INTERNATIONAL STANDARD

ISO/IEC 23008-3

First edition
2015-10-15

AMENDMENT 3
2017-01

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

Part 3: 3D audio

AMENDMENT 3: MPEG-H 3D Audio Phase 2

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

Partie 3: Audio 3D

AMENDEMENT 3: Phase 2 de l'audio 3D MPEG-H

Reference number
ISO/IEC 23008-3:2015/Amd.3:2017(E)

ISO copyright office
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

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 World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see the following URL: www.iso.org/iso/foreword.html.

Amendment 3 to ISO/IEC 23008-3:2015 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

Introduction

The following text describes the Amendment 3 to the specification ISO/IEC 23008-3:2015 MPEG-H 3D Audio in an "Amendment"-style, i.e. in "Replace A with B"-style. It includes additions and changes that serve a number of purposes:

- improving the coding efficiency especially for low bitrate coding modes (for scene based as well as for object based and for multichannel based content)
- adding descriptive metadata
- updating the MHAS description
- some improvements for usage of MPEG-H 3D Audio in broadcasting applications
- a tool for Advanced Loudness Control
- a layered coding mode for coding of scene based content

It is envisioned that this amendment will be merged with the current version of the MPEG-H 3D Audio specification resulting in a second edition of the standard. Text with yellow highlight shall be adjusted by the editor of a new edition of ISO/IEC 23008-3.

For ease of review the document is structured by clauses, each of which reflect an approved set of changes.

New Clauses, Tables and Figures are typically labelled "AMDX.Y", where X is the number of the clause it appears in in this document and Y is an increasing integer counter.

Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 3: 3D audio, AMENDMENT 3: MPEG-H 3D Audio Phase 2

1 Profiles and Levels

Add the following definition of profiles and levels to clause 4 Technical Overview:

4.X MPEG-H 3D Audio profiles and levels

4.X.1 Introduction

This subclause defines profiles and their levels for MPEG-H 3D Audio.

Complexity units are defined to give an approximation of the decoder complexity in terms of processing power required for the decoding process. The approximated processing power is given in "Processor Complexity Units" (PCU), specified in Millions Operations Per Second (MOPS).

4.X.2 Profiles

The following Audio Profiles are defined:

1. The Main Profile of MPEG-H 3D Audio provides a complete set of features for low-bitrate and high-quality coding, and rendering for all playback scenarios, exclusively based on the first edition of the MPEG-H 3D Audio specification ISO/IEC 23008-3:2015 3D Audio.

2. The High Profile of MPEG-H 3D Audio provides a complete set of features for low-bitrate and high-quality coding, and rendering for all playback scenarios.
   The High Profile is a superset of the Low-complexity Profile.

3. The Low Complexity Profile provides features for broadcasting and streaming with a reduced complexity of the decoder;

Table P1 — Summary of the Location of and Normative Reference to the Definitions of MPEG-H 3D Audio profiles. USAC and MPEG-H 3DA Main Profile are provided for information only

Tool / Module		defined in ISO/IEC	sub-clause	USAC 23003-3	MPEG-H 3DA Main Profile	MPEG-H 3DA High Profile	MPEG-H 3DA Low-Complexity Profile
block switching		14496-3	4.6.11	X	X	X	X
window shapes	AAC based	14496-3	4.6.11	X	X	X	X
	additional windows	23003-3	6.2.9.3	X	X	X	X
filter bank	AAC based	14496-3	4.6.11	X	X	X	X
	additional USAC	23003-3	7.9	X	X	X	X

