ITUPublications

International Telecommunication Union
Radiocommunication Sector

Report ITU-R BS.2493-1

(11/2024)

BS Series: Broadcasting service (sound)

Practical implementation of broadcast systems using audio codecs for ITU advanced sound systems

Foreword

The role of the Radiocommunication Sector is to ensure the rational, equitable, efficient and economical use of the radio-frequency spectrum by all radiocommunication services, including satellite services, and carry out studies without limit of frequency range on the basis of which Recommendations are adopted.

The regulatory and policy functions of the Radiocommunication Sector are performed by World and Regional Radiocommunication Conferences and Radiocommunication Assemblies supported by Study Groups.

Policy on Intellectual Property Right (IPR)

ITU-R policy on IPR is described in the Common Patent Policy for ITU-T/ITU-R/ISO/IEC referenced in Resolution ITU-R 1. Forms to be used for the submission of patent statements and licensing declarations by patent holders are available from https://www.itu.int/ITU-R/go/patents/en where the Guidelines for Implementation of the Common Patent Policy for ITU-T/ITU-R/ISO/IEC and the ITU-R patent information database can also be found.

Series of ITU-R Reports

(Also available online at https://www.itu.int/publ/R-REP/en)

Series	Title
BO	Satellite delivery
BR	Recording for production, archival and play-out; film for television
BS	**Broadcasting service (sound)**
BT	Broadcasting service (television)
F	Fixed service
M	Mobile, radiodetermination, amateur and related satellite services
P	Radiowave propagation
RA	Radio astronomy
RS	Remote sensing systems
S	Fixed-satellite service
SA	Space applications and meteorology
SF	Frequency sharing and coordination between fixed-satellite and fixed service systems
SM	Spectrum management
TF	Time signals and frequency standards emissions

***Note**: This ITU-R Report was approved in English by the Study Group under the procedure detailed in Resolution ITU-R 1.*

Electronic Publication
Geneva, 2025

REPORT ITU-R BS.2493-1

Practical implementation of broadcast systems using audio codecs for ITU advanced sound systems

(Question ITU-R 19-1/6)

(2021-2024)

This Report describes how new broadcast industry requirements can be addressed with Next Generation/Advanced sound systems. It is focussed on the system level approach to addressing these needs in practical broadcast systems, rather than on the low-level technical detail.

The Annexes provide guidance to broadcasters on the system specifications for encoding bitstreams for transmission.

TABLE OF CONTENTS

Page

Annex 1

AC-4 digital audio compression standard

Introduction

This Annex provides guidelines for broadcasters on the coding of a bitstream of audio information using the AC-4 digital audio compression standard, which relates to the decoding in broadcast receivers in a harmonised process. The coded representation specified herein is suitable for use in digital audio transmissions.

Building on specified technology, this Annex provides a harmonised representation of functionality that may be used for presentation to listeners of the digital audio transmissions.

AC-4 is a digital audio compression standard, openly published as ETSI TS 103 190-1 [1] and ETSI TS 103 190-2 [2]. AC-4 is adopted for use in Digital Video Broadcasting (DVB) systems as defined in ETSI TS 101 154 [3] and in ATSC 3.0 Systems as defined in A/342 Parts 1 [4] and 2 [5]. AC-4 has also been adopted as one of the NGA systems included in the Society of Cable and Television Engineers (SCTE) [6][7][8][9].

In addition to inclusion in the DASH-based ATSC3.0 specifications, AC-4 is also supported for HbbTV streaming and interactive applications per version 2.0.2 of the HbbTV specification – ETSI TS 102 796 [10].

AC-4 has been included as the sole next generation audio codec in several regional broadcast specifications, including North America ATSC3.0, NorDig v3.1.1 (unified specification for DVB-T2 covering Denmark, Norway, Finland, Iceland, Sweden and Ireland) [13], Italy UHD-Book v2.0 [14], and Poland DVB-T2 [15]. It is also supported in the ISDB-T specification for terrestrial television in Brazil. AC-4 broadcasts are on the air for ATSC 3.0 stations in the United States of America, and broadcast trials have been completed in several other countries including Spain, France, Italy and Poland.

This Annex does not specify an object audio renderer, which would be needed to decode object-based audio to a channel-based representation. Audio renderers are specified in ETSI TS 103 448 [11] that accompanies the AC-4 specification defined in ETSI TS 103 190 and Recommendation ITU-R BS.2127 – Audio Definition Model renderer for advanced sound systems.

1 Immersive audio

1.1 Channel-based approach

A broadcaster may wish to preserve current stereo and 5.1 audio configuration within the broadcast plant that uses the first eight channels. The third audio programme would be the immersive version. To keep all immersive audio tracks in sync within the plant’s infrastructure, it would be best practice to keep all audio elements in a single quadrant.

1.2 Object-based approach

1.2.1 Encoding and decoding (core and full) scenarios for immersive audio

This section highlights the availability of two decoding modes – core decoding and full decoding – by giving examples of encoding and decoding scenarios for channel-based and spatial object groups-based immersive audio.

The core decoding mode enables a lower-complexity, reduced-channel-count or reduced object-count output from the decoder, while the full decoding mode does the full decoding of the stream. These two decoding modes are made possible by the structure of the bitstream and the coding methods employed. They are an essential feature of AC-4, allowing a stream to play on lower-cost, lower-complexity devices, typically outputting stereo or 5.1, as well as on full-capability devices with high-channel-count output or other needs for full fidelity.

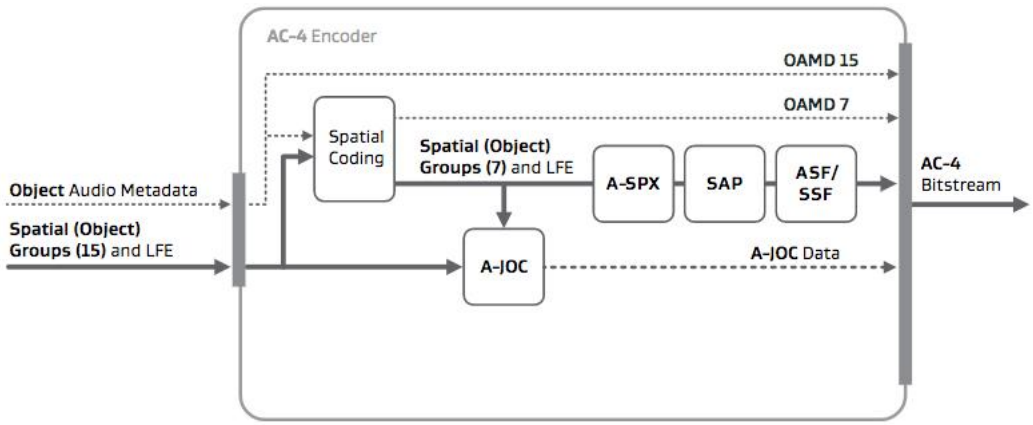
1.2.2 Object-based (spatial object groups) immersive

In the case of object-based audio using spatial object groups, the input to the encoder consists of spatial object groups (15 of which are in the example shown in Fig. 1) and the LFE channel, as well as their corresponding object audio metadata (OAMD). An Advanced Joint Object Coding (A-JOC) module on the encoder side is used to provide the A-JOC data to the bitstream.

In the example shown in Fig. 1, the spatial coding module further reduces the spatial object groups to seven while creating associated OAMD. The seven spatial groups are then encoded using the core encoder modules.

FIGURE 1

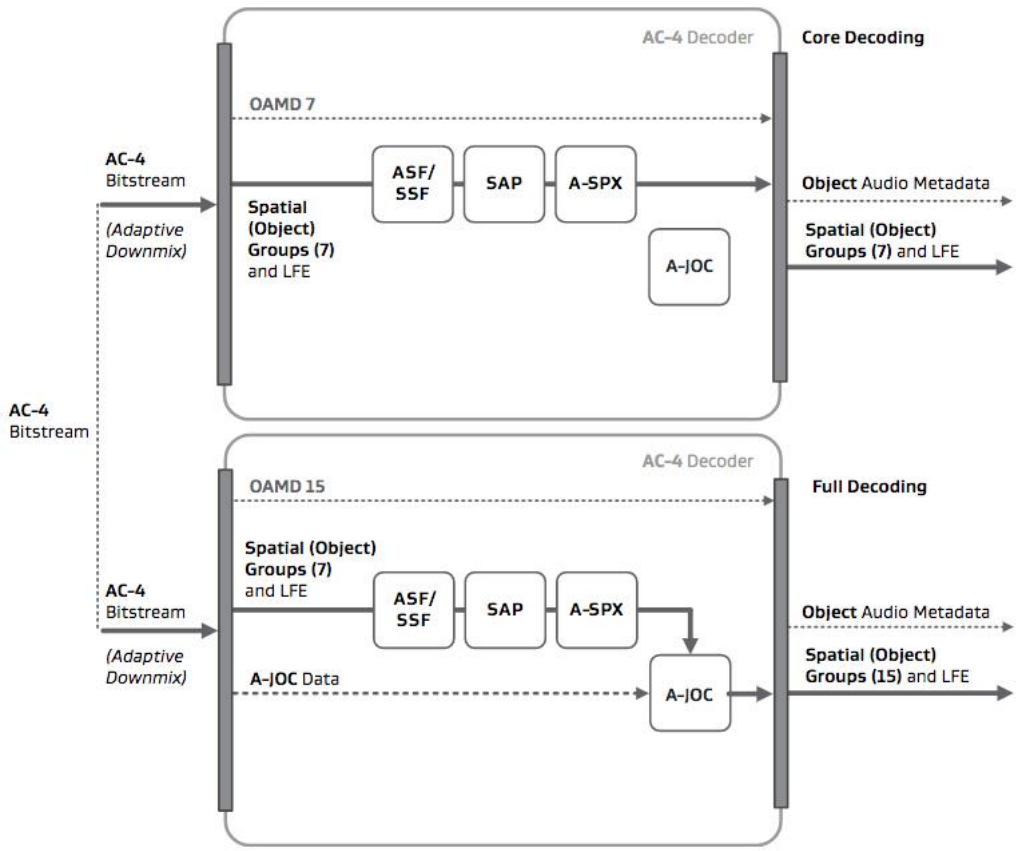
Encoding object-based audio (adaptive downmix)



As outlined in Fig. 2, the same bitstream can then be decoded by a playback device running in core decoding or full decoding. The difference is that for full decoding the A-JOC decoding module is used, resulting in 15 spatial object groups being output by the decoder. In both decoder modes there is a need for a renderer.

FIGURE 2

A-JOC core and full decoding



1.2.3 Channel-based immersive

For channel-based immersive audio (which in the example below is 7.1.4), different tools are used depending on the bit rate.

These tools, that code the spatial properties of the audio signal, aim to reduce the number of waveforms to be coded by the subsequent tools (A-SPX, SAP and ASF/SSF), and in doing so create a parametric representation. When A-CPL is used, 'side signals' are also created; these are conceptually similar to side signals in traditional mid-side coding.

– Advanced joint channel coding (A-JCC) is used for the lowest possible bit rates; a core of 5.1 channels is coded.

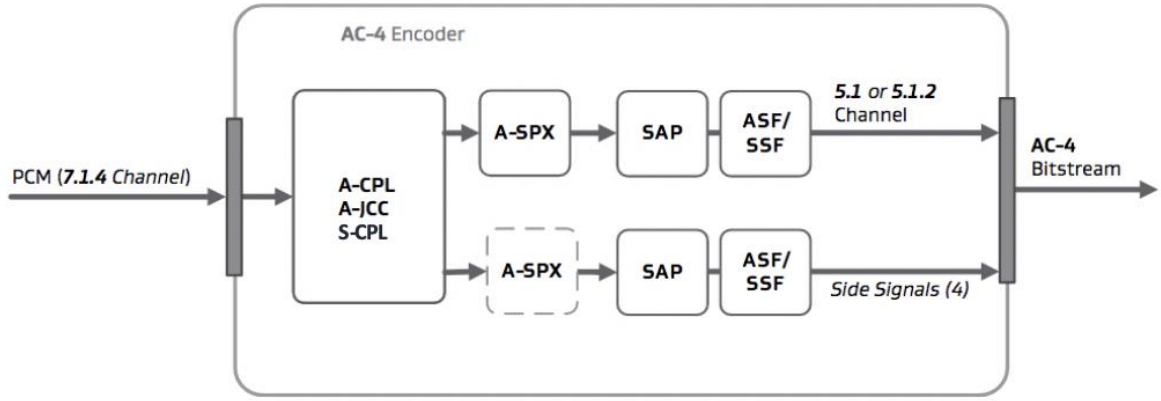
– Advanced coupling (A-CPL) is used for intermediate to high rates; a core of 5.1.2 audio channels is coded, and the audio side signal can optionally be coded to further increase audio quality.

– Simple coupling (S-CPL) is associated with highest bit rates. It uses fullband side signals to provide highest audio quality up to transparency.

Figure 3 illustrates the channel-based immersive encoder.

FIGURE 3

AC-4 encoder 7.1.4 immersive channel



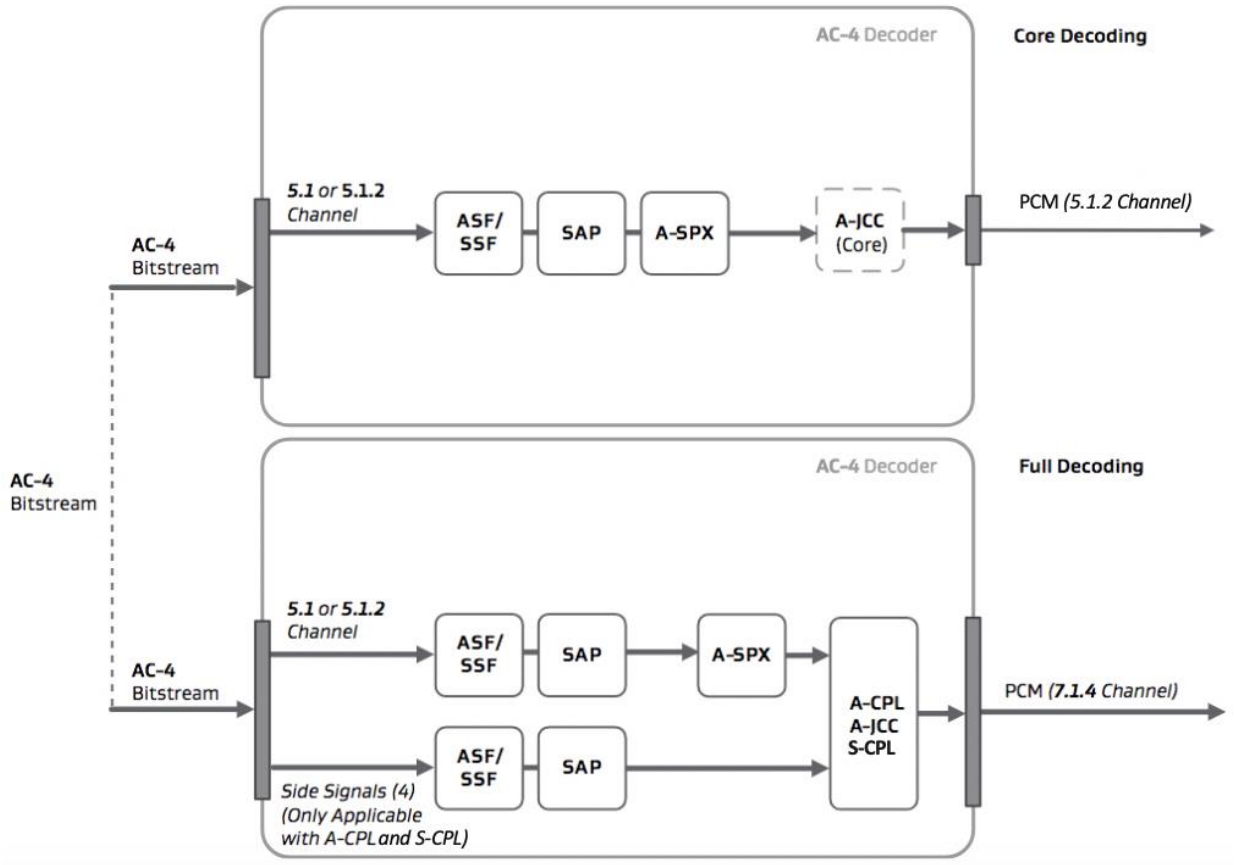
If the decoder is configured to do core decoding, the 5.1 or 5.1.2 waveform-coded channels are decoded by the core decoding tools – Audio Spectral Frontend (ASF), SAP, Advanced Spectral Extension (A-SPX) and, optionally, the core part of A-JCC, depending on coding configuration.

If the decoder is configured to do full decoding, the 5.1 or 5.1.2 waveform-coded channels, along with optional side signals, are decoded by the core decoding tools ASF, SAP, A-SPX, and the A-CPL or A-JCC, depending on coding configuration.

Figure 4 illustrates the differences in the pulse-code modulation (PCM) output decoding scenario when doing core decoding and full decoding.

FIGURE 4

PCM output channel decoding



2 Audio personalization

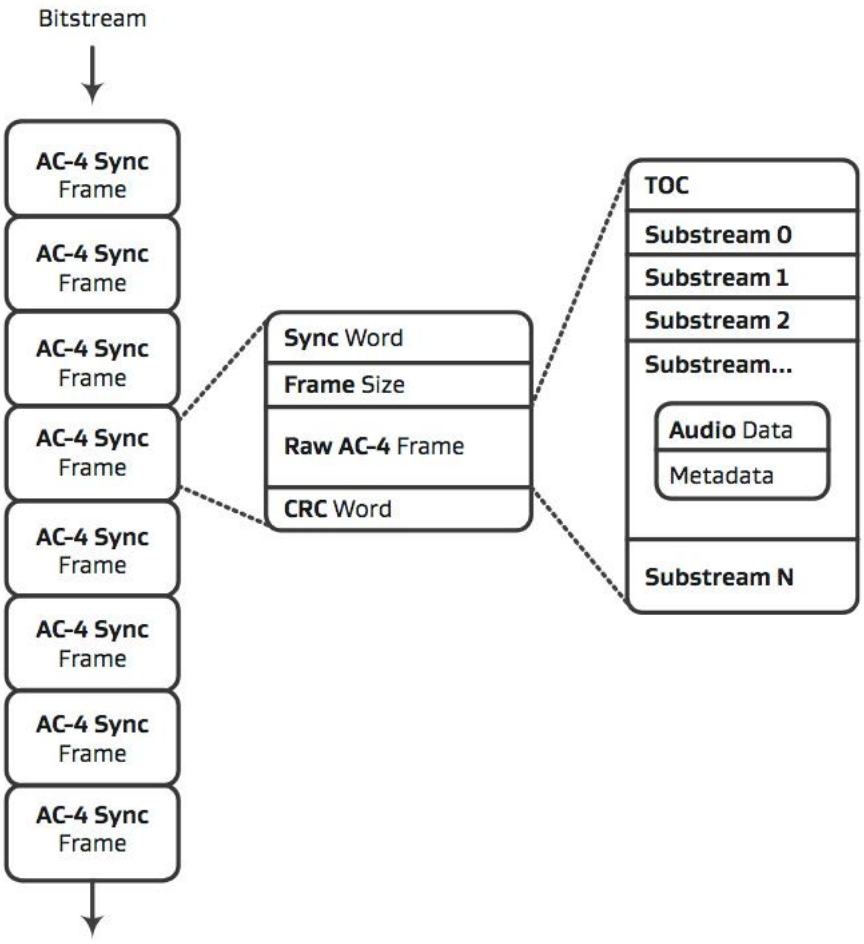
2.1 Overview of presentation-based approach

An AC-4 elementary stream consists of synchronization frames, each beginning with a sync word and optionally ending with a cyclic redundancy check (CRC) word. The sync word allows a decoder to easily identify frame boundaries and begin decoding. The CRC word allows a decoder to detect the occurrence of bitstream errors and perform error concealment when it detects an error.

The data carried within each synchronization frame is referred to as the raw AC-4 frame. Each raw frame contains a Table of Contents (TOC) and at least one substream containing audio and related metadata. Figure 5 shows the high-level bitstream structure.

