INTERNATIONAL STANDARD

ISO/IEC 23008-3

First edition
2015-10-15

Corrected version
2016-03-01

Information technology — High efficiency coding and media delivery in heterogeneous environments —

Part 3:
3D audio

Technologies de l'information — Codage à haute efficacité et livraison des medias dans des environnements hétérogènes —

Partie 3: Audio 3D

Reference number
ISO/IEC 23008-3:2015(E)

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the WTO principles in the Technical Barriers to Trade (TBT) see the following URL: Foreword - Supplementary information

The committee responsible for this document is ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

ISO/IEC 23008 consists of the following parts, under the general title *Information technology — High efficiency coding and media delivery in heterogeneous environments*:

— *Part 1:* MPEG media transport (MMT)

— *Part 2:* High efficiency video coding

— *Part 3: 3D audio*

— *Part 4: MMT Reference and Conformance Software*

— *Part 5: Reference software for high efficiency video coding*

— *Part 8: HEVC conformance testing*

— *Part 10: MPEG media transport forward error correction (FEC) codes*

— *Part 11: MPEG media transport composition information*

— *Part 12: Image file format*

— *Part 13: MMT Implementation Guidelines*

This corrected version of ISO 23008-3:2015 incorporates the following corrections: unreadable equations have been improved.

Introduction

3D sound systems are able to realize a significantly enhanced sound experience relative to current widespread 5.1 channel audio programs and playback systems. These systems demand high quality audio coding and error-free transmission in order to keep the timbre, sound localization and sound envelopment of the original audio program. Presentation over headphones with suitable spatialization are also considered.

This part of ISO/IEC 23008-3 "High Efficiency Coding and Media Delivery in Heterogeneous Environments — Part 3: 3D Audio" provides means for all scenarios where there is a need to compress a multi-channel audio program (e.g. 22.2 channel program) and to render it to the native target number of loudspeakers. In order to reach a wide market, a 3D Audio program is able to be downmixed to a lower hierarchy of loudspeakers, for example 10.1 or 8.1 channels. In addition, all scenarios support a level of random access to facilitate broadcast break-in, and "trick modes" such as fast forward when playing from stored media.

The main focus of this specification are applications such as audio for Home Theatres where the audio presentation is immersive, involving many loudspeakers (e.g. from 10 to more than 20) surrounding the listener and placed below, at and above ear-level. Moreover applications as Personal TV, TV for SmartPhones and Multi-channel Audio-only Programs are envisioned. These require that 3D Audio encoding/decoding systems are able to operate at low bitrates appropriate for efficient transmission over a cellular channel. At the same time the sense of envelopment and accurate sonic localization even for systems having a tablet-sized visual displays with speakers built into the device and headphone listening are maintained.

Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 3: 3D audio

1 Scope

This part of ISO/IEC 23008-3 specifies technology which supports the efficient transmission of 3D audio signals and flexible rendering for the playback of 3D audio in a wide variety of listening scenarios. These include 3D home theater setups, 22.2 loudspeaker systems, automotive entertainment systems and playback over headphones connected to a tablet or smartphone.

2 Normative references

The following documents, in whole or in part, are normatively referenced in this document and are indispensible for its application. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 13818-1:2013, *Information technology — Generic Coding of moving pictures and associated audio information: Systems*

ISO/IEC 14496-3:2009, *Information technology — Coding of audio-visual objects — Part 3: Audio*

ISO/IEC 14496-11, *Information technology — Coding of audio-visual objects — Part 11: Scene description and application engine*

ISO/IEC 23001-8:2013, *Information technology — MPEG systems technologies — Part 8: Coding-independent code-points*

ISO/IEC 23001-8:2013/Amd.1, *Information technology — MPEG systems technologies — Part 8: Coding-independent code-points, AMENDMENT 1: New audio code points*

ISO/IEC 23003-1:2007, *Information technology — MPEG audio technologies — Part 1: MPEG Surround*