TNS		14496-3	4.6.9	X	X	X	X
intensity		14496-3	4.6.8.2				
coupling		14496-3	4.6.8.3				
perceptual noise synthesis	PNS	14496-3	4.6.13				
	noise filling	23003-3	7.2	X	X	X	X
MS	basic mid/side coding	14496-3	4.6.8.1	X	X	X	X
	MDCT based complex prediction	23003-3	7.7.2	X	X	X	X
quantization	non-uniform	14496-3	4.6.1	X	X	X	X
	uniform	23003-3	7.1	X	X	X	X
entropy coding	Huffman	14496-3	4.6.3				
	context adaptive arithmetic coding	23003-3	7.4	X	X	X	X
SBR	base	14496-3	4.6.18	X	X	X	
	enhanced	23003-3	7.5	X	X	X	
parametric stereo extension	Parametric Stereo	14496-3	8.6.4 / 8.A				
	MPEG Surround 2-1-2 (incl. residual coding)	23003-3	6.2.13	X	X	X	
	Quad Channel Element	23008-3	5.5		X	X	
ACELP		23003-3	7.14	X	X	X	X
frequency domain noise shaping	scale factor based	14496-3	4.6.2	X	X	X	X
	LPC based	23003-3		X	X	X	X
Intelligent Gap Filling	IGF for FD	23008-3			X	X	X
Improved LPD coding	IGF for TCX and TBE in ACELP	23008-3 Amd3				X	X
	LPD stereo	23008-3 Amd3				X	X
Predictors for FD and TCX	frequency-domain prediction and time-domain post-filtering	23008-3 Amd3				X	X
Discrete Multi-channel coding	MCT	23008-3 Amd3				X	X
Format Converter	Generic downmix	23008-3	10, Amd3.1		X	X	X (Note 4)
Immersive Rendering	Immersive rendering within format converter	23008-3	11, Amd3.2		X	X	X (Note 4)
Static metadata	Metadata Audio Elements (MAE) and Audio Scene Information (ASI) Decoder and Renderer	23008-3	15		X	X	X
Dynamic object metadata	Object Audio Metadata (OAM) Decoder and Renderer	23008-3	7, 8		X	X	X
	MPEG Surround Extension	23003-1 Amd 3	9			X	
SAOC-3D	Decoder and Renderer	23008-3	9		X	X	
HOA	Decoder and Renderer	23008-3 and Amd3	12		X	X	X (Note 5)

	Near Field Compensation	23008-3			X	X	X (Note1)
	Subband Directional Prediction	23008-3 Amd3				X	
	Parametric Ambiance Replication (PAR)	23008-3 Amd3				X	
	Phase-based decorrelation	23008-3 Amd3				X	
Binaural	FD-binaural, TD-binaural	23008-3	13		X	X	X (Note2)
	HOA2Binaural H2B	23008-3			X	X	X (Note2)
DRC	DRC-1	23003-4			X	X	X (Note3)
	DRC-2 (single band)	23003-4			X	X	X
	DRC-2 (multi band)	23003-4					
	DRC-3 (single band)	23003-4			X	X	X
Sample Rate Converter		23008-3 Amd3	Amd3.3			X	X
Peak Limiter	Unguided clipping prevention	23008-3 23003-4	D			X	X
Loudness	Loudness metadata and handling	23003-4	6		X	X	X
	Loudness compensation	23008-3 Amd3				X	X
MHAS	MPEG-H 3D audio stream	23008-3	14		X	X	X
	Truncation message and CRC packet type, ASI packet type	23008-3 Amd3				X	X
File Format	Carriage of MPEG-H 3D Audio in ISO base media file format	23008-3 Amd2			(Note 6)		
Interfaces and processing	Interfaces and processing for Interaction data and local setup info	23008-3	17,18		X	X	X
Carriage of system data	Carriage of System Data for the interaction with System Engine	23008-3 Amd4				X	X
TCC	Tonal Component Coding	23008-3 Amd3				X	
IC	Internal Channel	23008-3 Amd3				X	
HREP	High Resolution Envelope Processing	23008-3 Amd3				X	

Note 1: Restrictions apply dependent on the levels
Note 2: Implementation of binaural rendering is only mandated if headphone reproduction is supported.
Note 3: Multi-band DRC-1 shall be applied in the STFT domain of the TD format converter.
Note 4: The TD format converter downmix shall be applied for downmixing.
Note 5: In order to achieve target complexity for the LC profile at a given level implementers should study Annex G.
Note 6: File Format encapsulation is independent of the profile that is used for the bitstream. A profile level indicator is part of the file format specification (see XXX).