FIGURE 5

High-level bitstream syntax for AC-4 (Part 2)



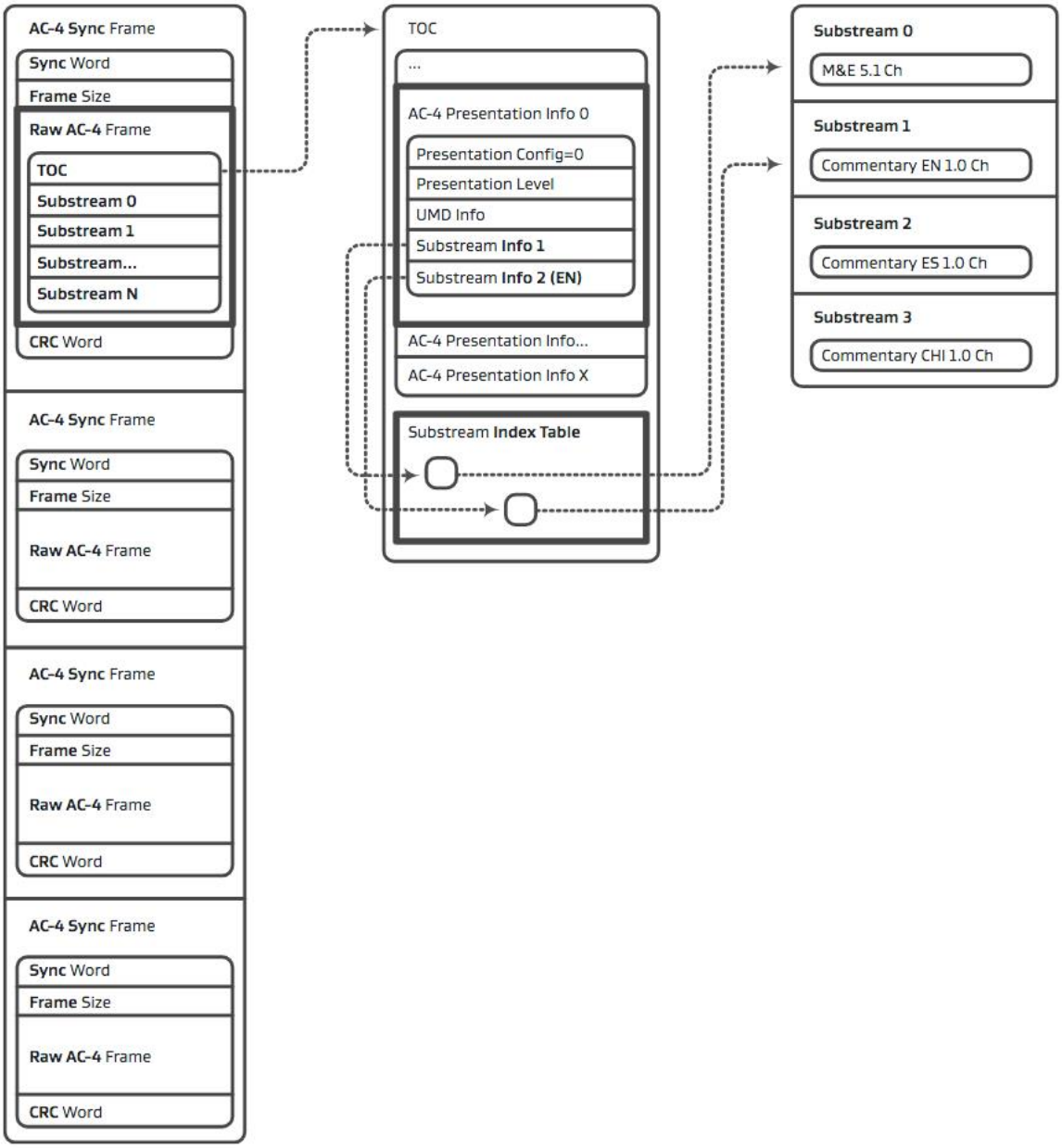
The TOC contains the inventory of the bitstream. Each audio substream can carry either one or more audio channels or an individual audio object. This structure provides flexibility and extensibility that allows the AC-4 format to meet future requirements.

AC-4 also allows multiple Presentations to be carried in a single bitstream. Each Presentation defines a way of mixing a set of audio substreams to create a unique rendering of the programme. Instructions for which substreams to use and how to combine them for each Presentation are specified in a Presentation info element carried in the TOC.

Presentations enable multiple versions of the audio experience, such as different languages or commentary, to be delivered in a single bitstream in a convenient, bandwidth-efficient manner. An example is shown in Fig. 6, where four versions of a live 5.1 sports broadcast – the original English version, two alternate languages (Spanish and mandarin Chinese), and a commentary-free version – are combined into a single AC-4 bitstream.

FIGURE 6

Live 5.1 sports broadcast with four presentations



Rather than transmitting four separate 5.1-channel streams, as would be done with current technologies, the AC-4 bitstream contains four substreams:

– Music/effects mix without commentary (5.1 channels);
– English commentary (mono);
– Spanish commentary (mono);
– Mandarin Chinese commentary (mono).

The TOC contains four Presentation info elements, one for each playback experience. Figure 6 above shows the English Presentation (Presentation 0) selected, which instructs the decoder to combine the music/effects and English commentary substreams to create the English output. Similarly, the Spanish and Mandarin Chinese Presentation info elements instruct the decoder to render the common music/effects substream with the relevant language commentary, while the commentary-free Presentation info element renders only the music/effects substream. All necessary mixing is done in the decoder, eliminating the need to implement this with surrounding system components.

Table 1 demonstrates how this Presentation-based approach can offer data-rate savings in excess of 50% compared with using the same format to deliver multiple 5.1 streams.

TABLE 1

Conventional approach		**Presentation approach**	
Stream	**Data rate (kbit/s)**	**Stream**	**Data rate (kbit/s)**
English 5.1	144	English commentary 1.0	40
Spanish 5.1	144	Spanish commentary 1.0	40
Chinese 5.1	144	Chinese commentary 1.0	40
Commentary-free 5.1	144	Commentary-free 5.1	144
Total	576	Total	264
		Savings	54%

Comparing the Presentation approach of AC-4 with delivery of multiple 5.1 streams using AC-3 (as used by many HDTV services today), the data-rate savings exceed 80%.

Use of Presentations also provides a way to deliver optimal audio experiences to devices with very different capabilities using a single audio bitstream. For example, decoders in some devices might be able to decode only up to 5.1 channels, where the additional channels are unnecessary for that device and would impose an unrealistic processing load.

A service provider wishing to offer, say, a 7.1.4-channel service could ensure compatibility with both simple and advanced devices by incorporating a 7.1.4-channel and a 5.1-channel Presentation within a single AC-4 bitstream. The formal interoperability test programme for AC-4 ensures that decoders conform to defined functional levels so that advanced services can easily be configured to suit the target device base.

Table 2 provides typical AC-4 bitrate ranges for legacy configurations, stereo and 5.1, as well as for an NGA immersive configuration with addition audio object for alternate languages or commentary.

TABLE 2

Configuration	**Typical ranges (kbit/s)**
0+2+0 (Stereo)	48-64
0+5+0 (5.0/5.1)	96-144
4+7+0 (7.1.4) plus four objects for alternative languages or commentary	320-512

Two mechanisms can be considered for selection of a presentation for playback.

In the first, a user defined preference setting drives selection of the most appropriate version. This may be desirable, for example, with language settings, where a default preferred language is typically chosen by the user on installation. This may be used to automatically select a presentation in that language from the list of presentations in the AC-4 table of contents.

In the second, the broadcaster may wish to offer the user a choice of audio versions – for example, sports games with biased 'home' and 'away' commentaries. In that case, it may be desirable to provide a user interface for selection of options while watching the programme. For example, the DVB signalling specification ETSI EN 300 468 [12] describes the Audio Preselection Descriptor (APD), which can be used to flag the available AC-4 presentations at an UI level. The data fields for each presentation included in this descriptor cover all the properties addressed by the typical preference settings and can be enhanced by a textual description which is conveyed in a separate descriptor for efficiency reasons. Based on this signalling, the user interface can inform the AC-4 decoder about the user's presentation selection through a unique identifier. If the signalled choice is temporarily unavailable – for instance during an advertisement break – the AC-4 decoder will instantaneously fall back on the user-preferences based presentation selection to provide a seamless experience to the listener and will revert the personalized choice once it becomes available again.

In applications where the available presentations are intended to be signalled at an EPG level, the DVB signalling specification uses the Component Descriptor. A first instance of this descriptor can be used to identify the used audio codec, while an additional instance can be used to signal the properties of each presentation. These cover the same parameters as with the Audio Preselection descriptor including the optional text string for an enhanced experience description.

For the use in the American territory, SCTE has adopted the DVB-defined descriptors [8], while for MPEG-DASH based emissions (as used by ATSC 3.0), the preselection element is functionally equivalent.

For HbbTV based systems, presentation selection is mapped to the HTML AudioTrack control, enabling presentation selection to be performed within a broadcaster app. This enables a list of available presentations to be populated into an app and enables the app to perform selection of the desired presentation.

User interface requirements are outside the scope of most regional broadcast receiver specifications. However, several DVB-based specifications do refer to the use of elementary stream level signalling for controls that will be driven by user preset preferences, and to the use of APD for presentations that will be manually selected.

2.2 Multi language application

AC-4 provides native support for multi-lingual applications as described above. To achieve high levels of bandwidth efficiency in these applications, AC-4 also contains a dedicated Speech Spectral Frontend (SSF). This prediction-based speech coding tool achieves very low data rates for speech content. Unlike most common speech coders, it operates in the MDCT domain, which enables seamless switching between the ASF and the SSF as the characteristics of the content change. The SSF is especially important for efficiently delivering multilingual and secondary commentary content where many independent dialogue substreams are encoded and carried in a single AC-4 stream.

2.3 Audio description application

AC-4 provides native support for description services in a similar manner to multiple language support. This can be used to selectively mix a separate descriptive commentary track over the main audio in the receiver where desired. This is sometimes referred to as 'receiver-mixed' audio description. An advantage of this receiver mixed approach is that any features of the main audio track

such as immersive audio or choice of experience can be maintained for the descriptive commentary version in a bandwidth-efficient manner.

AC-4 can also be used to delivery pre-mixed soundtracks that incorporate descriptive commentary, sometimes referred to as ‘broadcast-mixed’ audio description.

2.4 Dialogue enhancement application

2.4.1 Basic modes for legacy content

In practical tests, viewers have been found to have widely differing preferences for dialogue or commentary levels. AC-4 addresses this challenge by providing mechanisms for viewers to tailor the dialogue level to suit their individual preferences. These flexible mechanisms work with both legacy content that contains dialogue already mixed into the main audio and new content where a separate dialogue track is available to the AC-4 encoder.

Conventional dialogue enhancement techniques in products such as TVs and tablets have applied single-ended processing in the playback device to attempt to detect and adjust dialogue elements of the mixed audio. While these techniques have the advantage that they do not require specially produced content, they have the disadvantage of not being wholly predictable, and their effectiveness is limited by the processing power available in the playback device.

With AC-4, dialogue enhancement is instead implemented by utilizing the dramatically higher processing power of the audio encoder to analyse the audio stream and generate a highly reliable parametric description of the dialogue, whether or not a separate dialogue track is available. These parameters are sent with the audio in the AC-4 stream and used by the playback device to adjust the dialogue level under user control.

FIGURE 7

Dialogue enhancement for channel-based content (encoder functionality)

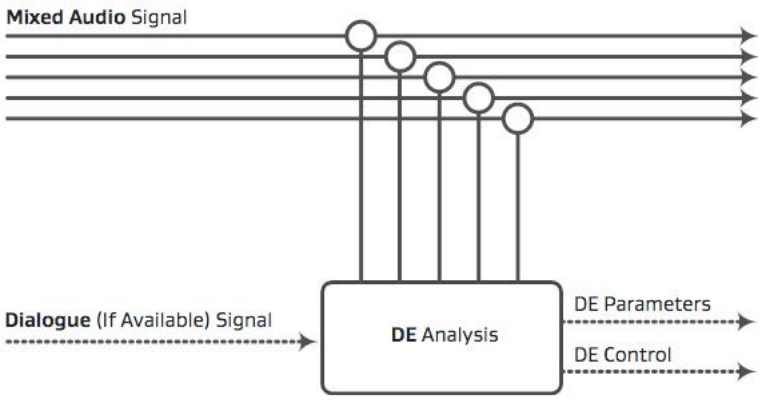
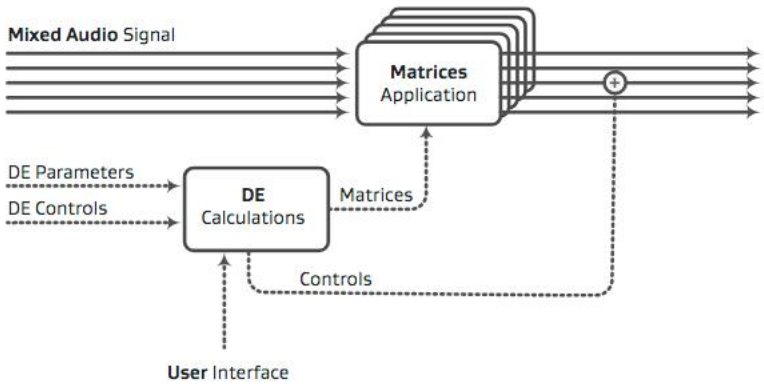


FIGURE 8

Dialogue enhancement for channel-based content (decoder functionality)



For legacy content, this can be performed with solely the mixed content in 'Independent' mode.

If the dialogue is available as a separate audio track, the encoder creates the parameters based on the joint analysis of the mixed audio signal and the separate dialogue signal. This is referred to as 'Guided' mode. These parameters are more precise than those extracted from the mixed audio signals as described previously and allow more precise dialogue adjustments in the decoder. In this mode, still only the mix and the control parameters are transmitted – good results can be achieved without having to send separate music/effects and dialogue signals.

2.4.2 Separate dialogue mode

Alternatively, if desired (for example, to perform language substitution), the dialogue and music/effects tracks can be sent in the AC-4 bitstream as separate substreams for optimum performance and maximum flexibility.

The combination of guided and automatic modes in the AC-4 system means that dialogue enhancement functionality may be implemented by the service provider in a broad, predictable, and effective manner. The automatic analysis method (at the encoder) provides a solution that is easy to deploy with legacy content and workflows as the industry transitions to interchange of separate dialogue tracks in the long term.

3 Loudness management

3.1 Intelligent loudness management in cascaded workflows

Over the last decade, approaches to managing the loudness of broadcast and streaming services have changed considerably. The broadcast industry, for example, has made significant steps towards using a long-term loudness measure, rather than just a short-term peak measure, to align the levels of programming and provide a more consistent and pleasant experience for viewers. This has resulted in Recommendations from the International Telecommunication Union (ITU) on international programme interchange levels, from the European Broadcasting Union (EBU) on broadcast levels, and from several other national and international groups on local requirements.

However, the need to achieve loudness consistency and meet regional loudness mandates has often led to the introduction of loudness processing at multiple points in the chain, for example, in content creation, in the broadcast station, and at the operator. In many cases, this redundant processing results in compromised sound quality.

To help services ensure loudness consistency and compliance with regulations, the AC-4 encoder incorporates integrated intelligent loudness management. The encoder assesses the loudness of

incoming audio and can, if desired, update the loudness metadata (dialnorm) to the correct value or signal the multiband processing required to bring the programme to the target loudness level. Rather than processing the audio in the encoder, this information is added to the bitstream in the DRC metadata so processing can be applied downstream in the consumer device appropriately for the playback scenario. The process is therefore non-destructive; the original audio is carried in the bitstream and available for future applications.

To avoid the problems associated with cascaded levelling processes, AC-4 makes use of the extended metadata framework standardized in ETSI TS 102 366 [16], Annex H. This framework carries information about the loudness processing history of the content so that downstream devices can intelligently disable or adjust their processing accordingly, maximizing quality while maintaining consistency. Annex H metadata can be carried throughout the programme chain, either with the baseband audio prior to final encoding or inserted into transmission bitstreams including AC-4.

If the incoming audio presented to the AC-4 encoder has previously been produced or adjusted to a target loudness level by a trusted device, this can be signalled to the encoder using Annex H metadata. In this case, the integrated loudness levelling processing of the encoder will be automatically disabled, so that the audio is delivered without further adjustment, maximizing quality and preserving the original creative decisions.

Because the Annex H metadata in the AC-4 bitstream also indicates any loudness processing that has been applied, this can be used to automatically disable unnecessary loudness processing that might be in place downstream - for example, in cable or Internet Protocol Television (IPTV) headends. The extended Annex H metadata also includes additional loudness measures such as short-term loudness, which can assist compliance in regions that regulate based on these characteristics.

4 Dynamic range management

4.1 Managing dynamic range for different device types

4.1.1 Dynamic range control (DRC)

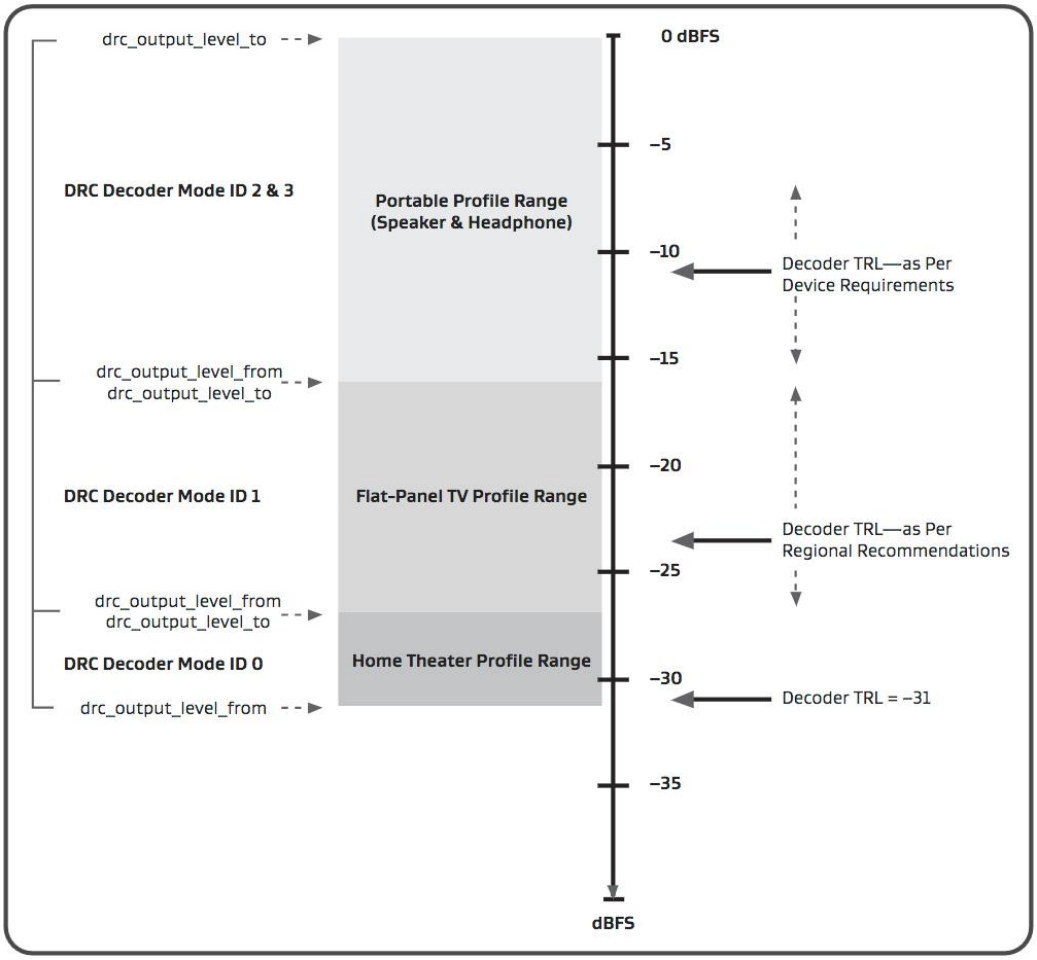
The AC-4 decoder applies DRC to tailor the dynamic range and the typical output level to suit the listening scenario. Implementing DRC in the QMF domain enables powerful multiband and multichannel processing which improves quality over previous wideband approaches.

AC-4 supports a number of DRC modes to adapt the content to different listening environments and playback scenarios. Each mode is associated with a type of playback device and has guidelines for decoder-defined playback reference levels.

Four standard DRC decoder modes have been defined, each with a corresponding output level range. In addition to the standard modes, it is possible to add up to four user definable modes to support future or proprietary device types. Figure 9 shows the four decoder modes, in relation to output level, indicating the output device type.

