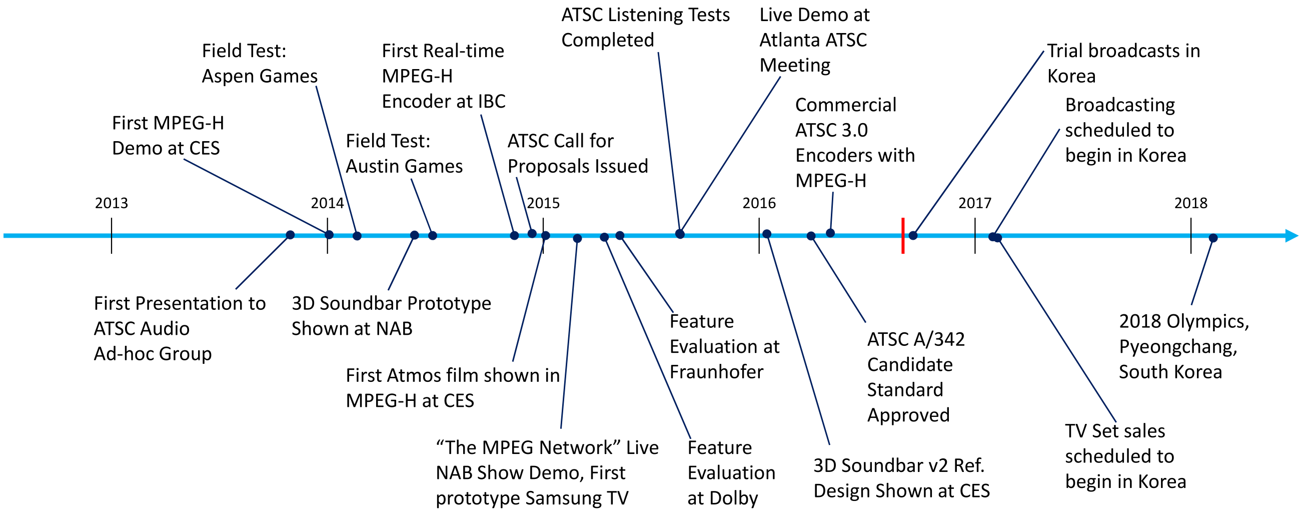

2013
2014
2015
2016
2017
2018
First MPEG-H Demo at CES
Field Test: Aspen Games
First Real-time MPEG-H Encoder at IBC
Field Test: Austin Games
ATSC Call for Proposals Issued
ATSC Listening Tests Completed
Live Demo at Atlanta ATSC Meeting
Commercial ATSC 3.0 Encoders with MPEG-H
Trial broadcasts in Korea
Broadcasting scheduled to begin in Korea
First Presentation to ATSC Audio Ad-hoc Group
3D Soundbar Prototype Shown at NAB
First Atmos film shown in MPEG-H at CES
"The MPEG Network" Live NAB Show Demo, First prototype Samsung TV
Feature Evaluation at Fraunhofer
Feature Evaluation at Dolby
ATSC A/342 Candidate Standard Approved
3D Soundbar v2 Ref. Design Shown at CES
TV Set sales scheduled to begin in Korea
2018 Olympics, Pyeongchang, South Korea

Fig. 47. Timeline of MPEG-H development in ATSC 3.0.

This was later reduced to a traditional soundbar form suitable for consumer manufacturing. The soundbar connects to the TV or other program sources over traditional HDMI 1.4 connections and provides a realistic immersive sound image within a wide listening area.

E. Rendering for Tablet Playback and Virtual Reality

In addition to the rendering available in the MPEG-H standard, the system offers the possibility to interface to external renderers. For example, to demonstrate playback of immersive content on tablet computers, the MPEG-H system decoder was combined with Fraunhofer Cingo, a proprietary binaural rendering product used in tablet computers and virtual reality headsets. This allowed rendering using the tablet computer's internal loudspeakers for handheld playback. In virtual reality applications, for realistic playback the rendering must be adjusted for the user's head position to align the audio image with the video viewport. This rendering to head tracking data is provided by Cingo for all MPEG-H formats – channels, objects, and scene-based components. MPEG-H also includes the metadata needed to mark certain signals as non-diegetic so they are rendered without head tracking.

X. MPEG-H and the ATSC 3.0 Process

A. Early Planning

In 2010, the ATSC held a symposium on new technologies for next-generation broadcast television. At that time, some of the authors wrote in a submitted paper:

> *A completely different approach is to leave the concept of channel-based audio production behind, and use an object-based approach instead. In this solution, a number of audio objects are described, along with a scene description, and these parameters are transmitted to the receiver, which renders the audio signal in accordance with the physical audio reproduction setup in the user's environment. This solution will provide for all possible setups, including perhaps those not yet envisioned at the time of the system's introduction. The optimal solution would support reproduction techniques like wave field synthesis or ambisonics.*

In 2012 a working group of the ATSC was formed to consider audio systems for the ATSC 3.0 standard. A consortium of companies, including those of the authors, made a proposal to the working group in November 2013 for developing a system based on the upcoming MPEG-H 3D Audio standard, as shown in Fig. 47.