4.X.2.1 Levels of the Low Complexity Profile

Table P2 — Levels and their corresponding restrictions for the Low Complexity Profile

Level	Max. Sampling rate	Max. no. of core ch. in compressed data stream	Max. no. of decoder processed core ch.	Max. no. of loudspeaker output ch.	Example of max. loudspeaker configuration	Max. no. of decoded objects	Example of a max. Config C+O	Max. HOA order	Example of max. HOA order + O
1	48000	10	5	2	2.0	5	2 ch. + 3 static obj. NOTE 1	2	2nd order + 3 static obj. NOTE 1
2	48000	18	9	8	7.1	9	6 ch. + 3 static obj. NOTE 1	4	4th order + 3 static obj. NOTE 1
3	48000	32	16	12	11.1	16	12 ch. + 4 obj.	6	6th order + 4 obj.
4	48000	56	28	24	22.2	28	24 ch. + 4 obj.	6	6th order + 4 obj.
5	96000	56	28	24	22.2	28	24 ch. + 4 obj.	6	6th order + 4 obj.
NOTE 1 – In this context "static objects" are understood as channel-based signals without accompanying OAM data which are not also associated to a channel bed									

— The use of switch groups determines the subset of core channels out of the core channels in the bitstream that shall be decoded.

— If the mae_AudioSceneInfo() contains switch groups (mae_numSwitchGroups>0), then the elementLengthPresent flag shall be 1

— The number of channels of the signaled referenceLayout shall not exceed the maximum number of loudspeaker output channels as defined in the levels Table P2

Table P3 — Approximated worst case processing power (PCU) of decoder modules and the whole decoder for the different Levels of the Low Complexity Profile given in MOPS

Level	Core LC	Format Converter	Object Renderer	HOA[2]	Objects only Renderer	DRC	Limiter	Binaural[1]	Worst case PCU
1	33	3	0	3	9	6	4	7	58
2	59	10	0	17	16	18	5	19	118
3	106	36	7	36	29	24	6	27	206
4	186	113	7	93	50	30	9	46	392
5	373	226	14	186	50	34	19	92	758

[1] NOTE: The complexity numbers for binaural processing are calculated on the basis of BRIR filters of 1 second length measured in a BS.1116 compliant room.

[2] NOTE: The complexity numbers for the HOA spatial decoding and rendering are based on the Low

Complexity Combined HOA Spatial Decoding and Rendering described in Annex G.

4.X.2.2 Restrictions for the Low Complexity Profile and Levels

In the low complexity profile the core decoder, format converter, object renderer, HOA renderer and DRC and peak limiter operate in time domain, MDCT-domain or STFT-domain.

The following restrictions apply for HOA renderer and decoder:

Table P4 — Restrictions for the HOA spatial Decoding and Rendering according to the Level of the Low Complexity Profile

Restriction applies to	Maximum allowed value depending on Mpegh3daProfileLevelIndication				
	Lvl 1	Lvl 2	Lvl 3	Lvl 4	Lvl 5
HOA order (max)	2	4	6	6	6
Number of Predominant Sounds (max)	3	5	7	8	8
Number of directional signals used in prediction (max)	2	3	4	5	5
The Near Field Compensation (NFC) processing may be applied to HOA content of an order which is smaller or equal to:	N/A (NFC not allowed)	1	2	3	3

NFC may be employed in not more than one signal group of type SignalGroupTypeHOA.

The following restrictions apply to MPEG-D DRC (ISO/IEC 23003-4) when employed as part of MPEG-H 3D audio:

— drcFrameSizePresent and timeDeltaMinPresent shall be set to 0.