FIGURE 9

Playback device target reference level and DRC profile mapping



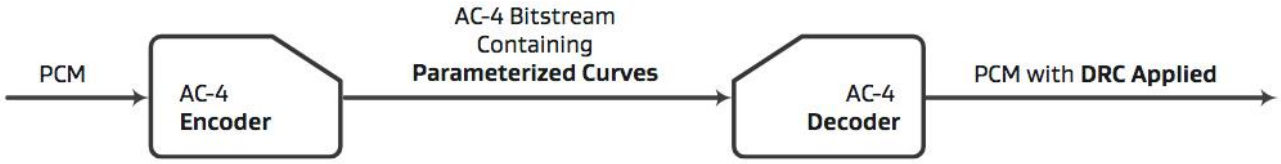
An AC-4 decoder may be provisioned to align with a number of device categories and applications along with how the Target Reference Level (TRL) parameter maps to the four decoder dynamic range control modes as shown in Fig. 9. For example, if −23 dBFS is selected for the target reference level (TRL), the AC-4 DRC Decoder Mode ID (representing the parameterized compression curve for that playback mode) overlapping with −23 will be used to generate the DRC information in the decoder, which in this example is the Flat Panel TV profile. For Home Theatre AVR use cases, the decoder user/system would provision the decoder target reference loudness to −31.

Selecting −31 TRL will place the DRC Profile selection into the DRC Decoder Mode ID 0 range, and the system will use the Home Theatre Profile to generate the DRC info in the decoder. Portable/Mobile devices have flexibility to set the TRL value to best align with their internal gain structure needs.

DRC parameters for each output mode are generated by the AC-4 encoder (see Fig. 10) or by an external third-party processor (see Fig. 11) and transported in the AC-4 bitstream as DRC gain values (wideband or multiband). Alternatively, the desired DRC characteristic can be expressed in the bitstream as a parameterized compression curve in a bandwidth efficient manner.

FIGURE 10

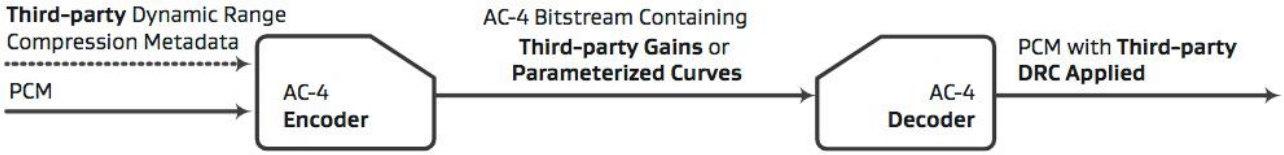
DRC parametrized curves



This curve can be created by the service provider or content creator to suit the content and their house style. These curves may be selected from a number of presets in the encoder or may be customized if desired. Parameterized compression curves provide benefits such as lower bit-rate overhead and higher audio quality for traditional channel audio content, with even larger gains for OBA and immersive audio content, where a DRC gain per channel or object, as used in other systems, becomes costly.

FIGURE 11

DRC third-party gains



The AC-4 decoder calculates gains based on the compression curve transmitted for the selected mode and the target playback reference level of the device. The target playback reference level for each mode is not fixed but instead can be defined in the decoder. This enables flexibility to match the loudness of other content sources depending on the listening scenario.

4.1.2 Simultaneous outputs

To enable an AC-4 decoder to support multiple outputs - for example, to feed several different devices simultaneously – the decoding process is split into two stages: the input stage and the output stage.

The input stage decodes the selected channels and objects, provides the configuration of the current frame, and performs error detection.

Each output stage adapts the audio to the particular playback scenario:

– Output mapping, for example, stereo, multichannel, or objects;

– User preferences, for example, the selection of dialogue enhancement or the main and/or associated audio;

– Environmental conditions, for example, late-night mode;

– Other device-specific playback optimizations.

An arbitrary number of output stages can be connected to a single input stage, which allows rendering different output configurations from one input bitstream simultaneously. Each output stage can be freely configured – for example, one for headphone output at a high output level and another for 5.1-channel mode at line level. This saves computational load compared with running an input decode stage for each output.

Figures 12 and 13 show examples of a decoder with a single output and a decoder with three independent outputs, each of which is optimized for a different kind of device.

FIGURE 12

Single output

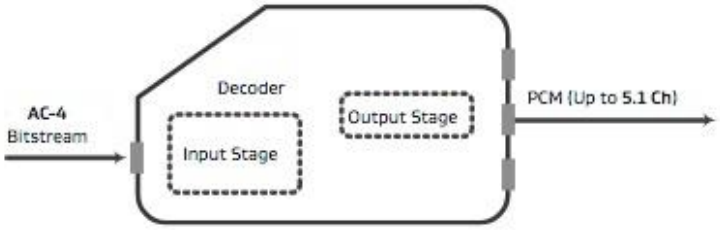
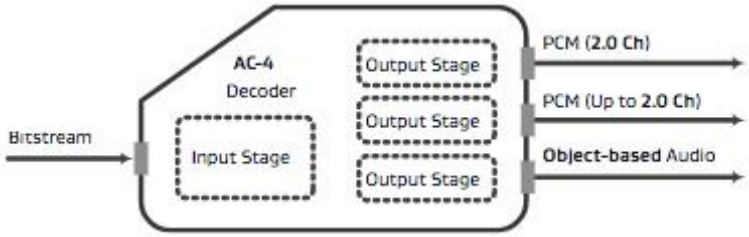


FIGURE 13

Three independent outputs



5 Content replacement and insertion

5.1 Audio/Video frame alignment

In current digital broadcast systems, encoded audio and video utilize different frame rates.

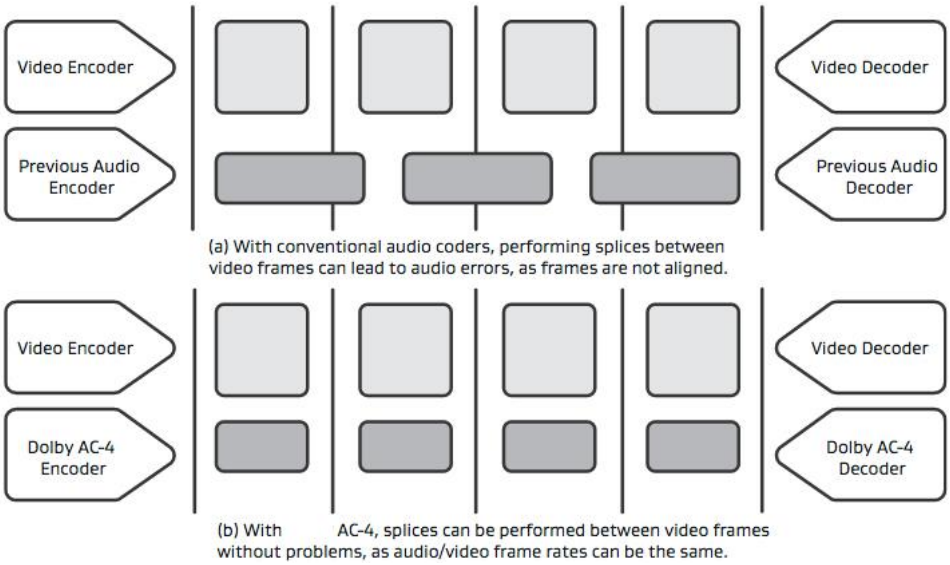
Although in isolation these rates make sense, the combination of two different rates in a final delivery stream or package makes further manipulation of the programme in the transport domain complex. Applications such as editing, ad insertion, and international turnarounds become challenging to implement, as the switching points at the end of video frames do not align with the ends of audio frames.

If not implemented carefully, this can result in sync errors between video and audio or audible audio errors. Current solutions to this involve decoding and re-encoding the audio, which introduces potential sync errors, quality loss, and, in the case of OBA, misalignment of audio and time-critical positional metadata.

In AC-4, a new approach is taken. The AC-4 encoder features an optional video reference input to align the audio and video frames. The encoded audio frame rate can therefore be set to match the video frame rate, and as a result, the boundaries of the audio frames can be precisely aligned with the boundaries of the video frames. The AC-4 system accommodates current broadcast standards, which specify video rates from 23.97 to 60 Hz as well as support for rates up to 120 Hz for new ultra high-definition specifications.

FIGURE 14

Audio/Video frame alignment in AC-4



This approach simplifies implementation for developers of downstream systems and reduces the risk of sync errors and other artefacts caused by trimming or cutting a stream. It does not require that the cut point be known when encoding the material, which provides flexibility for downstream manipulation within headends, the delivery network and consumer devices.

5.2 Seamless switching

The AC-4 decoder supports seamless switching between data rates and configurations. This is achieved by a decoder that reconfigures itself according to the input frame. A smooth transition is ensured because all relevant metadata is carried in precise alignment with the audio to which it refers.

Seamless switching also enables glitch-free transitions between different audio streams. To facilitate detection, the sequence number of each frame is checked in the decoder; if the frame numbers are non-consecutive, a crossfade is triggered. Because the codec operates on the overlap-add principle, there is sufficient audio available to perform a short crossfade so as to avoid any artefacts that would otherwise arise from a hard switch. If the content is the same before and after a switch – as in, for example, a switch of the bit rate in an adaptive streaming scenario – that transition becomes unnoticeable.

This allows a service provider using adaptive streaming to utilize a wide range of bit rates / operation points with one audio coding system, as the AC-4 decoder will smoothly switch at transition points between rates. Furthermore, to ease the integration in playback devices, the AC-4 decoder has a fixed latency across all configurations. This simplifies integration for developers, as it is not necessary to add code to reconfigure the system depending on the input type. It also removes the risk of sync errors and audible glitches when switching configurations.

References

[1] ETSI TS 103 190-1 v1.3.1 (2018-02), "Digital Audio Compression (AC-4) Standard Part 1; Channel based coding", https://www.etsi.org/deliver/etsi_ts/103100_103199/10319001/01.03.01_60/ts_10319001v010301p.pdf.

[2] ETSI TS 103 190-2 v1.2.1 (2018-02), "Digital Audio Compression (AC-4) Standard Part 2: Immersive and personalized audio", https://www.etsi.org/deliver/etsi_ts/103100_103199/10319002/01.02.01_60/ts_10319002v010201p.pdf.

[3] ETSI TS 101 154 v2.6.1 (2019-09), "Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcasting Application based on the MPEG-2 Transport Stream", https://www.etsi.org/deliver/etsi_ts/101100_101199/101154/02.06.01_60/ts_101154v020601p.pdf.

[4] ATSC A/342-1:2021, "ATSC Standard: A/342 Part 1, Audio Common Elements", March 9, 2021, https://www.atsc.org/wp-content/uploads/2021/03/A342-2021-Part-1.pdf.

[5] ATSC: A/342-2:2021, "ATSC Standard: A/342 Part 2, AC-4 System", March 10, 2021, https://www.atsc.org/wp-content/uploads/2021/03/A342-2021-Part-2-AC-4.pdf.

[6] ANSI/SCTE 242-1:2017, "Next Generation Audio Coding Constraints for Cable Systems: Part 1 – Introduction and Common Constraints", https://www.scte.org/standards-development/library/standards-catalog/ansiscte-242-1-2017/

[7] ANSI/SCTE 242-2 2017, "Next Generation Audio Coding Constraints for Cable Systems: Part 2 – AC-4 Audio Coding Constraints", https://www.scte.org/standards-development/library/standards-catalog/ansiscte-242-2-2017/

[8] ANSI/SCTE 243-1 2017, "Next Generation Audio Carriage Constraints for Cable Systems: Part 1 – Common Transport Signaling", https://www.scte.org/standards-development/library/standards-catalog/ansiscte-243-1-2017/

[9] ANSI/SCTE 243-2 2017, "Next Generation Audio Carriage Constraints For Cable Systems: Part 2 - AC-4 Audio Carriage Constraints", https://www.scte.org/standards-development/library/standards-catalog/ansiscte-243-2-2017/

[10] ETSI TS 102 796 v1.5.1 (2018-09), "Hybrid Broadcast Broadband TV", https://www.etsi.org/deliver/etsi_ts/102700_102799/102796/01.05.01_60/ts_102796v010501p.pdf.

[11] ETSI TS 103 448 v1.1.1 (2016-09), AC-4 Object Audio Renderer for Consumer Use, https://www.etsi.org/deliver/etsi_ts/103400_103499/103448/01.01.01_60/ts_103448v010101p.pdf.

[12] ETSI EN 300 468 v1.16.1 (2019-08), "Digital Video Broadcasting (DVB); Specification for Service Information (SI) in DVB systems", https://www.etsi.org/deliver/etsi_en/300400_300499/300468/01.16.01_60/en_300468v011601p.pdf.

[13] "NorDig Unified Requirements for Integrated Receiver Decoders" v3.1.2 (2021-04-13), https://nordig.org/wp-content/uploads/2016/03/NorDig-Unified-Requirements-ver.-3.1.2.pdf

[14] "UHD Book – Compatible High Definition and Ultra High Definition receivers for the Italian market: baseline requirements – Version 2.0" (2020-12-01), https://www.hdforumitalia.it/documents/uhd-book-2-0/?lang=en

[15] "Regulation of the minister for digitalization on the technical and operational requirements for digital receivers", https://ec.europa.eu/growth/tools-databases/tris/en/search/?trisaction=search.detail&year=2019&num=213

[16] ETSI TS 102 366 v1.4.1 (2017-09), "Digital Audio Compression (AC-3, Enhanced AC-3) Standard", https://www.etsi.org/deliver/etsi_ts/102300_102399/102366/01.04.01_60/ts_102366v010401p.pdf

Annex 2

MPEG-H Audio System

Introduction

The MPEG-H Audio System is a Next Generation Audio (NGA) system offering true immersive sound and advanced user interactivity features. It is based on MPEG-H 3D Audio (short MPEG-H Audio), which is an open international ISO standard and standardized in ISO/IEC 23008-3 [1]. Its object-based concept of delivering separate audio elements with metadata within one audio stream enables personalization and universal delivery. The MPEG-H 3D Audio Low Complexity Profile Level 3 is adopted by DVB in ETSI TS 101 154 v.2.3.1 [2] and is one of the audio systems standardized for use in ATSC 3.0 Systems as defined in A/342 Part 3 [3]. SCTE has included the MPEG-H Audio System into the suite of NGA standards for cable applications as specified in SCTE 242-3 [4].

The MPEG-H Audio System was selected by the Telecommunications Technology Association (TTA) in South Korea as the sole audio codec for the terrestrial UHDTV broadcasting specification TTAK.KO- 07.0127 [5] that is based on ATSC 3.0. On May 31, 2017, South Korea launched its 4K UHD TV service using the MPEG-H Audio System.

The MPEG-H 3D Audio Low Complexity Profile Level 3 was selected by 3GPP as the sole audio codec for distribution of channel, object and scene-based 3D audio for virtual reality and streaming applications over 5G networks, as described in the ETSI TS 126 118 specification [6]. The MPEG-H Audio System was adopted by the SBTVD (Sistema Brasileiro de Televisão Digital / Brazilian Digital Television System) Forum and included in the Brazilian specification for broadcast over ISDB-Tb: ABNT NBR 15602-2:2020, Digital terrestrial television – Video coding, audio coding and multiplexing Part 2: Audio coding [7]. TV Globo has conducted the first broadcast with MPEG-H Audio over ISDB-Tb in Brazil in September 2019. Additionally, broadcasters all over Europe have tested MPEG-H Audio over DVB-T2, DVB-S2 as well as streaming during major sports and music events.

The MPEG-H Audio System was adopted by all major specifications for streaming and hybrid delivery, including: MPEG CMAF [8], CTA-WAVE [9], DASH-IF [10], DVB-DASH [11], HbbTV [12] and DTV Play [13].

Besides the MPEG-H 3D Audio Low Complexity Profile, MPEG has standardized the MPEG-H 3D Audio Baseline profile as a subset of the Low Complexity Profile which supports channel- and object-based audio signals as specified in Section 4.8 of ISO/IEC 23008-3 [1]. The Baseline profile was designed for broadcast, streaming and immersive music streaming applications and MPEG has validated its high quality as documented in the MPEG-H 3D Audio Baseline Verification Test Report [14]. The Low Complexity profile was designed for broadcast and streaming with reduced decoder complexity and MPEG has validated its high quality in the MPEG-H 3D Audio Verification Test Report [16].

1 Overview of the MPEG-H Audio System

Efficient delivery and reproduction of immersive sound is a key feature of MPEG-H Audio. Distinguished from surround sound by expanding the sound image in the vertical dimension (i.e. the sound can come from all directions, including above or below the listener's head), immersive sound offers the listener a more enveloping and realistic experience. This increases the spatial impression and makes the user feel more 'part of the scene', diminishing the awareness of being a remote viewer.

1.1 Immersive sound

MPEG-H Audio can natively deliver immersive sound using any combination of the three well-established formats:

– Channel-based: Each transmission channel is traditionally associated with a precisely defined and fixed target location of the loudspeakers relative to the listener. An immersive sound experience is usually created by enhancing the traditional 5.1 or 7.1 surround configurations with height loudspeakers. Typically, four height loudspeakers are added on an upper layer above the middle layer (i.e. the horizontal listener plane) in a home environment.

– Object-based: Each individual audio object may be positioned in three dimensions independently of loudspeaker positions based on the associated side information. The main difference between objects and channels is that the spatial position of an audio object can vary over time, and the positioning information is carried as side information amongst other metadata used to fully describe the object. The associated metadata enables the decoder to render the object to the final loudspeaker setup at the receiver side.

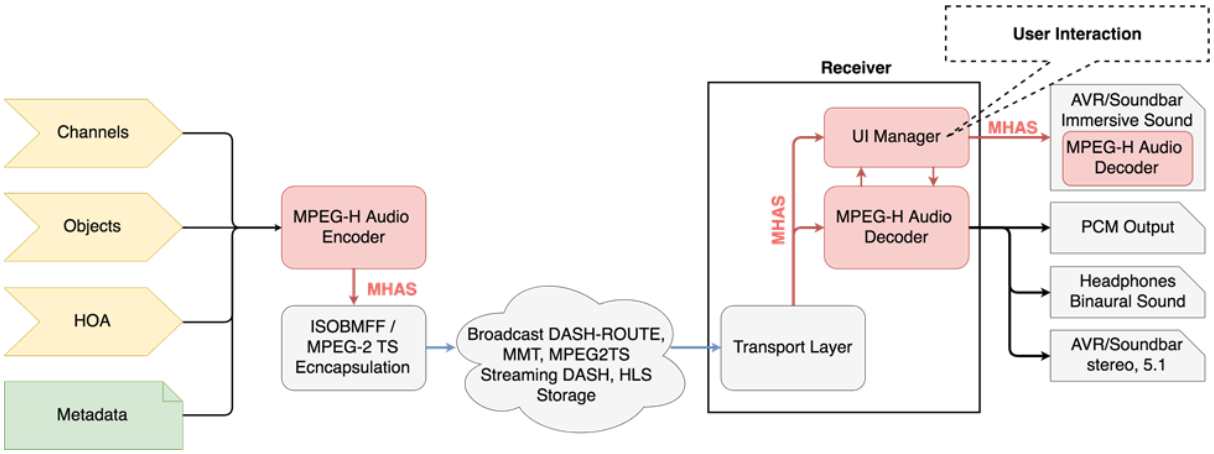
– Scene-based (or Ambisonics): A sound scene is represented by a set of spherical harmonic coefficients that have no direct relationship to channels or objects, but instead describe the sound field. Scene-based audio is natively carried in MPEG-H Audio as Higher Order Ambisonics (HOA). The HOA approach is designed for capturing and representing a sound field. The HOA coefficient signals can be easily manipulated using only matrix operations. This makes the sound field representation adaptable to different reproduction methods, such as, for immersive sound playback over headphones with enabled head tracking.

The most common use case for broadcast applications is using a mixture of a fixed immersive channel "bed" (e.g. 7.1.4) and several additional audio objects (e.g. several languages and Video Descriptive Services for broadcast, or overhead spatial effects for cinematic content).

As shown in Fig. 15, MPEG-H Audio can carry any combination of Channels, Objects and HOA signals in an efficient way, together with the metadata required for rendering, advanced loudness control, personalization and interactivity.