B. Call for Proposals and the Evaluation Process

In December 2014, the ATSC issued a call for proposals for an ATSC 3.0 audio system, describing the desired features of a new audio system and its evaluation process. The MPEG-H consortium system was proposed in March 2015 along with two other systems. These systems were evaluated for basic audio performance in a pre-screening phase and two systems continued into a formal multi-site double-blind listening test. This test showed both systems offered satisfactory audio quality, with a preference for the MPEG-H system on some content items, particularly items with speech where the speech coding tools of the MPEG-H codec could be employed.

In May and June 2015, evaluators from the ATSC conducted visits to the proponent's laboratories to evaluate demonstrations of the features required by the call for proposals. Both systems were evaluated to have met the ATSC requirements. In July 2015, the ATSC organized a special demonstration event where proponents could demonstrate their systems. The MPEG-H test bed (see Section VIII.F) was demonstrated at this event.

In November 2015, the ATSC elevated both systems to Candidate Standard status with a recommendation to use one system per region. In November 2016, both systems were elevated to ATSC Proposed Standards.

XI. Conclusion and the Future

Testing in the MPEG-H testbed and during the ATSC 3.0 standards development process has proven the MPEG-H

TV Audio System meets the requirements of broadcasters for a next-generation audio system. MPEG-H has been adopted as a Proposed Standard in ATSC 3.0 and has been selected by South Korea as the sole audio system for ATSC 3.0 broadcasting. Test broadcasts using MPEG-H in ATSC 3.0 began in Korea in November 2016 and full broadcasting service is planned for February 2017. It is expected that TV receivers for ATSC 3.0 (including the MPEG-H system) will go on sale to Korean consumers in February 2017 and that Korean broadcasters will broadcast the 2018 Olympics in Pyeongchang, South Korea in UHD and MPEG-H.

MPEG-H is also now a part of the DVB UHD standards family and is being considered for other TV standards.

Acknowledgment

A project of this dimension can only ever be realized by a joint effort of a considerable number of highly skilled experts. Therefore, the authors would like to extend their appreciation to the following list of contributors for their important and substantial work over the long duration of creation of the MPEG-H TV Audio System:

Bernhard Grill, Harald Popp, Nikolaus Rettelbach, Stefan Krägeloh, Marc Gayer, Rinat Zeh, Wolfgang Fiesel, Jan Plogsties, Andreas Silzle, Merrill Weiss, Jim De Filippis, Michael Meier, Oliver Wübbolt, Christian Borss, Jan Gregor Gatzmann, Bernhard Neugebauer, Emmanuel Ravelli, Matthias Neusinger, Nicolas Faecks, Matteo Agnelli, Bernd Herrmann, Christian Ertel, Ralf Geiger, Andreas Hölzer, Stephan Wilde, Markus Werner, Klaus Berghammer, Michael Fischer, M. Fischer, Pavan Kantharaju, Michael Härtl, Manuel Jander, Matthias Hildenbrand, Arthur Tritthart, Ingo Hofmann, Srikanth Korse, Herbert Thoma, Roland Clobus, Christian R. Helmrich, Markus Multrus, Jérémie Lecomte, Adrian Tomasek, Grzegorz Pietrzyk, Arne Borsum, Stephan Schreiner, Leonardo Sabino dos Santos, Felix Fleischmann, Debrupa Chakraborty, Matthias Felix, Kathrin Bogner, Matthias Rose, Johannes Schmidl, Michael Mauser, Stefan Meltzer, Jan Nordmann, Julien Robilliard, Axel Horndasch, Alexander Adami, Christian Neukam, Ulli Scuda, Valentin Schilling, Hanne Stenzel, Jim Hilson, Tobias Bliem, Florian Kolbeck, Sven Heinicke, Steven Wagner, Andreas Walther, Julian Klapp, Yannik Grewe, Michael Schnabel, Josef Höpfl, Matthias Hüttenberger, Youliy Ninov, Ferdinand Packi, Alfonso Pino Garcia, Lukas Schuster, Fabian Bauer, Thomas Blender, Lukas Lamprecht, Ralph Sperschneider, Martin Morrell, and Venkatraman Atti.

Furthermore, the authors would like to express their gratitude to all members of the MPEG Audio Subgroup for their support and collaborative work during the standardization process.