ISO/IEC 23003-2:2010, *Information technology — MPEG audio technologies — Part 2: Spatial Audio Object Coding (SAOC)*

ISO/IEC 23003-3:2012, *Information technology — MPEG audio technologies — Part 3: Unified speech and audio coding*

ISO/IEC 23003-4:2015, *Information technology — MPEG audio technologies — Part 4: Dynamic range control*

3 Terms, definitions and mnemonics

3.1 Terms and Definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 14496-3:2009, 1.3 (Terms and definitions), in ISO/IEC 14496-3:2009, 1.4 (Symbols and abbreviations) and in ISO/IEC 23003-3:2012, 3.1 (Terms and definitions) apply.

3.2 Mnemonics

The following mnemonics are defined to describe the different data types used in the coded bitstream payload.

bslbf	Bit string, left bit first, where "left" is the order in which bit strings are written in ISO/IEC 14496. Bit strings are written as a string of 1s and 0s within single quote marks, for example '1000 0001'. Blanks within a bit string are for ease of reading and have no significance.
uimsbf	Unsigned integer, most significant bit first.
vlclbf	Variable length code, left bit first, where "left" refers to the order in which the variable length codes are written.
tcimsbf	Two's complement integer, most significant (sign) bit first.

4 Technical Overview

4.1 Decoder block diagram

The 3D Audio Codec System consists of an MPEG-H 3D Audio Core Codec for coding of channel, object and Higher Order Ambisonics (HOA) signals. The core codec is based on the MPEG-D USAC codec. To increase the efficiency for coding a large amount of objects, MPEG SAOC technology has been adopted. Several types of renderers perform the tasks of rendering objects to channels, rendering channels to a different loudspeaker setup, rendering HOA signals to the loudspeaker setup or rendering virtual loudspeaker channels or HOA components to headphones.

When object signals are explicitly transmitted or parametrically encoded using SAOC, the corresponding Object Metadata information is compressed and multiplexed into the 3D-Audio bitstream.

Figure 1 shows the different algorithmic blocks of the 3D-Audio system.