FIGURE 15

Overview of the MPEG-H Audio System



The MPEG-H Audio Stream (MHAS), described in § 3, contains the audio bitstream and various types of metadata packets and represents a common layer for encapsulation into any transport layer format (e.g. MPEG-2 TS, ISOBMFF). The MPEG-H Audio enabled receiver can decode and render the audio to any loudspeaker configuration or a Binaural Audio representation for headphones reproduction. For enabling the advanced user interactivity features in cases where external playback

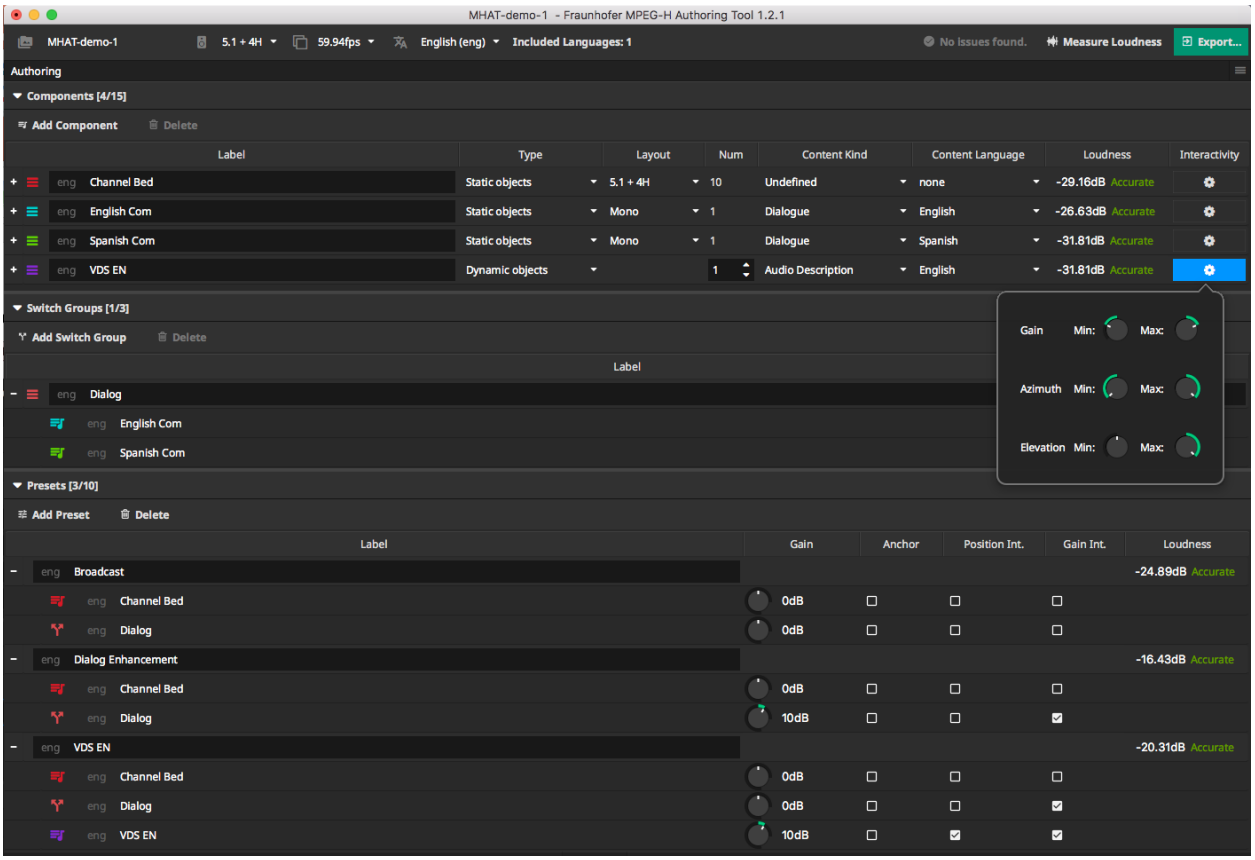
devices are used, the UI Manger can supply the user interactions by inserting new MHAS packets into the MHAS stream and further deliver this over HDMI to the subsequent immersive AVR/Soundbar with MPEG-H Audio decoding capabilities. This is described in more detail in §4.

1.2 Personalization and interactivity

MPEG-H Audio enables viewers to interact with the content and personalize it to their preference. The MPEG-H Audio Metadata carries all the information needed for personalization such as attenuating or increasing the level of objects, disabling them, or changing their position. The metadata also contains information to control and restrict the personalization options such as setting the limits in which the user can interact with the content, as illustrated in Fig. 16 (see also § 2 on MPEG-H Audio Metadata).

FIGURE 16

MPEG-H authoring tool example session

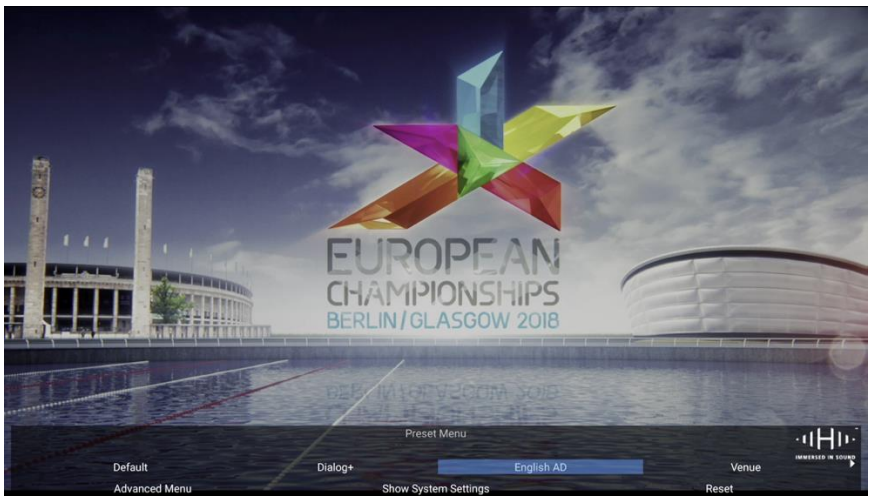


The MPEG-H Audio Metadata carries a rich set of information needed to enable viewers to manipulate the audio objects by attenuating or increasing their level, disabling them, or changing their position in the three-dimensional space. Usually, content providers and broadcasters desire control over the viewers' degrees of freedom to alter the way in which content is consumed. This is why the MPEG-H Audio Metadata gives broadcasters full control over the interactivity options they can offer, allowing them to strictly set the limits in which the user can interact with the content.

Additionally, with MPEG-H Audio, broadcasters can provide several versions of the content, as so-called ‘presets’, which describe how all channels, objects and HOA signals in one bitstream are mixed together and presented to the viewer. Choosing between different presets is the simplest way to interact with the content and will probably be used by most of the TV viewers.

FIGURE 17

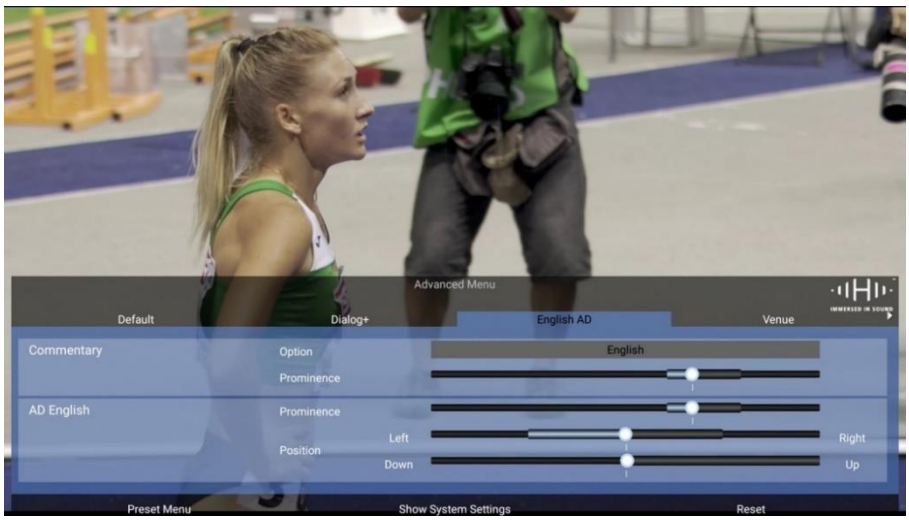
Example of user interface using presets



Furthermore, advanced interactivity settings can be offered to more experienced users interested in manipulating the individual objects. Figure 17 illustrates a typical user interface for providing a simple menu for selection of different presets, and Fig. 18 illustrates a more advanced interactivity menu after accessing the advanced menu.

FIGURE 18

Example of user interface using advanced interactivity options



1.3 Universal delivery

MPEG-H Audio provides a complete integrated audio solution for delivering the best possible audio experience, independently of the final reproduction system. It includes rendering and downmixing functionality, together with advanced Loudness and Dynamic Range Control (DRC).

The loudness normalization module ensures consistent loudness across programs and channels, for different presets and playback configurations, based on loudness information embedded in the MPEG-H Audio Stream. Providing loudness information for each preset allows for instantaneous and automated loudness normalization when the user switches between different presets. Additionally, downmix-specific loudness information can be provided for artist-controlled downmixes.

2 MPEG-H Audio Metadata

MPEG-H Audio enables NGA features such as personalization and interactivity with a set of static metadata, the Metadata Audio Elements (MAE). Audio Objects are associated with metadata that contain all information necessary for personalization, interactive reproduction, and rendering in flexible reproduction layouts. This metadata is part of the overall set-up and configuration information for each piece of content.

2.1 Metadata structure

The MAE is structured in several hierarchy levels. The top-level element is the Audio Scene Information or the "AudioSceneInfo" structure as shown in Fig. 19. Sub-structures of the AudioSceneInfo contain descriptive information about "Groups", "Switch Groups", and "Presets". An "ID" field uniquely identifies each group, switch group or preset, and is included in each sub-structure.

The group structures ("mae_GroupDefinition") contain descriptive information about the audio elements, such as:

– the group type (channels, objects or HOA);

– the content type (e.g. dialogue, music, effects);

– the language for dialogue objects; or

– the channel layout in case of channel-based content.

User interactivity can be enabled for the gain level or position of objects, including restrictions on the range of interaction (i.e. setting minimum and maximum values for gain and position offset). The minimum and maximum values can be set differently for each group.

Groups can be combined into switch groups ("mae_SwitchGroupDefinition"). All members of one switch group are mutually exclusive, i.e. during playback, only one member of the switch group can be active or selected. As an example, using a switch group for dialogue objects ensures that only one out of multiple dialogue objects with different languages is played back at the same time. Additionally, one member of the switch group is always marked as default to be used if there is no user preference setting and to make sure that the content is always played back with dialogue, for example.

The preset structures ("mae_GroupPresetData") can be used to define different "packages" of audio elements within the Audio Scene. It is not necessary to include all groups in every preset definition. Groups can be "on" or "off" by default and can have a default gain value. Describing only a sub-set of groups in a preset is allowed. The audio elements that are packaged into a preset are mixed together in the decoder, based on the metadata associated with the preset, and the group and switch group metadata.

From a user experience perspective, the presets behave as different complete mixes from which users can choose. The presets are based on the same set of audio elements in one Audio Scene and thus can share certain audio objects/elements, like a channel-bed. This results in bitrate savings compared to a simulcast of a number of dedicated complete mixes.

Textual descriptions (labels) can be associated with groups, switch groups and presets, for instance "Commentary" in the example below for a switch group. Those labels can be used to enable personalization in receiving devices with a user interface.

2.2 Metadata example

FIGURE 19

Example of an MPEG-H Audio Scene Information

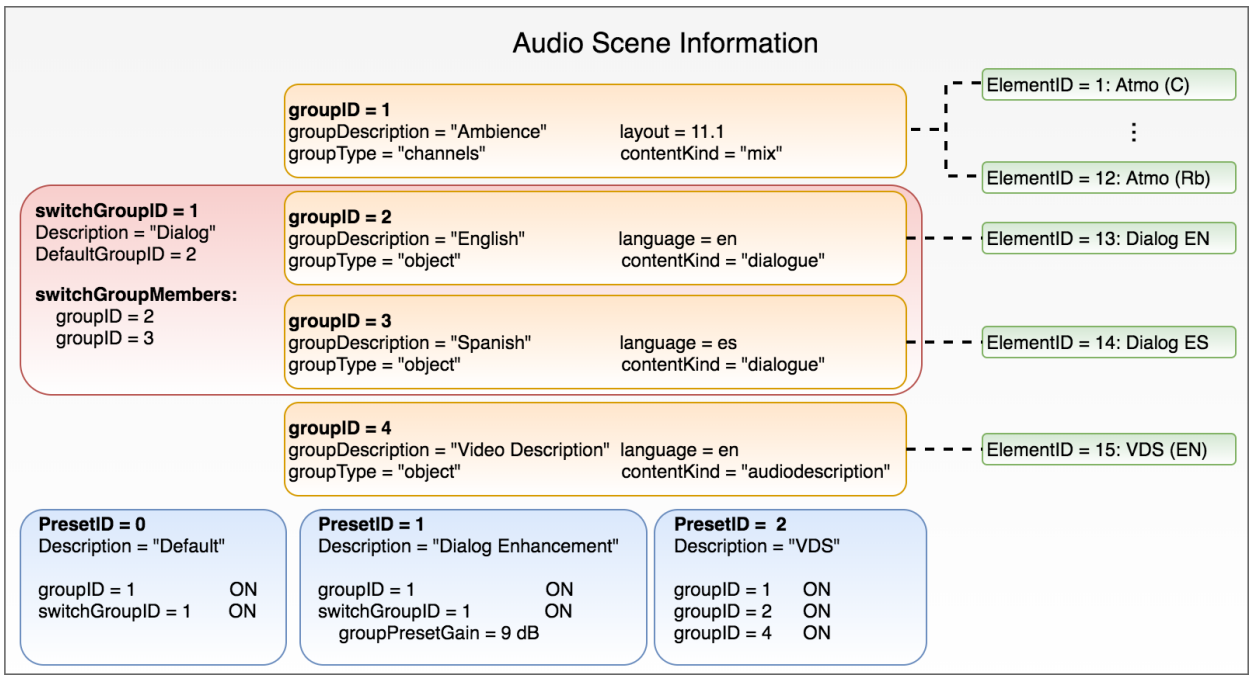


Figure 19 contains an example of MPEG-H Audio Scene Information with four different groups (orange), one switch group (red) and three presets (blue). In this example, the switch group contains two dialogues in different languages that the user can choose from, or the device can automatically select one dialogue based on the preference settings.

The “Default” preset (“PresetID = 0”) for this Audio Scene contains the “Ambience” group (“groupID = 1”) and the “Dialog” switch group (“switchGroupID = 1”) wherein the English dialogue (“groupID = 2”) is the default. Both the ambience group and the dialogue switch group are active (“ON”). This preset is automatically selected in the absence of any user or device automatic selection. The additional two presets in this example enable the advanced accessibility features as described in the following sub-sections.

The “Dialogue Enhancement” preset contains the same elements as the default preset, with the same status (“ON”) with the addition that the dialogue element (i.e. the switch group) is rendered with a 9 dB gain into the final mix. The gain parameter, determined by the content author, can be any value from −63 to +31 in 1 dB steps.

The "VDS" preset contains three groups, all active: the ambience (“groupID = 1”), the English dialogue (“groupID = 2”) and the Video Description (“groupID = 4”).

The “VDS” preset can be manually selected by the user or automatically selected by the device based on the preference settings (i.e. if Video Description Service is enabled in the device’s settings).

2.3 Personalization use case examples

2.3.1 Advanced Accessibility

Object-based audio delivery with MPEG-H Audio offers advanced and improved accessibility services, especially:

– Video Descriptive Services (VDS, also known as Audio Description) and

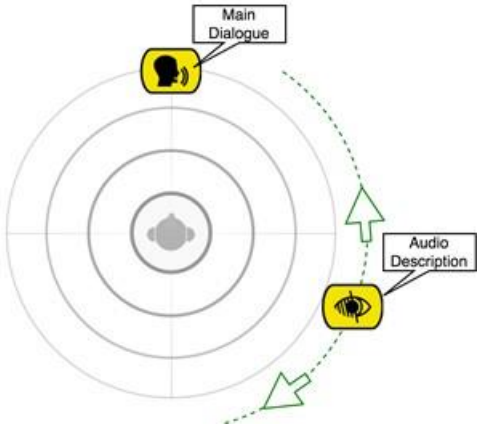
– Dialogue Enhancement (DE).

As described in the previous section, the dialogue elements and the Video Description are carried as separate audio objects (“groups”) that can be combined with a channel bed element in different ways and create different presets, such as a “default” preset without Video Description and a “VDS” preset.

Additionally, MPEG-H Audio allows the user to spatially move the Video Description object to a user selected position (e.g. to the left or right), enabling a spatial separation of main dialogue and the Video Description element, as shown in Fig. 20. This results in a better intelligibility of the main dialogue as well as the Video Description (e.g. in a typical 5.1 set-up the main dialogue is assigned to the centre speaker while the Video Description object could be assigned to a rear-surround speaker).

FIGURE 20

Audio description re-positioning example



2.3.2 Dialogue Enhancement (DE)

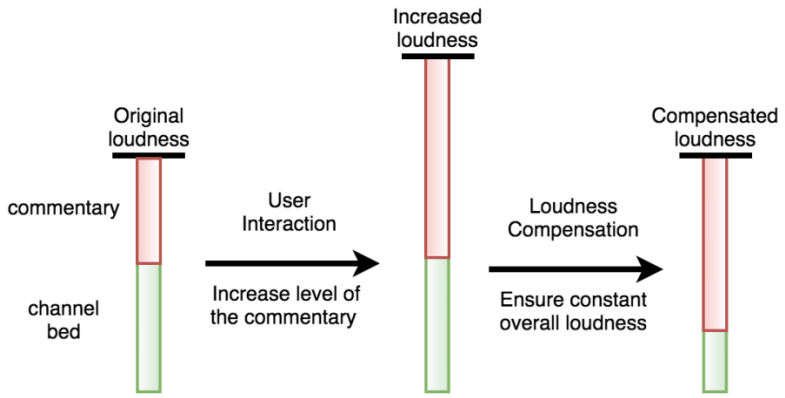
MPEG-H Audio includes a feature of DE that enables automatic device selection (prioritization) as well as user manipulation. For ease of user selection or for automatic device selection (e.g. enabling TV “Hard of Hearing” TV setting), a Dialogue Enhancement preset can be created, as illustrated in Fig. 19 Audio description re-positioning example using a broadcaster defined enhancement level for the dialogue element (e.g. 10 dB as shown in Fig. 16).

Moreover, if the broadcaster allows personalization of the enhancement level, MPEG-H Audio supports advanced DE which enables direct adjustment of the enhancement level via the user interface. The enhancement limitations (i.e. maximum level) are defined by the broadcaster/content creator as shown in Fig. 16 and carried in the metadata. This maximum value for the lower and upper end of the scale can be set differently for different elements as well as for different content.

The advanced loudness management tool of MPEG-H Audio automatically compensates loudness changes that result from user interaction (e.g. switching presets or enhancement of dialogue) to keep the overall loudness on the same level, as illustrated in Fig. 21. This ensures constant loudness level not only across programmes but also after user interactions.

FIGURE 21

Loudness compensation after user interaction



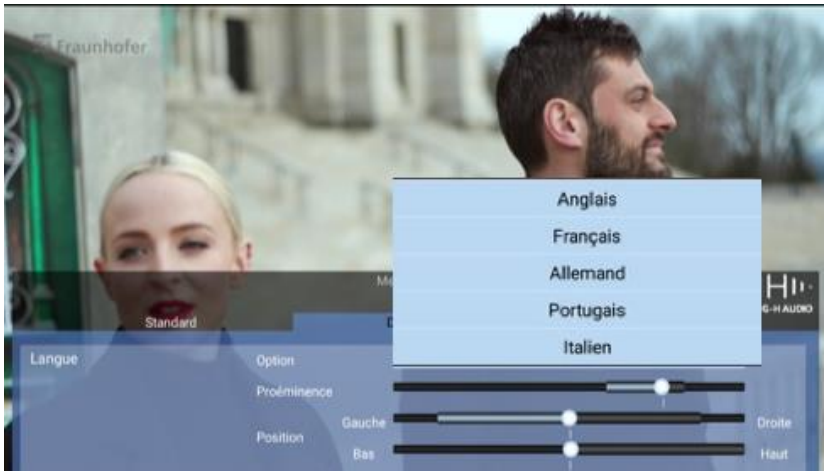
2.3.3 Multi-language services

With existing audio codecs, multi-language programs are broadcast as separate complete mixes in each language. Using one stream for each mix requires a high bitrate, directly proportional to the number of additional languages offered. Moreover, if Video Descriptive Services (VDS) have to be provided as additional complete mixes, the required bandwidth would increase even more.

MPEG-H Audio enables a much more efficient way of offering accessibility and multi-language services by making use of object-based audio, similar to the DE feature, as described above. With a common channel bed and individual audio objects for each language dialogue and audio description tracks, MPEG-H Audio requires a significantly lower bitrate than legacy systems. For example, a 5.1 programme is delivered in five different languages in a single stream using one audio object for each language. A legacy system would require transport of six complete 5.1 mixes in five different streams.