References

[1] *Information Technology—High Efficiency Coding and Media Delivery in Heterogeneous Environments—Part 3: 3D Audio*, ISO/IEC Standard ISO/IEC 23008-3:2015, 2015.

[2] *A/342 Part 3: ATSC Candidate Standard—MPEG-H System*, Standard A/342, 2016.

[3] A. Silzle, S. George, E. A. P. Habets, and T. Bachman, "Investigation on the quality of 3D sound reproduction," in *Proc. Int. Conf. Spatial Audio*, Detmold, Germany, 2011, pp. 334–341.

[4] *Information Technology—MPEG Systems Technologies—Part 8: Coding-Independent Code Points*, ISO/IEC Standard ISO/IEC 23001-8:2016, 2016.

[5] R. Bleidt. *Surround Sound Application Bulletin*. Accessed on Feb. 12, 2017. [Online]. Available: http://www.iis.fraunhofer.de/content/dam/iis/de/doc/ame/wp/FraunhoferIIS_Application-Bulletin_Fraunhofer-Surround-Codecs.pdf

[6] "Multichannel sound technology in home and broadcasting applications," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU R-REP-BS.2159-7-2015, 2015.

[7] J. Daniel, "Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia," Ph.D. dissertation, Laboratoire d'Acoustique Musicale, Univ. at Paris VI, Paris, France, 2000.

[8] M. Noisternig, T. Musil, A. Sontacchi, and R. Holdrich, "3D binaural sound reproduction using a virtual ambisonic approach," in *Proc. IEEE Int. Symp. Virtual Environ. Human Comput. Interfaces Meas. Syst.*, Lugano, Switzerland, 2003, pp. 174–178.

[9] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904–926, 2001.

[10] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *J. Audio Eng. Soc.*, vol. 53, no. 11, pp. 1004–1025, 2005.

[11] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *J. Acoust. Soc. America*, vol. 116, no. 4, pp. 2149–2157, 2004.

[12] F. Zotter, M. Frank, and H. Pomberger, "Comparison of energy-preserving and all-round ambisonic decoders," in *Fortschritte der Akustik AIA-DAGA*, Merano, Italy, 2013, pp. 2368–2371.

[13] A. J. Heller, E. M. Benjamin, and R. Lee, "A toolkit for the design of ambisonic decoders," in *Proc. Linux Audio Conf.*, Palo Alto, CA, USA, 2012, pp. 1–12.

[14] H. Fuchs, S. Tuff, and C. Bustad, "Dialogue enhancement—Technology and experiments," *EBU Technol. Rev.*, pp. 1–11, Jun. 2012. [Online]. Available: https://tech.ebu.ch/docs/techreview/trev_2012-Q2_Dialogue-Enhancement_Fuchs.pdf

[15] *Information Technology—MPEG Audio Technologies—Part 4: Dynamic Range Control*, ISO/IEC Standard ISO/IEC 23003-4:2015, 2015.

[16] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards*. Boston, MA, USA: Kluwer, 2002.

[17] S. Disch *et al.*, "Intelligent gap filling in perceptual transform coding of audio," in *Proc. Audio Eng. Soc. Conv.*, vol. 141. Los Angeles, CA, USA, Sep. 2016, p. 9661.

[18] C. R. Helmrich, A. Niedermeier, S. Disch, and F. Ghido, "Spectral envelope reconstruction via IGF for audio transform coding," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, South Brisbane, QLD, Australia, 2015, pp. 389–393.

[19] S. Disch, C. Neukam, and K. Schmidt, "Temporal tile shaping for spectral gap filling in audio transform coding in EVS," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, South Brisbane, QLD, Australia, 2015, pp. 5873–5877.

[20] F. Schuh *et al.*, "Efficient multichannel audio transform coding with low delay and complexity," *Audio Eng. Soc. Conv. Paper*, vol. 141. Los Angeles, CA, USA, Sep. 2016, p. 9660.

[21] C. R. Helmrich, A. Niedermeier, S. Bayer, and B. Edler, "Low-complexity semi-parametric joint-stereo audio transform coding," in *Proc. EURASIP EUSIPCO*, Nice, France, 2015, pp. 794–798.

[22] N. Peters, D. Sen, M.-Y. Kim, O. Wuebbolt, and S. M. Weiss, "Scene-based audio implemented with higher order ambisonics (HOA)," in *Proc. SMPTE Annu. Tech. Conf. Exhibit.*, Hollywood, CA, USA, 2015, pp. 1–13.

[23] V. K. Goyal, "Theoretical foundations of transform coding," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 9–21, Sep. 2001.

[24] D. Sen, N. Peters, M. Y. Kim, and M. J. Morrell, "Efficient compression and transportation of scene-based audio for television broadcast," in *Proc. AES Int. Conf. Sound Field Control*, Guildford, U.K., 2016, p. 2–1.