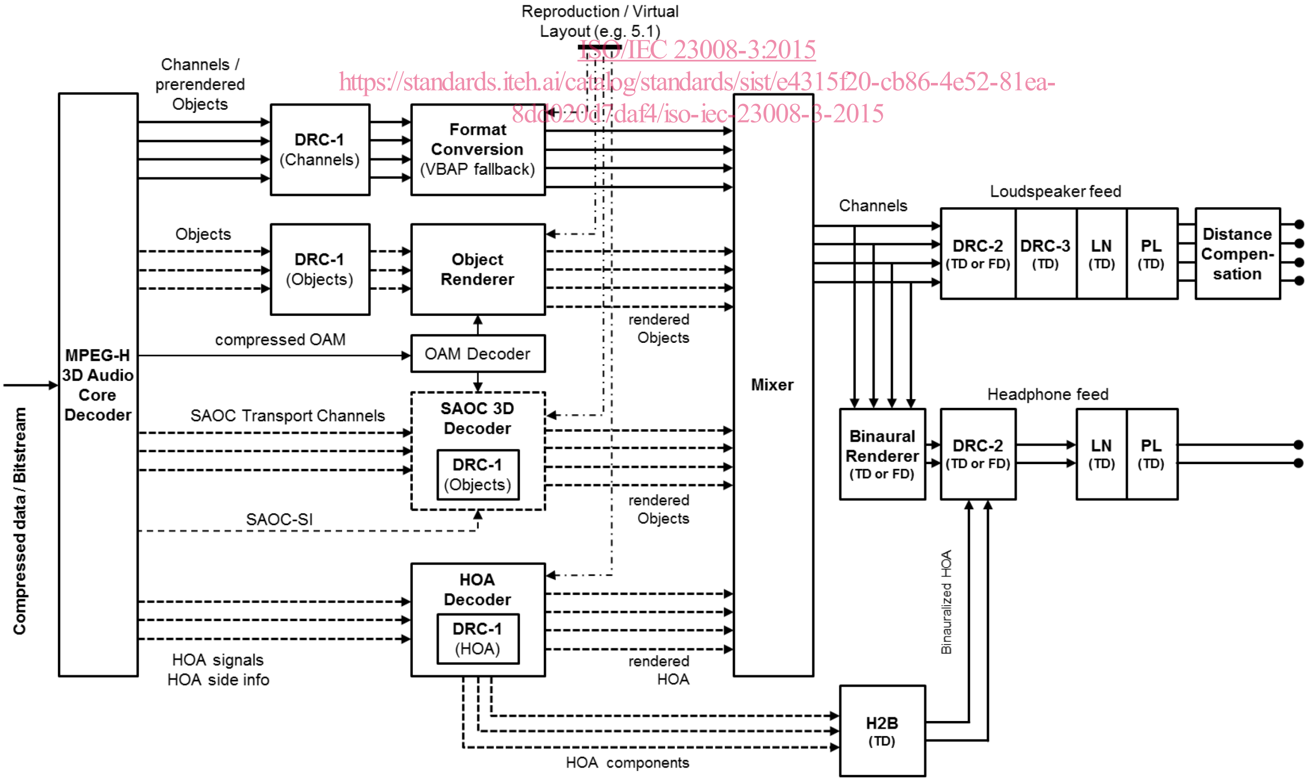

Figure 1 — Block diagram of the 3D-Audio decoder.
(DRC: Dynamc Range Control, SAOC: Spatial Audio Object Coding, HOA: Higher Order Ambisonics, LN: Loudness Normalization, PL: Peak Limiter)

4.2 Overview over the codec building blocks

The MPEG-H 3DA Core Codec for loudspeaker-channel signals, discrete object signals, object downmix signals and pre-rendered signals is based on MPEG-D USAC technology. It handles the coding of the multitude of signals by creating channel- and object-mapping information based on the geometric and semantic information of the input's channel and object assignment. This mapping information describes how input channels and objects are mapped to channel elements (CPEs, SCEs, LFEs) and the corresponding information is transmitted to the decoder.

The coding of objects is possible in different ways, depending on the rate/distortion requirements and the interactivity requirements for the renderer. The following object coding variants are possible:

- Prerendered objects: Object signals are pre-rendered and mixed to multi-channel or HOA signals before encoding, as appropriate. The subsequent coding chain then operates on multi-channel or HOA signals.
- Discrete object waveforms: Objects are supplied as monophonic waveforms to the encoder. The encoder uses single channel elements SCEs to transmit the objects in addition to the channel signals. The decoded objects are rendered and mixed at the receiver side. Compressed object metadata information is transmitted to the receiver/renderer alongside.
- Parametric object waveforms: Object properties and their relation to each other are described by means of SAOC parameters. The downmix of the object signals is coded with the MPEG-H 3D Audio Core codec. The parametric information is transmitted alongside. The number of downmix channels is chosen depending on the number of objects and the overall data rate. Compressed object metadata information is transmitted to the SAOC renderer.

The SAOC Encoder and Decoder for object signals are based on MPEG SAOC technology. The system is capable of recreating, modifying and rendering a number of audio objects based on a smaller number of transmitted channels and additional parametric data (OLDs, IOCs, DMGs).

The SAOC decoder reconstructs the object/channel signals from the decoded SAOC transport channels and parametric information, and generates the output audio scene based on the reproduction layout, the decompressed object metadata information and optionally on the user interaction information.

The Object Metadata Codec efficiently codes the associated metadata that specifies the geometrical position and volume of each object in 3D space by quantization of the object properties in time and space. The compressed object metadata is transmitted to the receiver as side information.

The Object Renderer utilizes the compressed object metadata to generate object waveforms according to the given reproduction format. Each object is rendered to certain output channels according to its metadata. The output of this block results from the sum of the partial results.