FIGURE 22

Example of multi-lingual service using MPEG-H Audio



Assuming typical bitrates for legacy 5.1 surround sound delivered using HE-AAC, as shown in Table 3, the object-based approach of MPEG-H Audio brings more than 50% in bitrate saving.

TABLE 3

Example of bit rates comparison with legacy codecs for multi-language services

5.1 multi-channel surround in five languages			
Bit rate using MPEG-H Audio (kbit/s)		**Bit rate using HE-AAC (kbit/s)**	
5.1 Bed	128	5.1 in Language 1	160
Language 1	32	5.1 in Language 2	160
Language 2	32	5.1 in Language 3	160
Language 3	32	5.1 in Language 4	160
Language 4	32	5.1 in Language 5	160
Language 5	32		
Total	**288**	**Total**	**800**

Moreover, all interactivity and personalization features can be enabled in a single stream. This simplifies the required signalling on the transport layer and the selection process on the receiver side.

2.3.4 Personalization for sports programmes

For various programme types, such as sports programmes, MPEG-H Audio provides additional advanced interactivity and personalization options, such as choosing between 'home team' and 'away team' commentaries of the same game, listening to the team radio communication between the driver and his team during a car race, or listening only to the crowd (or home/away crowd) with no commentary during a sports program.

2.3.5 Presentation of MPEG-H Audio services

The MPEG-H Audio Metadata describes the presets (“labels”) in multiple languages. The content creator can decide based on the regions where its content is distributed to author all labels in one or more languages. Based on the receiver’s preferred language setting the correct labels will be displayed to the viewer. For example, Fig.23 shows the labels authored during a live broadcast trial in two languages: English and French.

FIGURE 23

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

Using the MPEG-H Audio Metadata, the content creators or broadcasters can ensure that their artistic intent and the various features they want to enable are correctly displayed to the user. In this way

broadcasters are always in control of their content and the users will experience the content in the same way on all devices.

2.4 Interoperability with the Audio Definition Model

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

ADM profiles provide interoperability in ADM-based content ecosystems by incorporating application-specific and platform-dependent requirements for production, distribution and emission.

To achieve interoperability with the MPEG-H Audio Metadata as outlined in § 2 and standardized in ISO/IEC 23008-3 [1], the MPEG-H ADM Profile provides constraints on Recommendations ITU-R BS.2076 and ITU-R BS.2125 that enable transparent and reliable conversion of ADM-based content for distribution with the MPEG-H Audio System in file-based and linear workflows.

3 MPEG-H Audio Stream

The MPEG-H Audio Stream (MHAS) format is a self-contained, packetized, and extensible byte stream format to carry MPEG-H Audio data. The basic principle of the MHAS format is to separate encapsulation of coded audio data, configuration data and any additional metadata or control data into different MHAS packets. Therefore, it is easier to access configuration data or other metadata on the MHAS stream level without the need to parse the audio bitstream.

FIGURE 24

MHAS packet structure

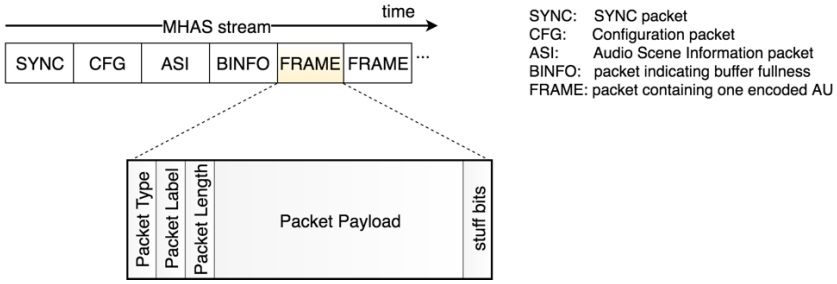


Figure 24 shows the high-level structure of an MHAS packet, which contains the header with the packet type to identify each MHAS packet, a packet label and length information, followed by the payload and potential stuffing bits for byte alignment.

The packet label has the purpose of differentiating between packets that belong either to different configurations in the same stream, or different streams in a multi-stream environment.

3.1 Random Access Points

A Random Access Point (RAP) consists of all MHAS packets that are necessary to tune to a stream and enable start-up decoding: a potential sync packet, configuration data and an independently decodable audio data frame.

If the MHAS stream is encapsulated into an MPEG-2 Transport Stream, the RAP also needs to include a sync packet. For ISOBMFF encapsulation, the sync packet is not necessary, because the ISO file format structure provides external framing of file format samples.

The configuration data is necessary to initialize the decoder, and consists of two separate packets, the audio configuration data and the Audio Scene information metadata.

The encoded data frame of a RAP has to contain an "Immediate Playout Frame" (IPF), i.e. an Access Unit (AU) that is independent from all previous AUs. It additionally carries the previous AU's information, which is required by the decoder to compensate for its start-up delay. This information is embedded into the Audio Pre-Roll extension of the IPF and enables valid decoded PCM output equivalent to the AU at the time instance of the RAP.

3.2 Configuration changes and A/V Alignment

When the content set-up or the Audio Scene Information changes (e.g. the channel layout or the number of objects changes), a configuration change can be used in an audio stream for signalling the change and ensure seamless switching in the receiver.

Usually, these configuration changes happen at programme boundaries (e.g. corresponding to ad insertion), but may also occur within a program. The MHAS stream allows for seamless configuration changes at each RAP.

Audio and video streams usually use different frame rates for better encoding efficiency, which leads to streams that have different frame boundaries for audio and video. Some applications may require that audio and video streams are aligned at certain instances of time to enable stream splicing.

MPEG-H Audio enables sample-accurate configuration changes and stream splicing using a mechanism for truncating the audio frames before and after the splice point. This is signalled on MHAS level through the AUDIOTRUNCATION packet.

An AUDIOTRUNCATION packet, indicating that the truncation should not be applied, can be inserted at the time when the stream is generated. The truncation can be easily enabled at a later stage on a systems level.

FIGURE 25

Example of a configuration change from 7.1.4 (also known as 7.1+4H) to 2.0 in the MHAS stream

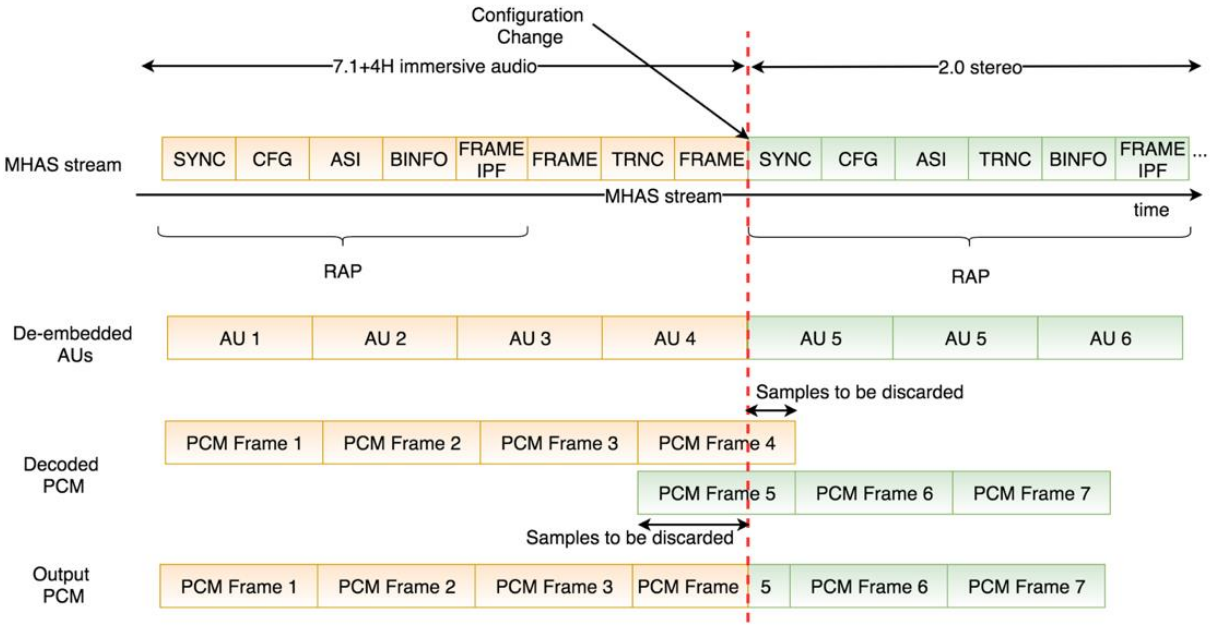


Figure 25 shows an example of a sample-accurate configuration change from an immersive audio set-up to stereo inside one MHAS stream (i.e. in the ad-insertion use case the inserted ad is stereo, while the rest of the programme is in 7.1.4).

The first AUDIOTRUNCATION packet ("TRNC") contained in the first stream indicates how many samples are to be discarded at the end of the last frame of the immersive audio signal, while the second AUDIOTRUNCATION packet ("TRNC") in the second stream indicates the number of audio samples to be discarded at the beginning of the first frame of the new immersive audio signal.

4 Distributed user interface processing

In order to take advantage of the advanced interactivity options, MPEG-H Audio enabled devices require User Interfaces (UIs). In typical home set-ups, the available devices are connected in various configurations such as:

– a Set-Top Box connecting to a TV over HDMI;

– a TV connecting to an AVR/Soundbar over HDMI or S/PDIF.

In all cases, it is desired to have the user interface located on the preferred device (i.e. the source device). For such use cases, the MPEG-H Audio System provides a unique way to separate the user interactivity processing from the decoding step. Therefore, all user interaction tasks are handled by the "UI Manager", in the source device, while the decoding is done in the sink device. This feature is enabled by the packetized structure of the MPEG-H Audio Stream, which allows for:

– easy stream parsing on system level;

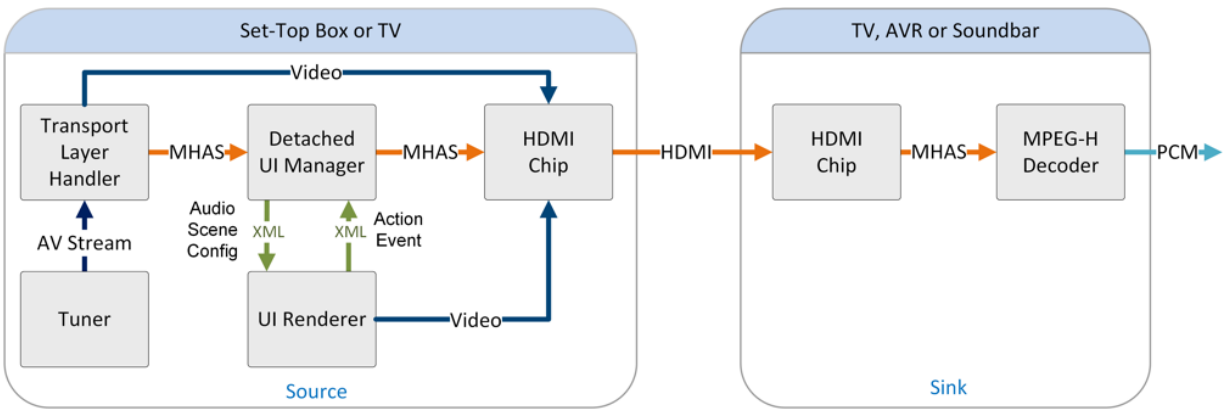
– insertion of new MHAS packets on the fly (e.g. "USERINTERACTION" packets).

Figure 26 provides a high-level block-diagram of such a distributed system between a source and a sink device connected over HDMI. The detached UI Manager has to parse only the MHAS packets containing the Audio Scene Information and provides this information to an UI Renderer to be displayed to the user. The UI Renderer is responsible for handling the user interactivity and passes the information about every user's action to the detached UI Manager, which embeds it into MHAS packets of type USERINTERACTION and inserts them into the MHAS stream.

The MHAS stream containing the USERINTERACTION packets is delivered over HDMI to the sink device which decodes the MHAS stream, including the information about the user interaction, and renders the Audio Scene accordingly.

FIGURE 26

Distributed UI processing with transmission of user commands over HDMI



The USERINTERACTION packet provides an interface for all allowed types of user interaction. Two interaction modes are defined in the interface:

– An advanced interaction mode, where the interaction can be signalled for each element group that is present in the Audio Scene. This mode enables the user to freely choose which groups to play back and to interact with all of them (within the restrictions of allowances and ranges defined in the metadata and the restrictions of switch group definitions).

– A basic interaction mode, where the user may choose one preset out of the available presets that are defined in the metadata audio element syntax.

5 Multi-stream environment

The MPEG-H Audio System can enable all NGA features described in the previous sections using only one stream. This is a much more efficient and robust solution compared to legacy codecs. For applications that involve a hybrid delivery system (i.e. a main MPEG-H Audio Stream delivered over broadcast and additional MPEG-H Audio Streams delivered over broadband), MPEG-H Audio allows distribution of the audio components composing an audio scene over several streams. The MPEG-H Audio Metadata assists the decoder to correctly decode all streams and present the various signalled presets.

When the content is delivered only in a linear fashion, for example over MPEG-2 TS, it is recommended to use only one MPEG-H Audio Stream. There might be applications that could benefit from a multi-stream approach, even for such linear delivery systems. For example, if the content is produced for distribution over various platforms using a multi-stream approach, a cable operator can choose to re-use the streams for delivery over MPEG-2 TS without any additional processing.

The MPEG-H Audio System specifies for multi-stream delivery, independent of the transport format, a mechanism for receivers to merge several MHAS streams into a single stream. This is realized based on the metadata available on the MHAS stream level, without the necessity of decoding any audio data. In this case, the merged stream can be fed into a single MPEG-H Audio decoder. Figure 27 provides a simple scenario with three MHAS streams:

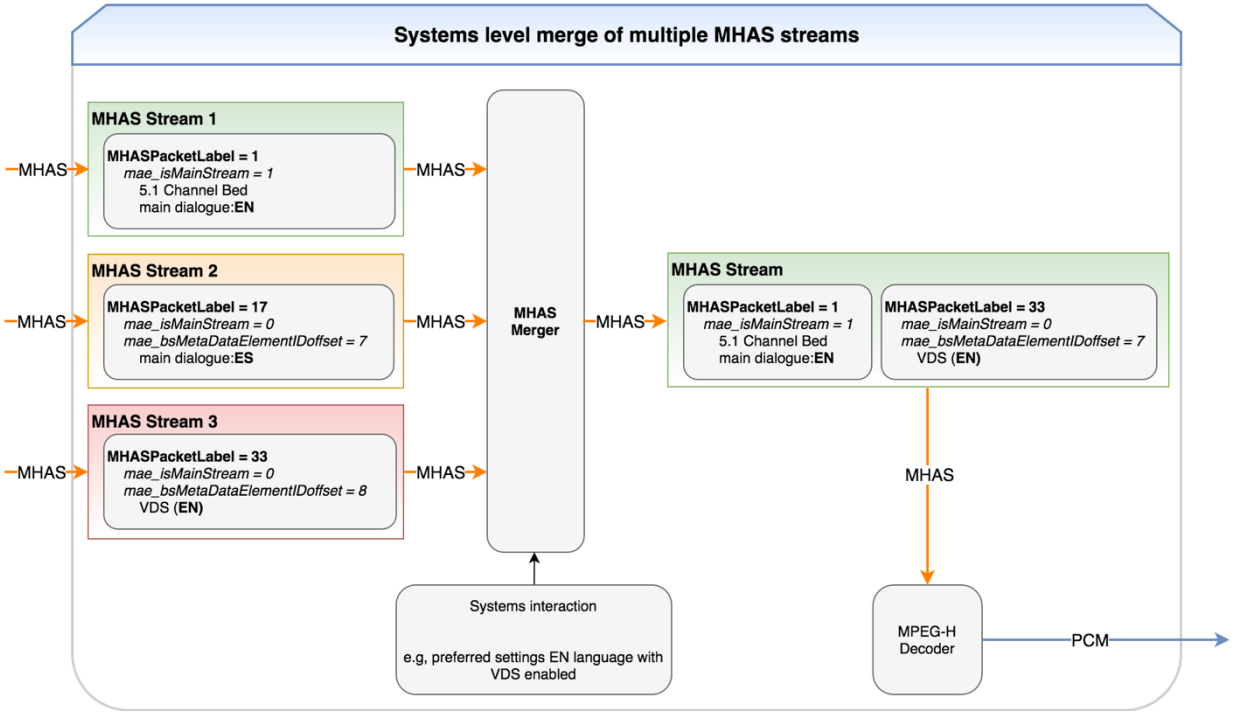
– Stream 1 contains the channel bed and the main dialogue in the original language,

– Stream 2 contains the main dialogue in a different language, and

– Stream 3 provides the VDS service with audio description in original language.

FIGURE 27

Example of selection and merge of multiple MPEG-H Audio Streams



Assuming a receiver selects, on the systems level, the original language and the VDS service, the second stream can be discarded while the first and third streams can be merged into one single stream that is provided to the decoder. The MHAS packets belonging to different streams are identified based on the *MHASPacketLabel* field and the streams are merged based on the MAE information.

6 System sounds and voice assistant sounds

The MPEG-H Audio System standardizes a dedicated solution for delivery of system sounds and voice assistant sounds to external devices, as specified in ISO/IEC 23008-3 [1]. The standard was already adopted by SBTVD in their receiver specification for ISDB-Tb broadcast [15].

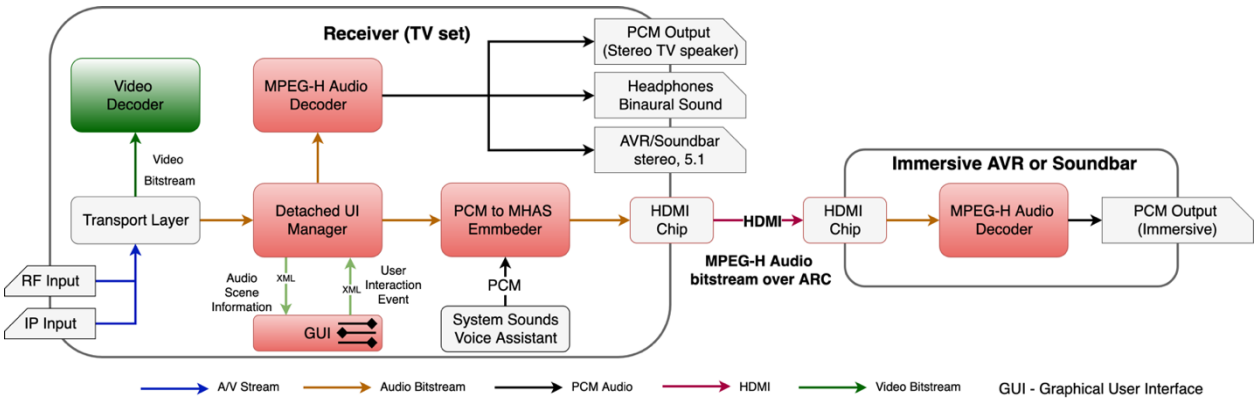
If a receiving device (e.g. TV set or set-top-box) is connected to an external device (e.g. Soundbar or AVR), system sounds and voice assistant sounds can be embedded on the fly in the MPEG-H Audio Stream and delivered to the external device which decodes the MPEG-H Audio Stream.

The process of embedding the system sounds or voice assistant sounds on the fly in an MPEG-H Audio Stream is illustrated in Fig. 28, where the “PCM to MHAS Embedder” module is responsible to receive the mono or stereo PCM data and create the MHAS packets of type:

– PACTYP_EARCON carrying the earconInfo() structure,
– PACTYP_PCMCONFIG carrying the pcmDataConfig() structure and
– PACTYP_PCMDATA carrying the pcmDataPayload() structure.

FIGURE 28

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



Additionally, the "PCM to MHAS Embedder" module embeds the MHAS packets of type PACTYP_EARCON, PACTYP_PCMCONFIG and PACTYP_PCMDATA into the received MHAS packet stream on the fly. The updated bitstream will be passthrough over HDMI to the external device which will decode and render the MPEG-H Audio Stream together with the system sounds and voice assistant sounds as described in ISO/IEC 23008-3 [1].