[25] S. Füg *et al.*, "Design, coding and processing of metadata for object-based interactive audio," in *Proc. 137th AES Conv.*, Los Angeles, CA, USA, 2014, p. 9097.

[26] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H audio—The new standard for universal spatial/3D audio coding," in *Proc. Audio Eng. Soc. 137th Conv.*, Los Angeles, CA, USA, 2014, pp. 479–490.

[27] S. Füg, D. Marston, and S. Norcross, "The audio definition model—A flexible standardized representation for next generation audio content in broadcasting and beyond," in *Proc. 141st AES Conv.*, Los Angeles, CA, USA, 2016, p. 9626.

[28] "Audio definition model," ITU-R, Geneva, Switzerland, Recommendation BS.2076, 2015.

[29] *Audio Definition Model—Metadata Specification*, European Broadcasting Union Standard EBU Tech 3364, 2014.

[30] *EBU Core Metadata Set*, European Broadcasting Union Standard EBU Tech 3293, 2015.

[31] "Long-form file format for the international exchange of audio programme materials with metadata," ITU-R, Geneva, Switzerland, Recommendation BS.2088, 2015.

[32] V. Pulkki, "Uniform spreading of amplitude panned virtual sources," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, 1999, pp. 187–190.

[33] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, Jun. 1997.

[34] *Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level*, ITU Standard ITU-R BS.1770-4, 2015.

[35] *Loudness Metering: "EBU Mode" Metering to Supplement EBU R 128 Loudness Normalization*, European Broadcasting Union Standard EBU-Tech 3341, 2016.

[36] *Techniques for Establishing and Maintaining Audio Loudness for Digital Television*, Advanced Television Systems Committee Standard ATSC A/85, 2013.

[37] *Loudness Range: A Measure to Supplement EBU R 128 Loudness Normalization*, European Broadcasting Union Standard EBU-Tech 3342, 2016.

[38] *Requirements for Loudness and True-Peak Indicating Meters*, ITU Standard ITU-R BS.1771-1, 2012.

[39] *Information Technology—Coding of Audio-Visual Objects—Part 12: ISO Base Media File Format, 5th Edition*, ISO/IEC Standard ISO/IEC 14496-12:2015, 2015.

[40] *Information Technology—Dynamic Adaptive Streaming Over HTTP (DASH)—Part 1: Media Presentation Description and Segment Formats, 2nd Edition*, ISO/IEC Standard ISO/IEC 23009-1:2014, 2014.

[41] R. Pantos, *HTTP Live Streaming*, Draft Version 20, IETF, Fremont, CA, USA, Sep. 2016. [Online]. Available: https://tools.ietf.org/html/draft-pantos-http-live-streaming-20

[42] *Information Technology—High Efficiency Coding and Media Delivery in Heterogeneous Environments—Part 3: 3D Audio Amendment 2: MPEG-H 3D Audio File Format*, ISO Standard ISO/IEC 23008-3:2015/Amd.2:2016, 2016.

[43] *Information Technology—Coding of Audio-Visual Objects—Part 15: Carriage of Network Abstraction Layer (NAL) Unit Structured Video in the ISO Base Media File Format*, ISO/IEC Standard ISO/IEC 14496-15:2014, 2014.

[44] *Information Technology—Coding of Audio-Visual Objects—Part 10: Advanced Video Coding, 8th Version*, ISO/IEC Standard ISO/IEC 14496-10:2014, 2014.

[45] *Information Technology—High Efficiency Coding and Media Delivery in Heterogeneous Environments—Part 2: High Efficiency Video Coding, 2nd Edition*, ISO/IEC Standard ISO/IEC 23008-2:2015, 2015.

[46] *Information Technology—High Efficiency Coding and Media Delivery in Heterogeneous Environments—Part 3: 3D Audio Amendment 3: Audio Phase 2*, ISO/IEC Standard 23008-3:2015/Amd3, 2015.

[47] *Information Technology—High Efficiency Coding and Media Delivery in Heterogeneous Environments—Part 1: MPEG Media Transport (MMT*), ISO/IEC Standard 23008-1, 2014.

[48] *ATSC Proposed Standard—Signaling, Delivery, Synchronization, and Error Protection*, Advanced Television Systems Committee Standard A331, 2016.