The Loudspeaker Renderer converts between the transmitted channel configuration and the desired reproduction format. It is thus called 'format converter'. In case of conversions to lower numbers of output channels it creates downmixes. The system automatically generates optimized downmix matrices for the given combination of input and output formats and applies these matrices in a downmix process. The format converter allows for standard loudspeaker configurations as well as for random configurations with non-standard loudspeaker positions.

The Higher Order Ambisonics (HOA) Decoder/Renderer reconstructs the HOA coefficient signals based on the HOA transport channels decoded by the 3D Audio Core Decoder and the HOA specific side information. The coding principle is based on a separate transmission of so-called predominant sounds and ambient sound scene components. Subsequently the HOA renderer generates the loudspeaker channel feeds based on the reproduction layout.

If two or more groups of channel based content, discrete/parametric objects or HOA based content are decoded, the corresponding waveforms are delay-aligned and sample-wise added by the Mixer before providing the resulting waveforms (or before feeding them to a postprocessor module such as the binaural renderer, DRC-2, DRC-3, the peak limiter PL, or the loudspeaker distance compensation).

The Binaural Renderer module produces a binaural downmix of the multichannel audio material, such that each input channel is represented by a virtual sound source. The processing is conducted frame-wise in the QMF domain or in the time domain. The binauralization is based on measured binaural room impulse responses.

Virtual layout information fed from an application shall be consistent with the corresponding BRIR set provided (expressed with MeasurementSetup, see in Table 157), in the sense that the set of positions corresponding to the virtual layout is a subset of the set of positions corresponding to the BRIRs.

4.3 Efficient combination of decoder processing blocks in time domain and QMF domain

There are numerous processing blocks in the MPEG-H 3DA decoding, rendering and processing framework. In general these blocks can be classified into different classes:

— Block operates in time domain (**TD**)

— Block operates in frequency domain (**FD**) – also called QMF domain

— Block is "**neutral**" and can operate in FD or TD.
Signal-processing-wise the same operation is carried out in either FD or TD. Operation in a different domain may cause no difference or only a small but perceptually negligible difference to the output signal

Table 1 lists all blocks with corresponding processing domain. For the sake of clarification the functional blocks are grouped into semantically differentiable contexts, which follow the MPEG-H 3D Audio signal processing as shown in Figure 2.

Table 1 — MPEG-H 3DA functional blocks and internal processing domain

Processing Context	Functional Block	Processing Domain
Audio Core	MPEG-H 3D Audio Core Coder	**FD** or **TD**[1)]
Rendering	DRC-1	if multiband: **FD** else: **neutral**
	Format Converter (FC)	**FD**
	Object Renderer	**neutral**
	SAOC 3D Decoder	**FD**
	HOA Decoder	**TD**
Mixing	Mixer	**neutral**
Post-processing	DRC-2	if multiband: **FD** else: **neutral**
	FD Binauralizer	**FD**
	TD Binauralizer	**TD**
End of chain	DRC-3 (only singleband)	**TD**
	Loudness Normalization	**TD**
	Peak Limiter	**TD**
	LS Distance Compensation	**TD**