References

[1] ISO/IEC: 23008-3:2019, Second Edition, "Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D audio", including ISO/IEC 23008-3:2019/AMD 1:2019 "Audio metadata enhancements" and ISO/IEC 23008-3:2019/AMD 2:2020, "3D Audio baseline profile, corrections and improvements", https://www.iso.org/standard/74430.html

[2] ETSI TS 101 154 v2.3.1 (2017-02), "Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcasting Application based on the MPEG-2 Transport Stream", February 14, 2017, http://www.etsi.org/deliver/etsi_ts/101100_101199/101154/02.03.01_60/ts_101154v020301p.pdf

[3] ATSC: A/342-3:2021, "ATSC Standard: A/342 Part 3, MPEG-H System", March 11, 2021, https://www.atsc.org/wp-content/uploads/2021/03/A342-2021-Part-3-MPEG-H.pdf

[4] SCTE 242-3:2017, "Next Generation Audio Coding Constraints for Cable Systems: Part 3 – MPEG-H Audio Coding Constraints", September 25, 2017, https://www.scte.org/standards-development/library/standards-catalog/ansiscte-242-3-2017/

[5] TTA TTAK.KO-07.0127R1 Transmission and Reception for Terrestrial UHDTV Broadcasting Service, December 27, 2016, https://www.tta.or.kr/eng/new/standardization/eng_ttastddesc.jsp?stdno=TTAK.KO-07.0127

[6] ETSI TS 126 118 v15.0.0 (2018-10), 5G; 3GPP Virtual reality profiles for streaming applications (3GPP TS 26.118 version 15.0.0 Release 15), https://www.etsi.org/deliver/etsi_TS/126100_126199/126118/15.00.00_60/ts_126118v150000p.pdf

[7] ABNT NBR 15602-2:2020, Digital terrestrial television – Video coding, audio coding and multiplexing Part 2: Audio coding

[8] ISO/IEC 23000-19:2020, Information technology – Multimedia application format (MPEG-A) – Part 19: Common media application format (CMAF) for segmented media, https://www.iso.org/standard/79106.html

[9] CTA-5001, Web Application Video Ecosystem – Content Specification, https://cdn.cta.tech/cta/media/media/resources/standards/pdfs/cta-5001-b-final_v2.pdf

[10] DASH-IF: Guidelines for Implementation: DASH-IF Interoperability Point for ATSC 3.0, https://dashif.org/docs/DASH-IF-IOP-for-ATSC3-0-v1.1.pdf

[11] ETSI TS 103 285 v1.3.1 (2020-02), Digital Video Broadcasting (DVB); MPEG-DASH Profile for Transport of ISO BMFF Based DVB Services over IP Based Networks, https://www.etsi.org/deliver/etsi_ts/103200_103299/103285/01.03.01_60/ts_103285v010301p.pdf

[12] ETSI TS 102 796 v1.5.1 (2018-09), "Hybrid Broadcast Broadband TV", https://www.etsi.org/deliver/etsi_ts/102700_102799/102796/01.05.01_60/ts_102796v010501p.pdf

[13] ABNT NBR 15606-1:2018, Digital terrestrial television – Data coding and transmission specification for digital broadcasting Part 1: Data coding specification

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

[15] ABNT NBR 15604:2020, Digital terrestrial television – Receivers

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

Annex 3

DTS-UHD Audio format

Introduction

DTS-UHD is a next generation audio format supporting Channel-based Audio (CBA), Object-based Audio (OBA) and Higher Order Ambisonics (HOA) audio encoding and decoding for delivery via a broadcast service. This Annex provides guidelines at a system level to assist the broadcaster in fulfilling current and future audio requirements.

DTS-UHD has been standardized with the European Telecommunications Standards Institute (ETSI) in TS 103 491 v1.2.1 [1] and is included in the DVB Specification TS 101 154 v2.6.1 [2] and the Society of Cable Television Engineers in SCTE 242-4 [3] and 243-4 [4]. DTS-UHD can be encapsulated in a number of transport formats including ISOBMFF, MPEG-2 Transport Stream and CMAF.

In order to deliver all of the solutions outlined in this Annex, a suitable renderer would be required. While one is not specified here, the DTS-UHD Point Source Renderer is fully described in ETSI TS 103 584 v1.1.1 [5].

1 Immersive audio

DTS-UHD audio coding system provides a robust solution to delivering immersive audio via a broadcast mechanism with the introduction of the Audio Compression Engine (ACE) at the core of its technology. This allows the broadcaster or content provider to deliver audio either by channels, objects or a mixture of the two. An example would be an effects and music bed, with dialogue and specific effects delivered as objects placed within a 3-dimensional sound field. This sound field is a development of the more traditional 'surround sound' previously delivered to a 5.1 system, adding 'height' to provide delivery to a 5.1.2, 5.1.4 or 7.1.4 home audio setup. This immersive experience is also suitable for delivery to headphones or soundbars producing a 'virtual' immersive experience.

1.1 Channel-based Audio

DTS-UHD is capable of providing a full channel-based broadcast stream of stereo and 5.1 audio. This has the added advantage of keeping associated metadata to a minimal level. The stream will be constrained to a limited number of playback options; however, renderers are capable of either downmixing a 5.1 to a clear stereo playback, or upmix to a virtualised immersive playback if the consumer requires this.

1.2 Object-based Audio

DTS-UHD is capable of supporting up to 224 discrete audio objects. These can be further organised into 32 object groups and 32 presentations within a single DTS-UHD bitstream. While OBA requires additional metadata to support the audio presentations and personalization control, there are two major advantages to DTS-UHD Object-based Audio:

– Adaptability to the listening environment. Audio programs mixed using OBA do not need to assume a particular listening environment (e.g. speaker layout or dynamic range). This allows the playback system to render the best experience for the listener.

– The ability to adapt to the listener's preference. OBA allows efficient support for features like alternate speech tracks and listener customizations such as changing the speech relative volume (without affecting anything else).

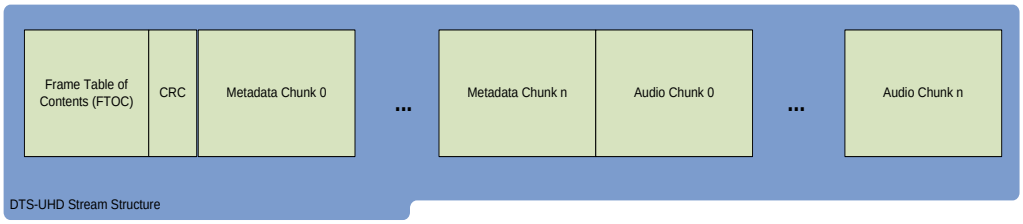
DTS-UHD has provisions for reducing the frequency at which metadata is repeated, thus reducing this burden. OTT streaming methods such as DASH and HLS can utilize larger media in blocks of samples that have guaranteed entry points and DTS-UHD permits encoding options to only update metadata when necessary.

2 Stream construction

A DTS-UHD stream is made of audio frames, which consist of a Frame Table of Contents (FTOC) that provides a description of the audio frame and allows a decoder to navigate directly to the required elements of metadata and audio within the frame.

FIGURE 29

DTS-UHD Stream Structure



The FTOC contains the Sync Word indicating whether this is a sync frame or non-sync frame, default presentations (see § 3.1) and additional dependency information.

The decoder does not need any information from previous or future frames to produce a frame of output Linear PCM samples from a sync frame. All parameters necessary to unpack metadata and audio chunks, describe audio chunks, render and process audio samples and generate a frame of Linear PCM samples can be found within the payload of a sync frame. A decoder establishes initial synchronization exclusively with a Sync frame. These frames represent the random-access points for random navigation to a particular location in the bitstream.

A non-sync frame permits both metadata chunks and audio chunks to minimize payload size by only sending parameters that have changed in value since the previous frame or sync frame, as stated in the introduction. All parameters that are not re-transmitted are assumed to maintain their previous value. Any value or set of values may be updated in a non-sync frame.

A decoder cannot establish initial synchronization using non-sync frames, nor can these non-sync frames be used as random-access points.

2.1 Metadata

Metadata carries the full description of the associated audio objects and how decoded audio shall be rendered for final audio presentation. Additional types of metadata that may be useful for categorization of an audio presentation and support of some post-processing functionality may also be carried within the metadata chunks.

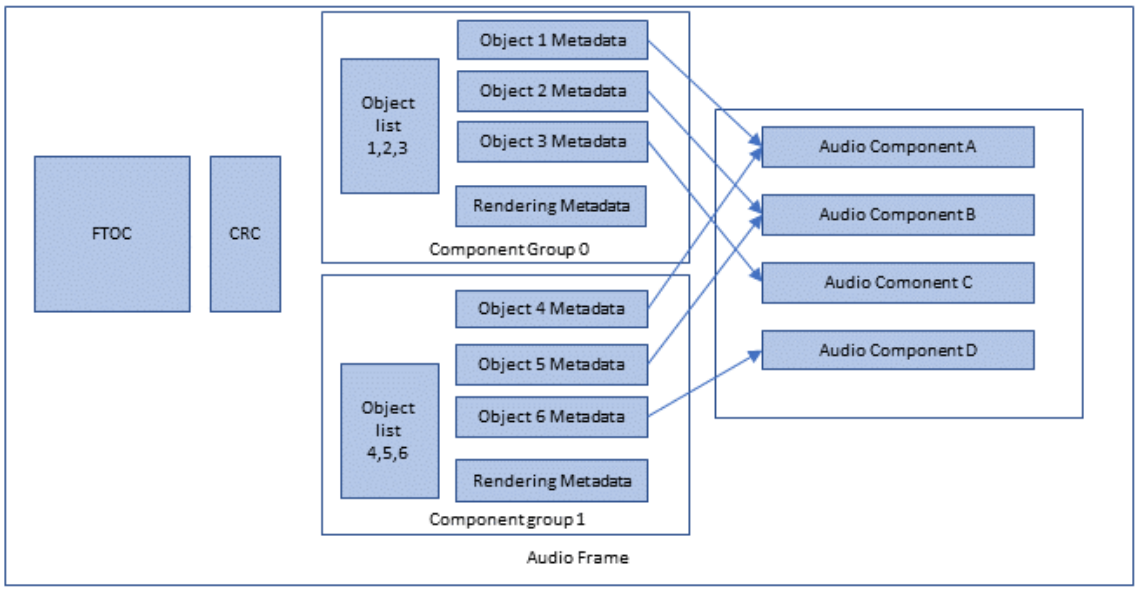
2.2 Audio Objects

Audio Objects carry the compressed audio samples. Audio samples may be representing speaker feeds, waveforms associated with a 3D audio object, waveforms associated with a sound field audio representation, or some other valid audio representation. The metadata and audio are packed immediately after the FTOC CRC word.

Nominally, an audio object points to a minimum collection of compressed waveforms that can be decoded without dependency on any other audio object. All compressed waveforms within an audio object that has been selected for decoding shall be decoded and played together. In some cases, an elementary stream may already have its own sub-division into individually decodable parts, in which case all encoded objects within one DTS-UHD stream can be packed into a single audio object.

FIGURE 30

Audio frame composition



The above diagram shows how the Metadata consists of Object lists, with each associated Object Metadata pointing to an Audio Object or Component.

Organization of the audio parameters into objects allows for the addition of new features and/or quality improvements by simply defining new objects. When new audio objects are defined, legacy

decoders shall recognize and extract audio objects that they are aware of (subject to the system constraints like the maximum number of decodable channels, the maximum sampling frequency, etc.) and ignore all other audio data objects.

3 Presentations

The fundamental unit of a DTS-UHD stream is the object. While this can be a single waveform with associated metadata providing information on placement, volume etc., an object can also be a group of waveforms, such as a stereo, 5.1 or 7.1 channel-based set of waveforms. In addition, Object groups are used, which consist of a number of objects that should always be used together. This can be managed with a single identifier.

Metadata associated to an object can uniquely identify that object within a stream, which points to an associated waveform or group of waveforms. This Metadata also describes the properties necessary to render the waveforms in a default presentation including whether the waveforms should be played or silent. In addition, it will describe the loudness and dynamic properties of the object.

Metadata associated with an Object group again can uniquely identify an object group within a stream with a unique ID. The Metadata will consist of an object list identifying each unique object and will again indicate in a default presentation whether to be played or silent. This can be done at an individual object level within the group and can overwrite the original default setting on any individual object. This ability to manage and target objects at an individual and group level allows for many ways to control and personalise the audio. In order to simplify this, audio is managed through presentations.

3.1 Presentation methodology

Presentations are composed of a selection of objects and/or object groups. An object or object group can appear in a number of different presentations, with the presentation also able to select objects uniquely if the object group is not desired within the presentation.

Multiple presentations may be present within a single DTS-UHD bitstream, with each presentation having a unique presentation index within the stream. Each presentation within a stream may be categorized as:

selectable: where the playback of this presentation does not need playback of any of the audio presentations with the higher audio presentation index; or

non-selectable: where the playback of this presentation requires playback of one or more audio presentations with the higher presentation index. Note that the non-selectable presentation is the default presentation if it is the only audio presentation in the stream.

Two selectable presentation examples are provided below. The first shows a simple default presentation consisting of two Object groups, while the second presentation is made up of Object groups and explicit objects, with the active and non-active flag shown.

FIGURE 31

Presentation formed from object groups

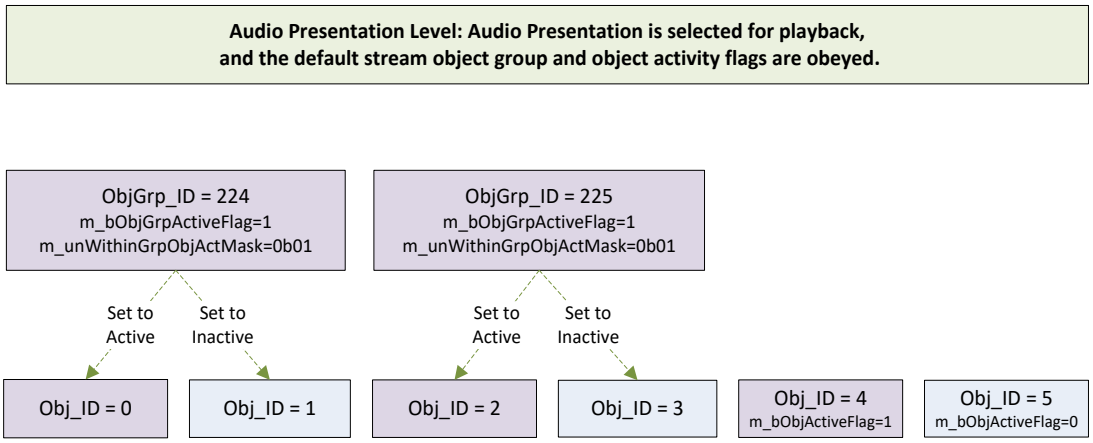
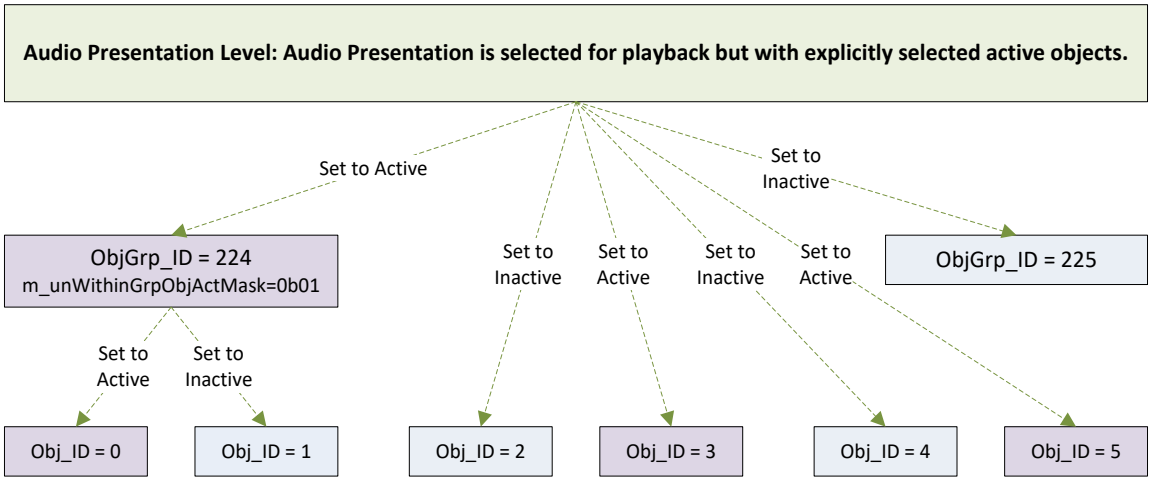


FIGURE 32

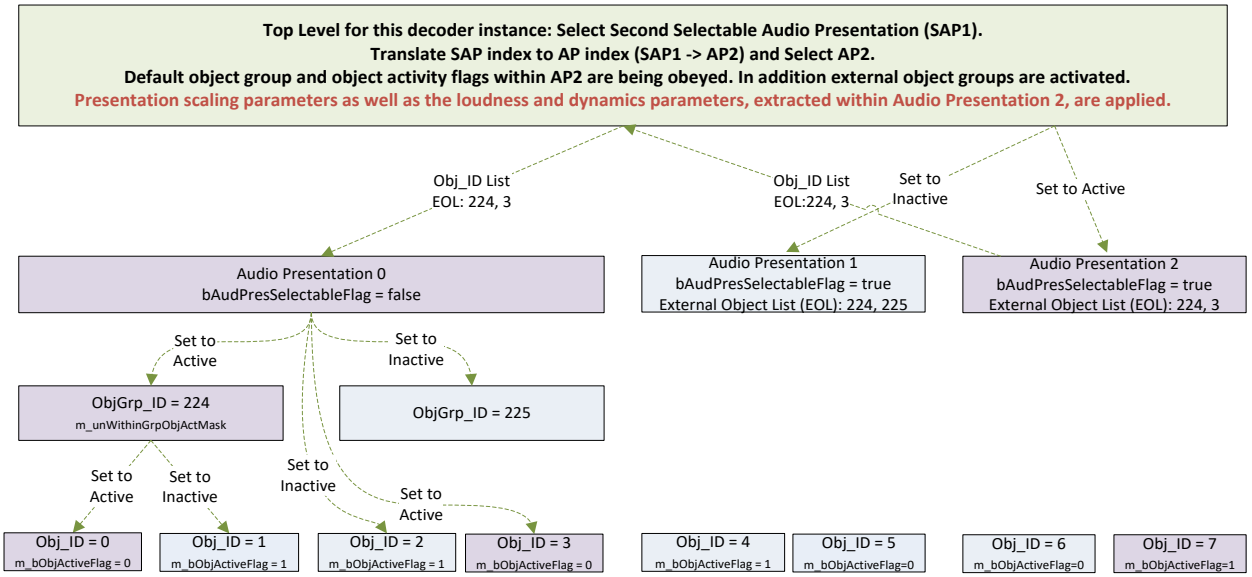
Presentation formed from object groups and objects



The final example shows playback requiring selected presentations which override the default playback.

FIGURE 33

Presentation formed from presentations with higher index

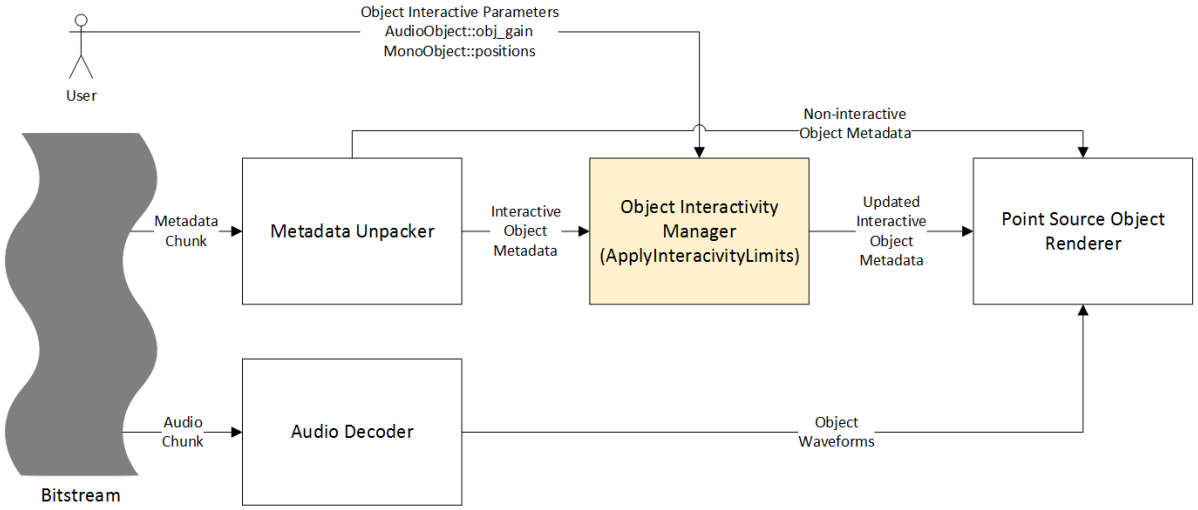


4 Personalization

Metadata in the DTS-UHD bitstream may be controlled by the user for a number of different purposes outlined below. These can be to change the relative level of a particular ‘track’ or object; or to change from one particular object, such as a commentary, to another; or provide an additional track such as Audio Description. In addition, the changes to the metadata and therefore the audio can be limited or disabled. The limits can be specified by the content creator and captured in the original bitstream metadata. The diagram below gives an outline of how the Object Interactivity Manager fits into the system. This object interactivity manager applies user changes to the metadata before delivering the bitstream to the renderer.