[49] Y. Xu, S. Xie, H. Chen, L. Yang, and J. Sun, "DASH and MMT and their applications in ATSC 3.0," *ZTE Commun.*, vol. 14, no. 1, pp. 29–49, Feb. 2016. [Online]. Available: http://wwwen.zte.com.cn/endata/magazine/ztecommunications/2016/1/articles/201603/P020160311295798100137.pdf

[50] *Guidelines for Implementation: DASH-IF Interoperability Point for ATSC 3.0, DASH-IF Community Review v0.9*, DASH Ind. Forum, Beaverton, OR, USA, Aug. 2016. [Online]. Available: http://dashif.org/wp-content/uploads/2016/08/DASH-IF-IOP-for-ATSC3.0-v0.90.pdf

[51] K. Park, Y. Lim, and D. Y. Suh, "Delivery of ATSC 3.0 services with MPEG media transport standard considering redistribution in MPEG-2 TS format," *IEEE Trans. Broadcast.*, vol. 62, no. 1, pp. 338–351, Mar. 2016.

[52] H. Stenzel and U. Scuda, "Producing interactive immersive sound for MPEG-H: A field test for sports broadcasting," in *Proc. 137th Conv. Audio Eng. Soc.*, Los Angeles, CA, USA, 2014, p. 9211.

[53] *SMPTE Standard—1.5 Gb/s Signal/Data Serial Interface*, SMPTE Standard ST291-1:2011, 2011.

[54] R. Bleidt. *Installing the MPEG-H Audio Alliance's New Interactive and Immersive TV Audio System in Professional Production and Distribution Facilities*. Accessed on Feb. 12, 2017. [Online]. Available: http://www.mpeghaa.com/papers

[55] *SMPTE Standard—Ancillary Data Packet and Space Formatting*, SMPTE Standard ST291:2006, 2006.

[56] *SMPTE Standard—Format for Non-PCM Audio and Data in an AES3 Serial Digital Audio Interface*, SMPTE Standard ST337:2015, 2015.

[57] *Digital Audio—Interface for Non-Linear PCM Encoded Audio Bitstreams Applying IEC 60958—Part 13: MPEG-H 3D Audio*, IEC Standard IEC 61937-13, 2016.

[58] *A DTV Profile for Uncompressed High Speed Digital Interfaces*, Standard ANSI/CTA-861-G, 2016.

[59] G. K. Walker, T. Stockhammer, G. Mandyam, Y.-K. Wang, and C. Lo, "ROUTE/DASH IP streaming-based system for delivery of broadcast, broadband, and hybrid services," *IEEE Trans. Broadcast.*, vol. 62, no. 1, pp. 328–337, Mar. 2016.

[60] R. Bleidt, A. Borsum, H. Fuchs, and S. M. Weiss, "Object-based audio: Opportunities for improved listening experience and increased listener involvement," *SMPTE Motion Imag. J.*, vol. 124, no. 5, pp. 1–13, Jul./Aug. 2015.

[61] R. Bleidt *et al.*, "Building the world's most complex TV network—A test bed for broadcasting immersive and interactive audio," in *Proc. Annu. Tech. Conf. (SMPTE)*, Hollywood, CA, USA, 2016, pp. 1–10.

[62] *MPEG-H Audio Alliance Live Broadcast Demonstration at NAB 2015 (Video Tour of the Test Bed)*, Fraunhofer IIS, Munich, Germany, Apr. 2015. [Online]. Available: https://youtu.be/wnBx9SjOOII

[63] "Codec for enhanced voice services (EVS); detailed algorithmic description," 3GPP Technical Specification 3GPP TS 26.445, V1.0.0, Release 12, 2014.

[64] *ATSC Proposed Standard—MPEG-H System*, ATSC Standard A/342, 2016.

[65] "A/331: ATSC candidate standard signaling, delivery, synchronization, and error protection," Adv. Television Syst. Committee, Washington, DC, USA, Tech. Rep., 2016.

[66] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–515, Apr. 2000.

[67] "White paper on MPEG-D dynamic range control," MPEG, White Paper ISO/IEC JTC1/SC29/WG11 N15071, 2015.

[68] S. Füg *et al.*, "Design, coding and processing of metadata for object-based interactive audio," in *Proc. Audio Eng. Soc. Conv.*, vol. 137. Los Angeles, CA, USA, Oct. 2014.

Robert L. Bleidt received the B.E.E., M.S.E.E., and M.B.A. degrees from the Georgia Institute of Technology, Atlanta. Since 2007, he has been Division General Manager with Fraunhofer USA Digital Media Technologies, San Jose, CA, USA.

He was the Director of Marketing and New Business Development and the Director of Mass Storage Technology for Sarnoff Real Time Corporation. From 1998 to 2000, he was a Senior Product Manager for Asset Management Products for Philips Digital Networks. From 2000 to 2002, he was the Director of Product Management and Business Strategy for the MP4Net Business of Philips Digital Networks. From 2002 to 2007, he was the President of Streamcrest Associates, a business and market strategy consulting firm. He has authored two SMPTE papers on the MPEG-H TV audio system and numerous industry papers and presentations on MPEG-H, MPEG-4, and related topics.