[1)] depends on bit stream configuration

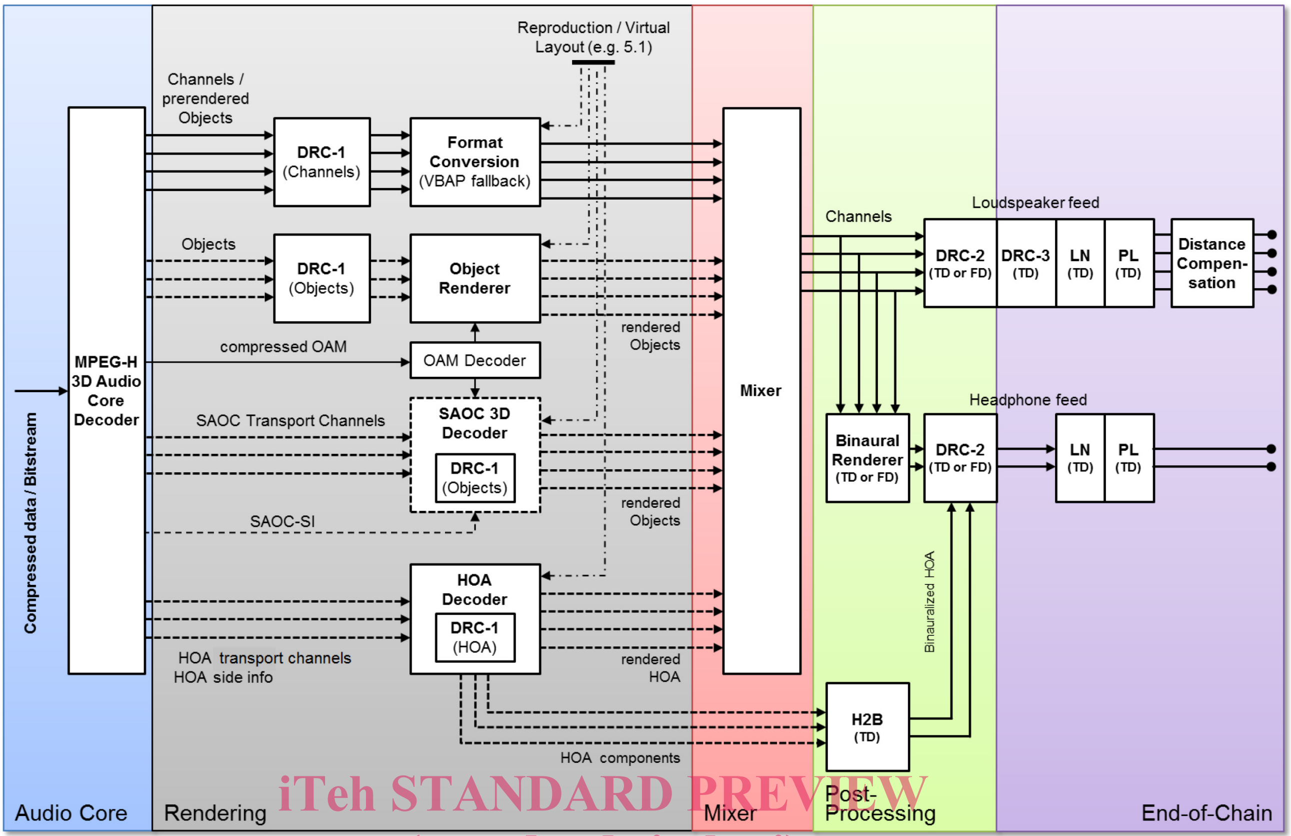

Figure 2 — MPEG-H 3D audio decoder overview with signal processing context

Operating blocks that are directly connected in the signal chain and which operate in the *same* domain can interface in that common domain.

Operating blocks that are directly connected in the signal chain and which operate in *different* domains require a transformation block which turns one signal representation into another, i.e. a QMF Analysis (TD to FD) or a QMF Synthesis (FD to TD). This transform causes additional delay and hence needs to be taken into account when determining the overall signal processing delay.

The following subclause describes formal generic rules which determine how the various functional can be connected.

4.4 Rule set for determining processing domains

4.4.1 Audio Core Codec, Processing Domain

The incoming bit stream configuration determines which core decoder tools are active and whether parts of the core codec operate in QMF domain / FD.

— If last decoding stage is in QMF domain (SBR and/or MPS212 active)
  (typical for mid to low bit rate coding)
  — formal Audio Core Codec processing domain is **FD**

— If last decoding stage is **not** in QMF domain (neither SBR nor MPS212 active)
  (typical for high rate coding)
  — formal Audio Core Codec processing domain is **TD**