FIGURE 34

Object interactivity manager



4.1 Multi language

The ability to control either single audio objects or use lower order presentations as part of a main presentation opens up significant opportunities in the delivery of content in multiple languages. In the first instance audio commentary for sports event could be delivered in multiple languages by changing the active commentary component within a presentation. This can be seen in the above Fig. 32 whereby the Object_ids would have their states changed from Active to Inactive, and vice versa, changing the different languages. This would allow all other aspects of the audio stream to remain the same, perhaps delivered in a 5.1 surround bed, while having a significant reduction on bandwidth overhead. If such an option was delivered traditionally each language option would require a full 5.1 premixed stream. The bandwidth savings can be seen in Table 4.

TABLE 4

Bandwidth requirements for Multi-language support

Current 5.1 Audio delivery solution		**DTS-UHS Presentation delivery solution**	
5.1English AAC	144 kbit/s	5.1Bed DTS-UHD	144 kbit/s
5.1Spanish AAC	144 kbit/s	English DTS-UHD	48 kbit/s
5.1French AAC	144 kbit/s	Spanish DTS-UHD	48 kbit/s
5.1German AAC	144 kbit/s	French DTS-UHD	48 kbit/s
		German DTS-UHD	48 kbit/s
Total	576 kbit/s	Total	336 kbit/s
Saving		240 kbit/s	

The second instance would allow substituted dialogue to be used for different languages with different presentations. This could allow required changes to the mix to be done in production, with the change to another language instigating a new presentation with the associated metadata for the new mix. Again, this would reduce capacity requirements as each separate mix would not need to be broadcast, just the presentation audio with metadata. This would lend itself to recorded content such as Drama or even cinematic content for distribution over multiple countries.

4.2 Sports personalization

As stated above a basic ability of DTS-UHD might be to change a commentary or language. In sport there are multiple examples where this could increase the viewer engagement significantly.

For a football match, this would allow a viewer to receive a commentary that could be delivered from partisan commentators, providing a perspective of the viewers own supported side. In addition, the viewer may have the ability to either control the volume level of the referee, or the crowd. This may also lead to providing a track formed from the crowd supporting the viewers home team, immersing them in the sound of their fellow supporters.

Other Sports such as Formula 1 Motor racing would also benefit from the ability to turn on different audio tracks or objects. Currently viewers are provided with a limited number of radio messages between drivers and their teams, many of which in real time are actually ongoing throughout the race. With personalized audio a viewer would be able to choose a single team or driver and listen solely to their communications. Using presentations and metadata this could continue in conjunction with the standard commentary allowing the dipping in volume of the commentary when audio from the radio communications was detected.

4.3 Accessibility services

When delivering accessibility services, the DTS-UHD audio format provides additional user control to enhance and improve the experience of the user.

4.3.1 Dialogue enhancement

There are sometimes issues raised by viewers about the audibility of the dialogue track within a programme. This can be something that occurs through creative intent or can be due to the viewer having a hearing impairment that makes the separation of sounds more difficult. In addition, there are now many listening environments that may be used for reproduction of content, not only due to the rooms that may be used but also the viewers equipment. Consumers now may have a TV with built-in speakers, a high end soundbar with or without subwoofer or a fully immersive 7.1.4 home cinema. This can mean that in reproduction prominence of the dialogue in a mix may be reduced or that a single mix is just not suitable for all environments. In addition, this can be helpful for viewers listening to different dialects or languages or listening in noisy environments. At its most basic level, the ability to raise the level of the dialogue track or object or increase its prominence in the mix by lowering accompanying effects or music can go some way to solving this problem. With DTS-UHD this can be achieved through use of the Object interactivity manager. In order to maintain artistic intent, limits on the relative volume of a particular track can be put in place when the audio is mixed or it is possible to provide multiple preset mixes that could be chosen by the viewer depending not the environment they are in.

4.3.2 Audio description

Audio description has been supported in channel-based audio within broadcast for a number of years. The addition of an audio description track to the main programme audio has been achieved by two methods: broadcaster mix, and receiver mix. Broadcaster mix requires the delivery of two audio mixes of the programme content. One mix will be the standard service, and the second consists of the main programme with an additional audio description track, with the required level fades within the mix. This method of delivery for audio description does require additional bandwidth as the complete audio mix is effectively delivered twice.

Receiver mix differs in that the audio description track is delivered as a single track. In the event that audio description is desired the receiver will perform the required mix with the main programme audio, using additional metadata to perform the required fades to ensure audio description audibility.

DTS-UHD provides greater control over the audio description track, most importantly allowing the user to move this track to a different point within the soundfield. This spatial separation of the audio description track from the rest of the audio can help differentiate it from the rest of the audio mix and can allow for the track to perhaps be at a lower level causing less intrusion for other viewers.

5 Dynamic management

5.1 Loudness

DTS-UHD allows both the user and content author to manage the loudness of content. This ensures the end user receives a uniform target loudness regardless of the incoming content loudness while maintaining as much as possible the original dynamic range of the content.

The DTS-UHD elementary stream is capable of carrying multiple loudness parameter sets, some of which include (nominally) the complete presentation, the speech components only, and composition of all components excluding the speech. The decoder can output any reasonable loudness level, e.g. from −31 to −16 LKFS. The system may apply DRC accordingly with reference to the output loudness.

The long-term loudness measure is associated with an asset type. Table 5 shows the different asset types associated with the long-term loudness measure, this is a measure of the long-term loudness of all objects, of this asset type, within the audio presentation.

TABLE 5

Defined asset types for long-term loudness measure

index	m_ucAssociatedAssetType
0	ASSET_TYPE_UNKNOWN
1	ASSET_TYPE_COMPLETE_AUDIO_PRESENTATION
2	ASSET_TYPE_COMPLETE_DIALOG
3	ASSET_TYPE_COMPLETE_AUDIO_PRES_EXCLUDING_DIALOG
4	ASSET_TYPE_BED_MIX_WITH_DIALOG
5	ASSET_TYPE_BED_MIX_EXCLUDING_DIALOG
6	ASSET_TYPE_DIALOG
7	ASSET_TYPE_MUSIC
8	ASSET_TYPE_EFFECTS
9	ASSET_TYPE_MUSIC_EFFECTS
10	ASSET_TYPE_COMMENTARY
11	ASSET_TYPE_VISUALLY_IMPAIRED
12	ASSET_TYPE_HEARING_IMPAIRED
13	ASSET_TYPE_AMBIENCE
14	ASSET_TYPE_ISOLATED_FOLEY
15	ASSET_TYPE_KARAOKE
16	ASSET_TYPE_NON_DIEGETIC
17	ASSET_TYPE_COMPOSITE_MULTI_SRC
18	ASSET_TYPE_NEARFIELD_BED
19	ASSET_TYPE_SPOKEN_SUBTITLE
20-31	Reserved

A field within the metadata provides an index of the long-term loudness measurement type for the audio, being either ATSC, EBU or ITU.

5.2 Dynamic Range Control (DRC)

Multiple selectable and custom dynamic range compression curves can be associated with an Audio Programme to facilitate adaptation to various listening environments. The presence of a selectable DRC curve is indicated by the bitstream metadata as defined in ETSI TS 103 491 v1.2.1 (2019-05) [1]. Different curves can be used to accommodate various playback environments. These curves are based on the DRC compression types and parameters based on the general symmetry between the amount of boost against attenuation. Specific slow/fast attack and release times are associated with each profile.

TABLE 6

Common DRC curves

DRC curve	Compression type	Boost vs attenuation parameter
Common 1	Low	A
Common 2	Low	B
Common 3	Low	C
Common 4	Medium	A
Common 5	Medium	B
Common 6	Medium	C
Common 7	High	A
Common 8	High	B
Common 9	High	C

For the boost vs attenuation parameter:

– A has less aggressive attenuation to loud content.

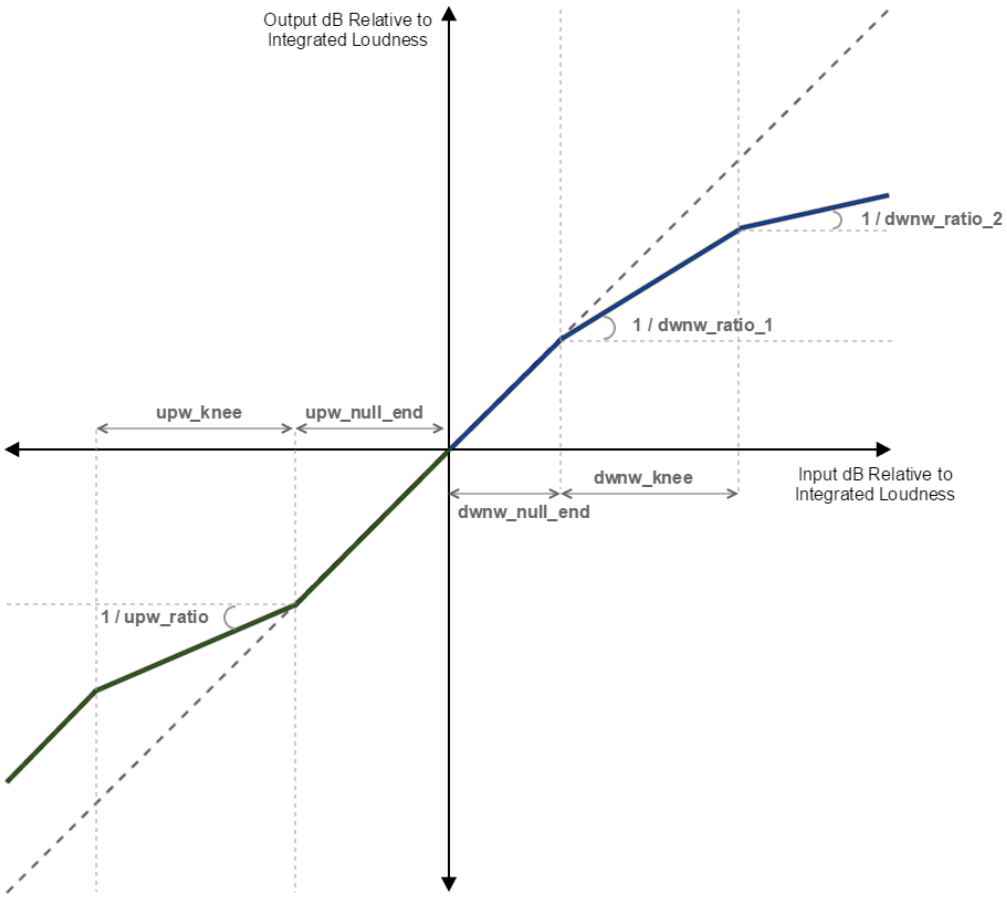
– B has less aggressive boost to quiet content.

– C has equal amount of attenuation and boost.

Other legacy DRC curves are also supported within the system for film, music and speech. Additionally, a fully customized curve can be included in the metadata, whereby a curve type is defined. This curve will then have additional parameters within the metadata to define a piece-wise linear DRC curve explicitly as below, this is further described in ETSI TS 103 491[1].

FIGURE 35

Explicitly defined piece-wise linear DRC curve



6 Multi-stream support

The previously shown presentation showed how to use numerous Objects, Object Groups and Presentations to create the final presentation. However, the presentation may be made up of multiple streams, with a main stream and additional auxiliary streams. For example, a primary audio and video service with the main audio bed, dialogue track and music could be delivered by IP Multicast. Then additional audio elements in the form of language tracks, audio description, etc., could be delivered via a secondary route such as DASH. There are two options available in this instance to decode the streams in this case:

A single decoder may process the audio frames from the various streams as required sequentially, then render all the waveforms from the given time interval together to generate the final output.

Separate decoder instances can be used to decode each stream, with each stream passing the associated metadata to the renderer. In this case the final rendering metadata for scaling the output shall always be provided by the highest ordered elementary stream in the sequence that contains such metadata.

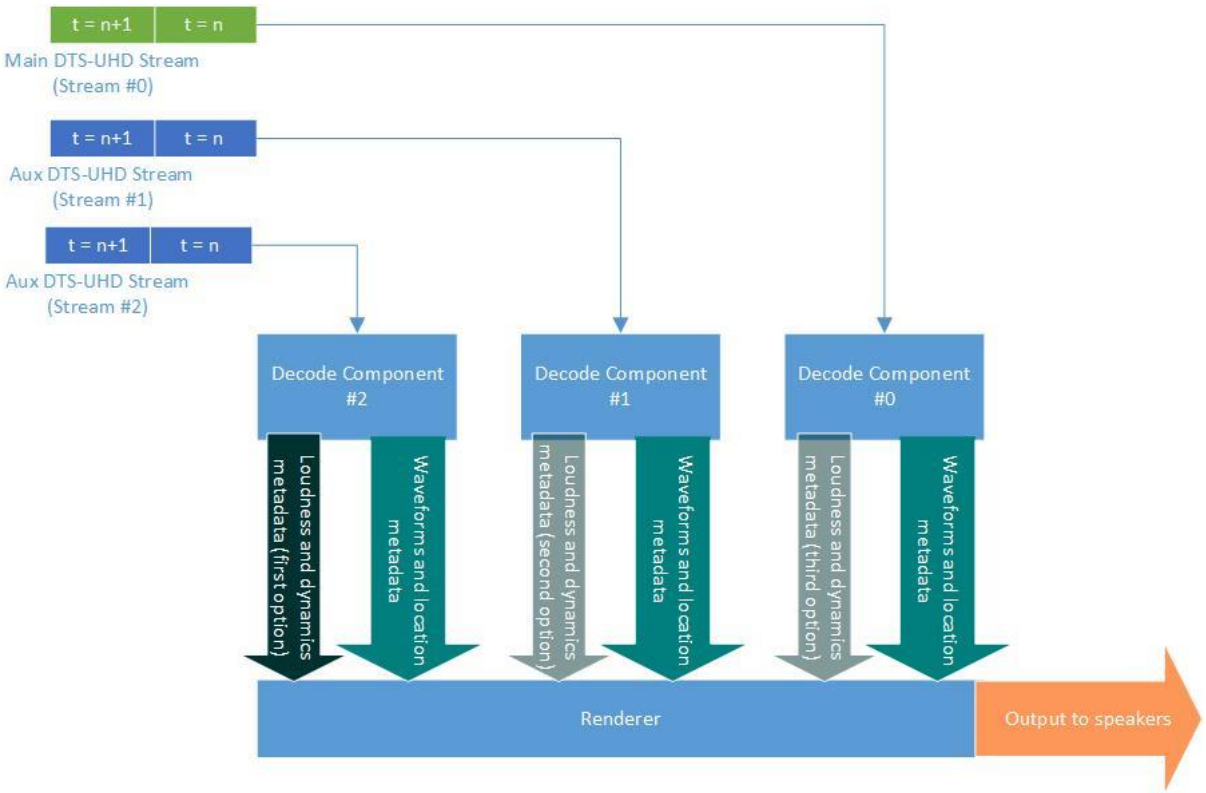
In the example below three elementary streams contribute to a particular preselection. Component #2 is from the highest ordered stream in a multi-stream preselection. The renderer will first look for metadata from Component #2 to perform the final scaling of the mix. If some metadata is missing,

then the renderer looks at the metadata delivered with Component #1, and finally Component #0, in order, to fill in the missing metadata.

To illustrate this example, consider that the component from elementary stream #0 carries music and effects, the component from elementary stream #1 carries dialogue, and the component from elementary stream #2 adds spoken subtitles. Multiple dialogue objects might be able to use the same music and effects, so the mixing metadata with the dialogue will be preferred when only these two components are selected. Since the spoken subtitle is stored in stream #2 and was mastered with the M&E plus dialogue, it was the only one mastered with the awareness of the other components. Therefore, the metadata in Component #2 can provide the best experience. In some scenarios, new mixing metadata may not be generated with the spoken subtitle, i.e. it was mastered in consideration of the stream #1 metadata. In this case, stream #1 metadata will be used for the final rendering.

FIGURE 36

Multi-stream support



References

[1] ETSI TS 103 491 V1.2.1 (2019-05) “DTS-UHD Audio Format; Delivery of Channels, Objects and Ambisonic Sound Fields”
https://www.etsi.org/deliver/etsi_ts/103400_103499/103491/01.02.01_60/ts_103491v010201p.pdf

[2] ETSI TS 101 154 v2.6.1 (2019-09), “Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcasting Application based on the MPEG-2 Transport Stream”
https://www.etsi.org/deliver/etsi_ts/101100_101199/101154/02.06.01_60/ts_101154v020601p.pdf

[3] ANSI/SCTE 242-4 2018, “Next Generation Audio Coding Constraints for Cable Systems: Part 4-DTS-UHD Audio Coding Constraints”
https://scte-cms-resource-storage.s3.amazonaws.com/ANSI_SCTE%20242-4%202018.pdf

[4] ANSI/SCTE 243-4, "Next Generation Audio Carriage for Cable Systems: Part 4 – DTS-UHD Audio Carriage Constraints"
https://scte-cms-resource-storage.s3.amazonaws.com/ANSI_SCTE%20243-4%202018.pdf

[5] ETSI TS 103 584 v1.1.1(2018-01) "DTS-UHD Point Source Renderer"
https://www.etsi.org/deliver/etsi_ts/103500_103599/103584/01.01.01_60/ts_103584v010101p.pdf

Annex 4

Application of the Audio Vivid format

Introduction

With the popularization of advanced sound systems for broadcasting, streaming services and 5G technologies together continued improvement of consumer sound reproduction systems and ultra-high definition (UHD) displays, there is a need for end-to-end sound systems capable of delivering the creative intent of programme makers and meeting audience expectations for advanced audio reproduction.

Compared with dual-channel stereo and multi-channel surround sound, 3D or advanced audio systems provide a richer spatial sound field and a sense of immersion. More immersive sound is one of the core experiences of UHD technology and a key component of spatial audio and virtual reality experiences. It includes not only sound channel information, but also object and scene information. It requires end-to-end collaboration of sound acquisition, post-production, rendering, encoding, distribution (including international content exchange) and reproduction technologies to provide optimal auditory experience.

Audio Vivid is a new generation audio encoding format released and used in China. It is a next-generation high-efficiency compression method for immersive audio in a wide variety of scenarios. It can be applied in home environments, broadcasting, on-demand, OTT, cinema, AR/VR applications and automotive audio systems.

Detailed technical information of the system specification (including encoding, decoding, bitrates and bitstream syntax and semantics) can be found in GY/T 363-2023 *3D audio coding and rendering* [1].

1 Characteristics of Audio Vivid

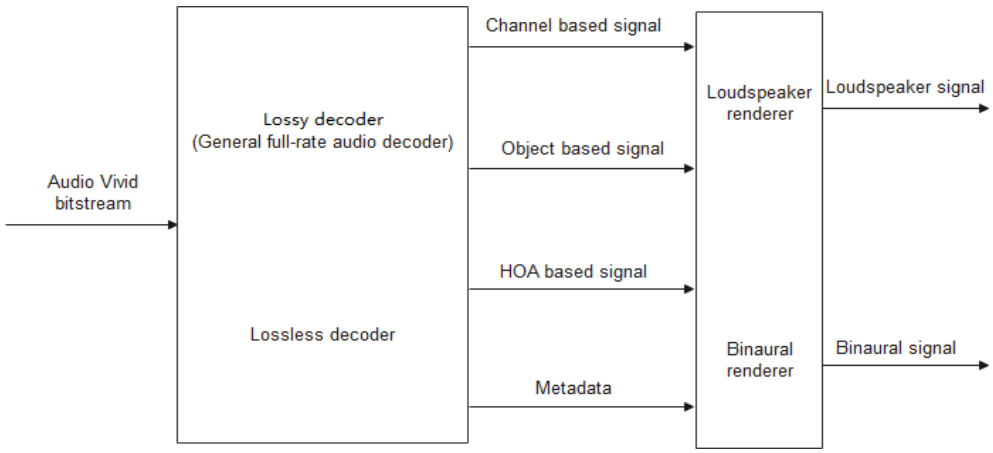
1.1 Overview of the coding system

The Audio Vivid coding system includes lossy coding, lossless coding, loudspeaker renderer and binaural renderer. Supported signal formats include channel-based signal, object-based signal, higher-order ambisonics (HOA) based signal along with metadata.