He managed the development of the Phillips asset management system which received a 2003 technical Emmy Award and created the product concept for the Sonnox Fraunhofer Pro-Codec which received a 2011 TEC Award nomination and the Sound on Sound Editor's Choice award for 2012. He is a Registered Professional Engineer with the state of Georgia, a member of AES and SMPTE, and holds four patents.

Deep Sen (M'92–SM'99) received the B.E. and Ph.D. degrees from the University of New South Wales, Sydney, Australia, in 1990 and 1994, respectively. Since 2011, he has been with Qualcomm Technologies, Inc., San Diego, CA, USA.

He is a Senior Director with the Multimedia Research and Development Laboratory, Qualcomm Technologies and leads the 3-D-Audio research team. From 1994 to 2003, he was with the Speech and Audio Laboratories, AT&T Bell Laboratories, NJ, USA, and from 2003 to 2011, he was a Faculty Member with the School of Electrical Engineering, University of New South Wales, Australia. He has over 60 peer-reviewed publications and multiple granted patents. His past experience includes speech and audio coding, perception, bio-mechanical modeling of the cochlea, blind source separation, beamforming, and hearing prosthetics.

Dr. Sen is a member of the Acoustical Society of America, Audio Engineering Society, and an Elected Member of the IEEE Speech and Language Technical Committee.

Andreas Niedermeier received the degree in mathematics from the Fern-Universität Hagen, Germany, and the Dipl. Math. degree in 2010 with a thesis on wavelet compression technique. In 2011, he joined the Fraunhofer Institute for Integrated Circuits, Erlangen, Germany. He is a Group Manager of the Audio Coding and Multimedia Software Group and focusing on the MPEG-H audio encoder.

Bernd Czelhan received the B.Sc. and M.Sc. degrees in computer science from the Technische Hochschule Nürnberg Georg Simon Ohm, Nuremberg, Germany.

In 2012, he joined the Fraunhofer Institute for Integrated Circuits, as a Research Engineer, where his main researching topic is the next generation audio codec MPEG-H. He has been involved in the standardization process at MPEG and DASH-IF and holds a patent in the area. In addition, he is supporting the practical implementation of MPEG-H including the ATSC 3.0 rollout in South Korea. He is especially interested in modern transport mechanism and system aspects of today's audio codecs, such as MMT, DASH/Route, and hybrid delivery.

Simone Füg received the degree in media technology from the University of Technologies Ilmenau, Germany, and the M.Sc. degree in 2012 with a thesis on controlled binaural distance perception.

She then joined the Fraunhofer Institute for Integrated Circuits, Erlangen, Germany. She is a Senior Engineer with the Department of Semantic Audio Processing. Her current field of interest comprises semantic audio rendering and spatial audio reproduction, focusing on binaural reproduction and object-based audio.

Ms. Füg is active in major international standardization bodies (MPEG audio subgroup and ITU-R.)

Sascha Disch received the Dipl.-Ing. degree in electrical engineering from the Technical University Hamburg-Harburg in 1999. He joined the Fraunhofer Institute for Integrated Circuits (IIS) in 1999.

He received the Doctoral (Dr.-Ing.) degree from Leibniz University Hannover (LUH) in 2011. Since 1999, he has been working in research and development of perceptual audio coding and audio processing. From 2007 to 2010, he was a Researcher with the Laboratory of Information Technology, LUH. He contributed to the standardization of MPEG Surround, MPEG Unified Speech and Audio Coding, and MPEG-H 3-D Audio.

His research interests as a Senior Scientist with Fraunhofer IIS and a member of the International Audio Laboratories Erlangen include waveform and parametric audio coding, audio bandwidth extension, and digital audio effects.

Jürgen Herre (SM'04) received the degree in electrical engineering from Friedrich-Alexander-Universität in 1989 and the Ph.D. degree for his work on error concealment of coded audio. He is a member of the IEEE Technical Committee on Audio and Acoustic Signal Processing, served as an Associate Editor of the IEEE Transactions on Speech and Audio Processing and is an active member of the MPEG audio subgroup.

In 1989, he joined the Fraunhofer Institute for Integrated Circuits (IIS), Erlangen, Germany. In 1995, he joined Bell Laboratories, as a Post-Doctoral Fellow, working on the development of MPEG-2 Advanced Audio Coding. By the end of 1996, he went back to Fraunhofer to work on the development of more advanced multimedia technology including MPEG-4, MPEG-7, and MPEG-D, currently as the Chief Scientist, for the audio/multimedia activities with Fraunhofer IIS, Erlangen. In 2010, he was an Appointed Professor with the University of Erlangen and the International Audio Laboratories Erlangen. He is an expert in low bit-rate audio coding/perceptual audio coding, spatial audio coding, parametric audio object coding, perceptual signal processing, and semantic audio processing.

Dr. Herre is a Fellow Member of the Audio Engineering Society, the Co-Chair of the AES Technical Committee on Coding of Audio Signals, and the Vice Chair of the AES Technical Council.

Johannes Hilpert received the Dipl.-Ing. degree in electrical engineering from the University of Erlangen-Nürnberg, Germany, in 1994.

He joined the Fraunhofer Institute for Integrated Circuits, Erlangen, Germany, where he worked on perceptual audio measurement and MPEG perceptual audio codecs such as MP3 and AAC. In 2000, he headed the team for real-time audio coding algorithms on digital signal processors, and since 2001, he has been in charge of the Audio Coding and Multimedia Software Group. From 2013, he had the position of a Chief Engineer for the Audio and Multimedia Division and has been the Head of the Audio Department, since 2016.

Mr. Hilpert's recent research topics are parametric multichannel, multiobject audio coding, and 3-D audio. He is a Co-Editor of the ISO/MPEG MPEG Surround and SAOC standards.

Max Neuendorf received the Diploma degree in electrical engineering and information technology with the Technical University of Munich in 2002. He is a Group Manager of the Audio and Speech Coding Group, Fraunhofer Institute for Integrated Circuits.

He has been working in the field of modern audio codec development for 14 years and is the Main Editor of the ISO/IEC MPEG-D USAC standard document and the ISO/IEC MPEG-H 3-D audio standard document.

Harald Fuchs received the Diploma degree in electrical engineering from the University of Erlangen, Germany, in 1997. He joined Fraunhofer Institute for Integrated Circuits in 1997.

From 1997 to 2002, he was a Software Developer for video codecs and multimedia streaming systems. From 2002, he concentrated on multimedia systems aspects and standardization, and from 2009 he focused on audio technologies for broadcasting and broadband streaming applications. He actively participated in several standardization organizations, including MPEG, DVB, ATSC, DLNA, and HbbTV, and contributed to several standard documents in those groups. More recently, he focused on MPEG-H audio, and more specifically, on the delivery and transport of MPEG-H audio in ATSC 3.0 systems and MPEG-2 transport stream-based DVB systems. Since 2011, he has been a Senior Business Development Manager, Audio for TV Broadcast, and since 2013, he has been a Group Manager of the Semantic Audio Coding Group.

Jochen Issing received the Dipl.-Ing. degree in electrical engineering from the University of Applied Sciences Amberg-Weiden in 2002. He joined the Fraunhofer Institute for Integrated Circuits (IIS) in 2002.

From 2009 to 2013, he was working in research and development with the Friedrich Alexander University Erlangen, and in 2013, he joined Skype as a Software Engineer. In 2015, he returned to Fraunhofer IIS.

Adrian Murtaza received the Diploma degree in electrical engineering electronics and information technology from the Politehnica University of Bucharest, Romania, in 2010, and the 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. In 2012, he joined Fraunhofer Institute for Integrated Circuits, where he was a Researcher on semantic audio coding, parametric multiobject and multichannel audio coding, and 3-D audio.

He actively participated in several standardization organizations, including MPEG, DVB, ATSC, and SCTE, and contributed to several standards documents in those groups. His recent activity is focused on MPEG-H audio, and more specifically, on the signaling and transport of MPEG-H audio.

Achim Kuntz received the Dipl.-Ing. degree in electrical engineering from the University of Erlangen-Nürnberg in 2002, and the Doctoral degree from the Telecommunications Laboratory, University of Erlangen-Nürnberg, researching on spatial sound reproduction, multidimensional signals and wave field analysis, in 2008.

He is a Chief Scientist with the Fraunhofer Institute for Integrated Circuits, Erlangen, Germany, and a member of the International Audio Laboratories Erlangen. His current field of interests comprises perceptually motivated signal processing algorithms for audio applications including the acquisition, manipulation, coding and reproduction of spatial sound.

Dr. Kuntz is active in the standardization of audio codecs within ISO/MPEG.

Michael Kratschmer received the Dipl.-Ing. degree in electrical engineering, electronics and information technology from the University of Erlangen-Nuernberg, Germany, in 2009. In 2010, he joined the Fraunhofer Institute for Integrated Circuits as a Research Engineer in the area of speech enhancement applications. Since 2015, he has been a Senior Engineer, where his current field of interest comprises audio processing and signal manipulation based on low bit rate metadata coding and transmission throughout the complete content production and reproduction chain. He is active in the standardization of audio codecs within the MPEG audio subgroup with a focus on dynamic range and loudness control.

Fabian Küch received the Dr.-Ing. degree in electrical engineering from the Friedrich-Alexander-University Erlangen-Nuremberg, Germany, in 2005. He was a member of the audio group with the Chair of Multimedia Communications and Signal Processing. Since 2006, he has been with the Audio Department, Fraunhofer Institute for Integrated Circuits, where he holds the position of a Group Manager. His responsibilities include research and development in the area of spatial audio processing as well as loudness and dynamic range control. He has contributed to the standardization of MPEG-D Dynamic Range Control and MPEG-H 3D audio.