Figure 37 shows the framework of the Audio Vivid decoding system.

FIGURE 37

Framework of the Audio Vivid decoding system

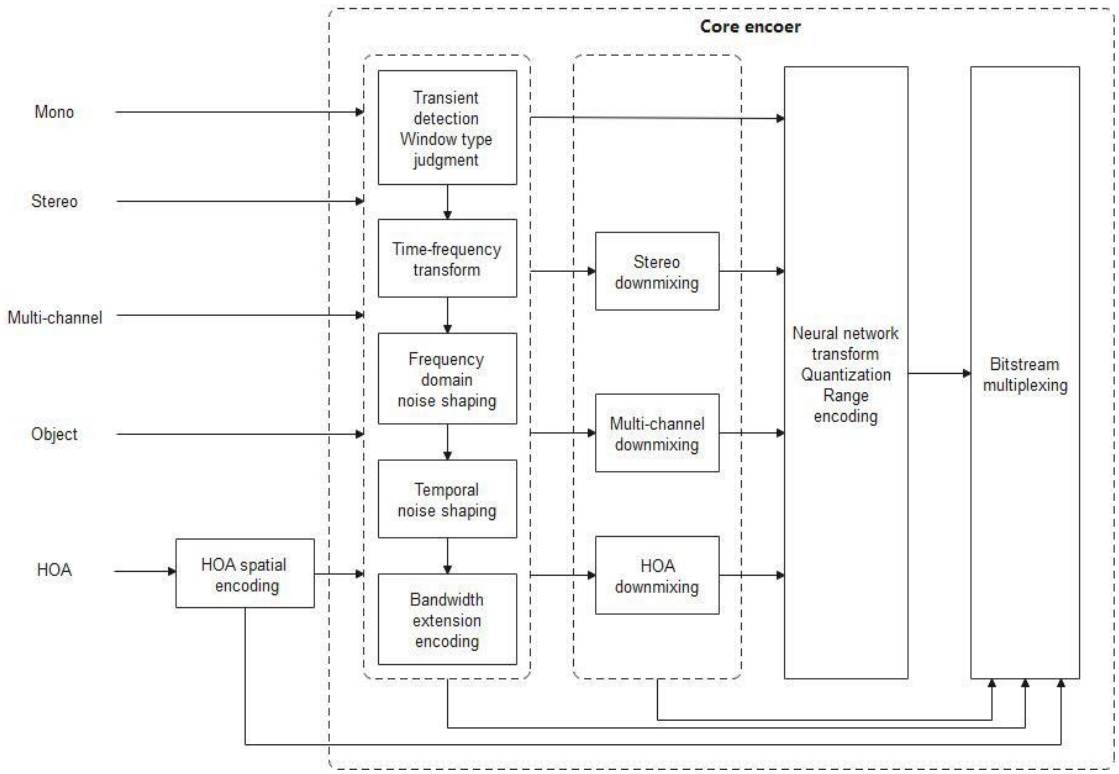


1.2 General Full-Rate Audio Encoding

A key component of the Audio Vivid system is General Full-Rate Audio Encoding. Figure 38 shows the basic architecture of a General Full-Rate Audio coder. It supports the mono, stereo, multi-channel, object, HOA along with metadata at the input and capable of handling sampling rate from 32 kHz to 192 kHz, and coding rates from 32 kbit/s to 1.6 Mbit/s. The audio encoder is compatible with 16 to 24-bit audio and supports mono, stereo, and surround sound formats. The general full-rate audio encoder consists of transient state detection, window type judgment, time-frequency transform, frequency-domain noise shaping, temporal noise shaping, bandwidth extension, downmixing, neural network transform, quantization, and context encoding. It encodes channel signals and object signals into bitstreams. The HOA spatial encoder and core encoder encode HOA signals into bitstreams.

FIGURE 38

Architecture of General Full-Rate Audio Coding



1.3 3D audio coding based on the use of Artificial Intelligence

Audio Vivid uses hybrid AI coding architecture to improve audio compression efficiency. Traditional encoding and decoding technologies are used in pre-processing and AI-based technologies are used for feature transformation, quantization and entropy encoding process.

This hybrid AI architecture not only combines the essence of traditional audio compression theory (psychoacoustics theory) with the advantages of deep learning to extract abstract features, but also strikes a reasonable balance between performance and processing overhead.

In this architecture, an audio signal is converted from time domain to frequency domain using Modified Discrete Cosine Transform (MDCT) during the pre-processing stage. Frequency domain noise shaping and time domain noise shaping are applied to the MDCT signal. Any required downmix processing can be performed on the MDCT signal before it proceeds to the AI processing phase.

A deep neural network is used during the AI processing phase to convert the MDCT signal into an intermediate 'feature' signal which is then scalar quantized and sent to an AI-based entropy coding module. The purpose of the intermediate feature signal is to identify features that are more favourable for efficient entropy coding. The entropy encoding module generates a context of the to-be-encoded implicit feature signal using a second deep neural network and selects a corresponding codebook according to the context, to perform entropy encoding on the implicit feature signal.

The two deep neural networks are jointly trained building a relationship between the to-be-encoded feature and its context. Each codebook is jointly searched under the constraint of minimizing information entropy, thereby fully utilizing a powerful abstraction capability of the deep neural network.

To facilitate deployment of the decoder on multiple platforms, especially mobile platforms, and to minimize storage and calculation overheads of the decoder, the two deep neural networks are designed asymmetrically at the encoder and decoder stages. The encoder uses a larger neural network to ensure relatively high compression efficiency, while the decoder uses a smaller neural network to reduce the overhead.

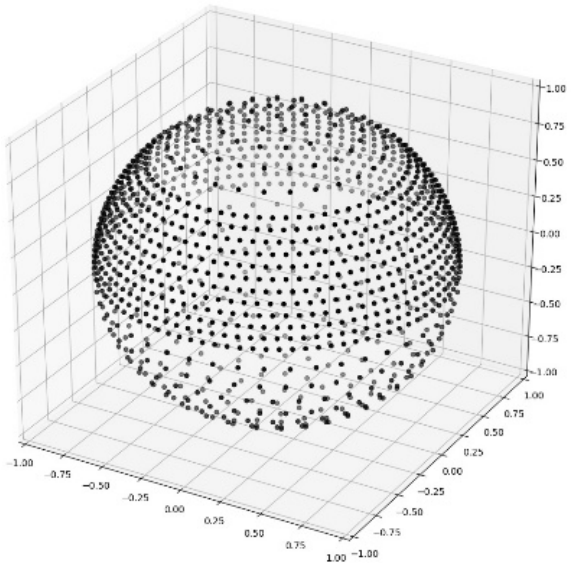
1.4 High Order Ambisonics – spatial coding based on virtual loudspeaker projection

Audio Vivid greatly improves the coding efficiency of HOA signals by providing an HOA spatial encoding algorithm. This algorithm assumes that several virtual speakers are distributed around a scene. Figure 39 shows an example of virtual loudspeaker distribution.

The HOA signal is approximated by a linear combination of several virtual loudspeaker signals, the difference between input signal and the reconstruction signal which is recovered by the selected virtual loudspeakers. Using selected virtual loudspeakers to represent the sound source in the scene minimizes residual signal difference between input signal and the reconstruction signal. With HOA spatial coding, the input signal is transformed into a few virtual loudspeaker signals, residual signals and side information, then sent to the encoder. The number of the virtual loudspeaker signals and residual signals is far less than the quantity of input signals, and the residual signal can be coded by fewer bits, therefore greatly improves the HOA coding efficiency.

FIGURE 39

Virtual loudspeaker distribution



1.5 Flexible metadata system

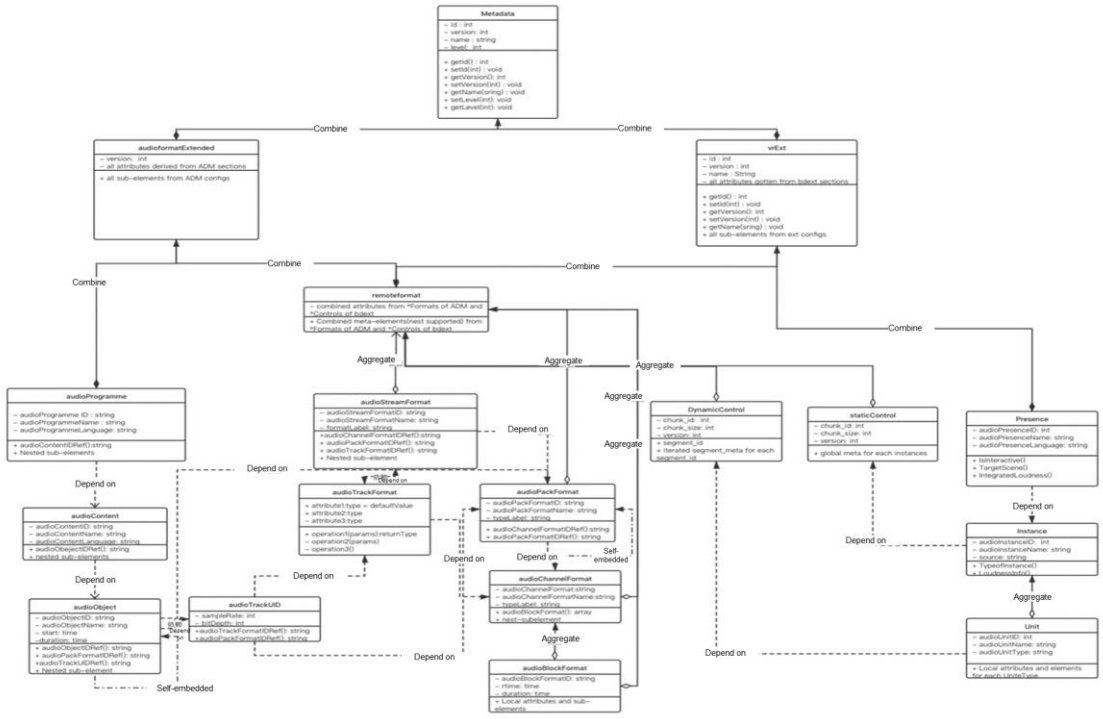
The metadata system is an important feature of the Audio Vivid system. Metadata describes audio signal format, audio signal content, a physical attribute of the audio signal in a playback space, a sound effect parameter of the audio signal, the mechanism for user interaction with the audio signal, and so on.

The Audio Vivid metadata system is designed as a hierarchical structure consisting of a base layer and an extension layer. The base layer of metadata multiplexes the attributes and elements of the audio signal content and format defined in Recommendation ITU-R BS.2076 (ADM). It is used to deliver the content and control information related to audio types such as direct speaker, matrix, object, HOA, and binaural.

The extension layer of metadata provides enhanced binaural rendering features and supports binaural rendering to better restore a director's artistic intent through metadata such as acoustic environment and sound effect post-processing used in rendering. Audio Vivid metadata base layer conformance with Recommendation ITU-R BS.2076 provides interoperability with the extension layer providing flexibility and extensibility. This dual-layer approach provides a powerful representation and interaction capabilities for the new generation of audio systems. The metadata system architecture and detailed description refer to Fig. 40.

FIGURE 40

Audio Vivid metadata system architecture



2 Information on the application of Audio Vivid

2.1 Rollout of Audio Vivid

"Hundred Cities and Thousand Screens" is a UHD video promotion activity jointly deployed by the Ministry of Industry and Information Technology, the Ministry of Transport, the Ministry of Culture and Tourism, the State Administration of Radio and Television, and the China Media Group (CMG).

At present, more than 100 UHD screens have been deployed in 35 cities across the country. The public can watch China Media Group 8K UHD TV channel on the public screens. At the same time, CMG provides the audio service for the UHD videos through mobile terminals on the basis of the original UHD videos without causing sound interference.

In August 2022, the headquarters used the Audio Vivid technology to listen to 3D audio with accurate synchronization between audio and video through the mobile application Ethereal Sound. On 10 September 2022, the CMG Mid-Autumn Festival Gala was broadcast simultaneously by Audio Vivid through the promotion "Hundred cities and Thousand screens" for the first time, bringing an audio-visual feast to the public.

2.2 Audio Vivid Application in the Thousands of Screens in Hundreds of Cities Program

2.2.1 Introduction to the Thousands of Screens in Hundreds of Cities Program

The Thousands of Screens in Hundreds of Cities Program is a campaign to promote UHD video. To date, more than 100 UHD screens have been installed in 35 cities across China, allowing the general public to watch the China Media Group's 8K UHD TV channels. An audio service – Portable Audio – is simultaneously being provided for mobile devices, in order to better promote UHD videos in the Thousands of Screens in Hundreds of Cities Program. This means users can enjoy audio services anytime and anywhere while watching UHD videos on large outdoor screens, free from sound disturbance. In August 2022, supported by Audio Vivid technologies, the China Media Group

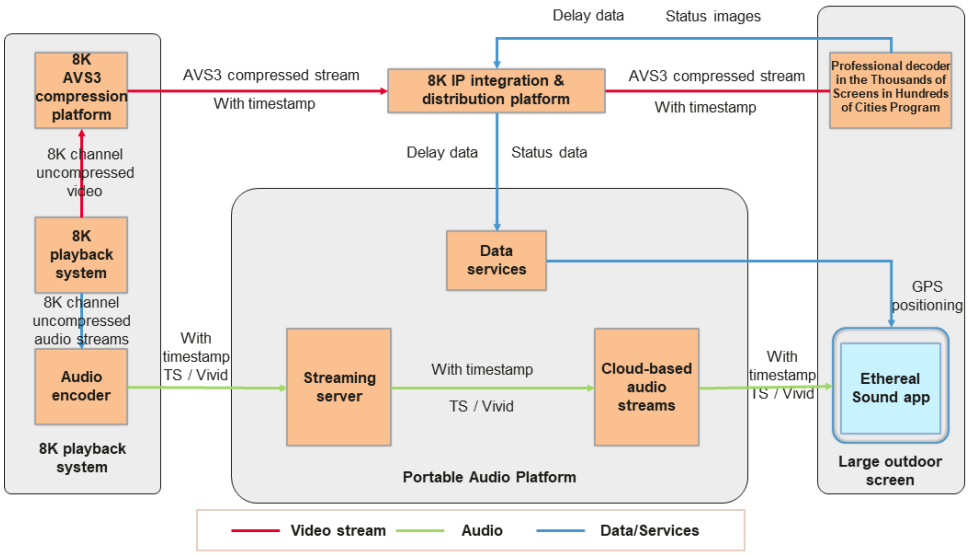
allowed users of the Portable Audio service to receive 3D signals that perfectly synced audio and video. On 10 September 2022, for the first time, the China Media Group Mid-Autumn Festival Gala was broadcast simultaneously to screens installed as part of the Thousands of Screens in Hundreds of Cities Program.

2.2.2 Portable Audio Platform in the Thousands of Screens in Hundreds of Cities Program

The primary purpose of the Portable Audio Platform is to support the encoding and transmission of television audio to mobile devices, thereby allowing the audio and video content of TV channels to be transmitted separately but presented in sync. This can greatly improve the viewing experience of users watching large outdoor screens. The audio service provided by the Portable Audio Platform was upgraded from stereo to Audio Vivid 3D audio on 18 August 2022, marking the beginning of 3D audio's commercial use in live broadcasts. Figure 41 shows the Portable Audio Platform's system architecture.

FIGURE 41

Portable Audio system architecture



The Portable Audio Platform uses Audio Vivid throughout the whole process, from encoding and encapsulation to transmission and distribution, and supports playback on a mobile device that integrates the Portable Audio service after decoding and binaural rendering.

2.2.3 Portable Audio Service in the Thousands of Screens in Hundreds of Cities Program

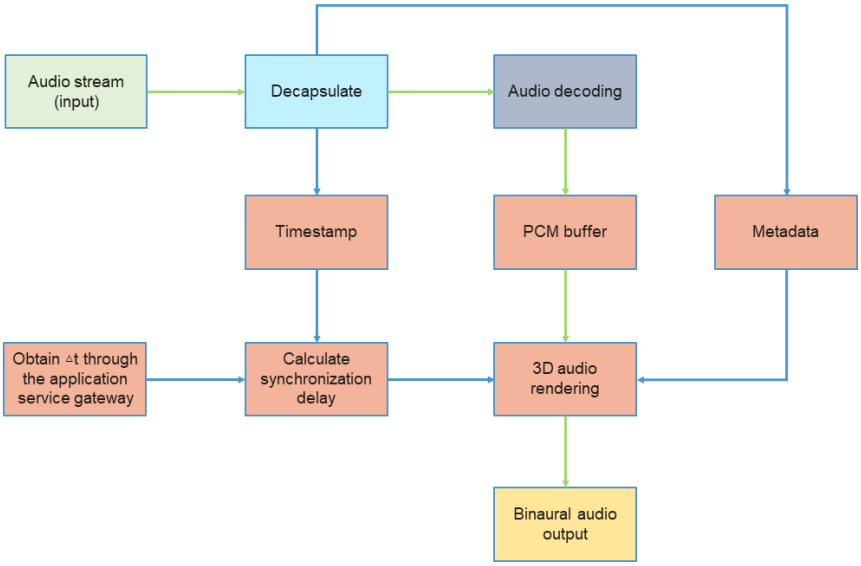
The Portable Audio service is integrated within the Ethereal Sound app. Ethereal Sound is a mobile audio app launched by the China Media Group based on 5G transmission technology. Focusing on content like news, general knowledge, and culture, Ethereal Sound collects high-quality content from the China Media Group and offers self-made audio programs and high-quality audio books. Ethereal Sound is committed to providing full-scenario audio services for a variety of users, including those that use mobile phones, head units, tablets and smart wearables.

In the Thousands of Screens in Hundreds of Cities Program, users can link the Ethereal Sound app to a designated large screen by scanning the relevant QR code or through manual selection. They can then obtain the audio stream URL for that large screen and the corresponding playback delay caused by video decoding through the China Media Group's media gateway interface. Following this, audio stream data that is encoded using Audio Vivid and encapsulated in transport streams is read using the obtained URL. Finally, Audio Vivid audio signals are played in sync with that large screen through

steps like encapsulation, decoding, and rendering. Specifically, the playback process includes the following steps show in Fig. 42:

FIGURE 42

The Ethereal Sound App playback workflow



2.3 Audio Vivid Application for OTT applications

Internet video content service companies such as Tencent Video, Migu Video, CMG Mobile, Himalayan Audiobook and Huawei Music have integrated Audio Vivid into their respective products to be able to offer Audio Vivid formatted content.

Tencent Video officially launched “Zhenyue Panoramic Sound” in July 2024, offering the audience very high quality three-dimensional sound using Audio Vivid encoding. The Tencent Video application adopted an efficient software decoder providing-Audio Vivid content on Android, iOS, and HarmonyOS smartphones covering popular movies and TV series. To date, the total duration of Audio Vivid format encoded content is approximately 70 hours. More Audio Vivid three-dimensional sound content will continue to be launched with the expectation that all Tencent Video high-quality original productions will be available in Audio Vivid by 2025. The current Tencent Video audio interface is shown in Fig. 43.

China Mobile Migu has applied Audio Vivid for live broadcasts of major sports events, The current state of Migu Video audio interface is shown in Fig. 44.

Live Audio Vivid broadcast content by China Mobile Migu

– November 2022 – Qatar World Cup (this was the first use of Audio Vivid for World Cup broadcasts.
– July 2023 – Chengdu Universiade and September 2023 – Hangzhou Asian Games, live broadcasts using Audio Vivid.

FIGURE 43

The current state of Tencent Video

FIGURE 44

The current state of Migu

2.4 In car Audio Vivid Application

Ximalaya (from May 2024) and NetEase Cloud Music (from July 2024) support in-car Audio Vivid broadcasting. Ximalaya has already launched 16000 minutes of dramatized audio books, including the popular “The Three-Body Problem”, “Dune” and “The Wandering Earth”. The current Ximalaya interface is shown in Fig. 45.

NetEase Cloud Music has released over 1 million Audio Vivid music tracks using automated conversion, leveraging the Audio Vivid encoding automation tools. It is anticipated that the entire library of NetEase Cloud Music will fully support Audio Vivid by the end of 2024.

Huawei Cloud Music has established a dedicated space for audio (Audio Vivid) content and collaborated with the China Central Philharmonic Orchestra to produce and launch over 2 000 high-quality original Audio Vivid music tracks.

FIGURE 45

The current state of Ximalaya

References

[1] GY/T 363-2023 3D audio coding and rendering, https://uhd-world-association.com/uwa-resources/