Richard Füg studied the degree in Informations- und Kommunikationstechnik (information and communication technology) from the Friedrich-Alexander-Universität Erlangen-Nürnberg, the B.Sc. degree in 2012, and the M.Sc. degree in 2014.

In 2014, he joined the Fraunhofer Institute for Integrated Circuits as a Research Engineer with the High Quality Audio Coding Group, where he is involved in the MPEG-H encoder development. His research interests include digital audio and music signal processing with emphasis on audio coding, source separation, and audio effects.

Benjamin Schubert received the Dipl.-Ing. degree in media technology from the University of Technology Ilmenau, Germany, with a thesis on parametric audio coding in 2010. He joined the Fraunhofer Institute for Integrated Circuits in 2011. He is a Senior Engineer with the High Quality Audio Coding Group with a focus on low bitrate audio and speech coding.

Sascha Dick received the Dipl.-Ing. degree in information and communication technologies from the University of Erlangen-Nuremberg, Germany, in 2011 with a thesis on an improved psychoacoustic model for spatial audio coding. He joined Fraunhofer Institute for Integrated Circuits as a Research Engineer in 2011. He has contributed to the development and standardization of audio codecs such as MPEG-H 3D audio. His current research interests include psychoacoustics, multichannel signal processing, and audio coding.

Guillaume Fuchs received the engineering degree from the INSA of Rennes, France, in 2001, and the Ph.D. degree from the University of Sherbrooke, Canada, in 2006, both in electrical engineering. He is a Senior Scientist with Fraunhofer Institute for Integrated Circuits (IIS). From 2001 to 2003, he worked on digital image compression as a Research Engineer with Canon Research, France. From 2003 to 2006, he worked on speech coding with VoiceAge, Canada, as a Scientific Collaborator. In 2006, he joined Fraunhofer IIS, and since then, has been working on developing speech and audio coders for different standards. His main research interests include speech and audio source coding and speech analysis.

Florian Schuh received the Dipl.-Ing. degree in electrical engineering, electronics and information technology from the University of Erlangen-Nuremberg, Germany, in 2012 with a thesis on multichannel acoustic echo cancellation using graphics cards. He joined the Fraunhofer Institute for Integrated Circuits as a Research Engineer in 2012. He is mainly enrolled in the audio core codec development of MPEG-H, extended HE-AAC, and HE-AAC. His current field of interest comprises multichannel audio processing.

Elena Burdiel received the bachelor's degree in sound and image engineering from the Technical University of Madrid in 2014 with the thesis entitled *Measurement of the Acoustic Properties of Hall 1 From Kinepolis Movie Theatre*.

She joined the Audio and Speech Coding Group, Fraunhofer Institute for Integrated Circuits, Erlangen, Germany, in 2014, where she is currently with the Audio Coding and Multimedia Software Group.

Nils Peters received the M.Sc. degree in electrical and audio engineering from the University of Technology, Graz, Austria, and the Ph.D. degree in music technology from McGill University, Montreal. He is a Senior Staff Research Engineer and a Manager with the Multimedia Research and Development Laboratory, Qualcomm Technologies, Inc. He was a Post-Doctoral Fellow with the International Computer Science Institute, Center for New Music and Audio Technologies, and the Parallel Computing Laboratory, University of California, Berkeley.

He has been a Researcher in the field of spatial audio for over a decade and researches on technical and perceptual aspects of spatial audio, including soundfield analysis and compression, acoustic sensing, spatial sound reproduction, auditory perception, and sound quality evaluation. He has published over 40 peer-reviewed papers. He was an Audio Engineer researches in the fields of recording and postproduction.

Dr. Peters serves as the Co-Chair of the Technical Committee for spatial audio at the Audio Engineering Society.

Moo-Young Kim received the M.Sc. degree in electrical engineering from Yonsei University, Seoul, South Korea, in 1995, and the Ph.D. degree in electrical engineering from KTH (the Royal Institute of Technology), Stockholm, Sweden, in 2004. He is a Principal Engineer with the Multimedia Research and Development, Qualcomm Technologies, Inc.

From 1995 to 2000, he was a Research Staff Member with the Human Computer Interaction Laboratory, Samsung Advanced Institute of Technology. From 2005 to 2006, he was a Senior Research Engineer with the Department of Multimedia Technologies, Ericsson Research. From 2006 to 2014, he was an Associate Professor with the Department of Information and Communication Engineering, Sejong University. His research interests include speech/audio coding, information theory, biometrics, speech enhancement, joint source and channel coding, and music information retrieval.