— gainInterpolationType shall be set to 1.

— dependsOnDrcSetPresent shall be set to 0 for drcInstructionsUniDrc() with downmixId == 0.

— HOA signal groups shall be restricted to one drcChannelGroup and DRC gains shall be applied to the HOA core channels (HOATransportChannels).

— The values of bsSequenceIndex within drcInstructionsUniDrc() shall be unique in simultaneously applied DRC sets except for bsSequenceIndex == 0.

— Multiband DRC shall be restricted to drcInstructionsUniDrc() with downmixId == 0. If the bitstream should contain multiband DRC, the number of multiband DRC core channels shall be restricted as follows:

(numAudioChannels +
numAudioObjects + numAudioObjectsMB +
numHOATransportChannels + numHOATransportChannelsMB)
≤ (numCoreChannelsMax(Lvl) – dependsOnDrcSetPresentFlag – 1)

, where:

— numAudioChannels, numAudioObjects and numHOATransportChannels are the number of C, O and HOA core channels as specified in Table 4.

  - numAudioObjectsMB and numHOATransportChannelsMB are the number of O and HOA core channels out of numAudioObjects and numHOATransportChannels that contain multiband DRC
  - numCoreChannelsMax is the maximum number of core channels ("No of Core ch") depending on the Mpegh3daProfileLevelIndication field as defined in Table P2
  - dependsOnDrcSetPresentFlag is set to one if the bitstream contains any configuration with dependsOnDrcSetPresent==1 (otherwise zero).
- nNodes shall be restricted to a maximum value of 32, where nNodes is the number of encoded gain values in the current DRC frame.

Table P5 — Restrictions applying to DRC processing according to the Levels of the Low Complexity Profile

Restriction applies to	Maximum allowed value depending on Mpegh3daProfileLevelIndication				
	Lvl 1	Lvl 2	Lvl 3	Lvl 4	Lvl 5
nDrcChannelGroupsTotal (Note 1)	5	9	16	28	28
drcCoefficientsUniDrcCount	4	4	4	4	4
bandCount (Note 2)	2	4	4	4	4
sequenceCountTotal (Note 3)	24	28	32	48	63
drcInstructionsUniDrcCount	16	16	32	32	32
Note 1: Maximum allowed number of simultaneously active DRC channel groups in all applied DRC sets. Note 2: Maximum allowed number of DRC bands for multiband DRC. Note 3: Sum of all nDrcBands in drcGainSequence() structures plus number of sequences with gainCodingProfile=3.					

The following tool specific restrictions apply:

- If the independent noise filling (INF) of the intelligent gap filling (IGF) is activated (i.e. if igfUseEnf==1), then the Complex Prediction tool shall be restricted to real-only prediction, i.e. complex_coef shall be 0.
- If Stereo Filling is activated (i.e. if stereo_filling==1), then the Complex Prediction tool shall be restricted to real-only prediction, i.e. complex_coef shall be 0.
- The independent noise filling of the intelligent gap filling shall not be employed in cases where igfBgn corresponds to an audio frequency higher than 8 kHz.
- The LPD mode shall only be employed at 3DA core coder sampling rates (as defined in Table 2 — Syntax of mpegh3daConfig()) ≤ 32000 Hz

EXAMPLE For a 48 kHz input signal, the encoder resamples the signal to a 32 kHz core coder sampling rate and the LPD decoder operates at this lower sampling rate. After the core decoding the signal is resampled to 48 kHz.

- The multi-channel coding tool (MCT) shall not employ more stereo boxes than specified in Table P9

Table P9 — Restrictions applying to MCT processing according to the Levels of the Low Complexity Profile

Restriction applies to	Maximum allowed value depending on Mpegh3daProfileLevelIndication				
	Lvl 1	Lvl 2	Lvl 3	Lvl 4	Lvl 5
Number of stereo boxes in MCT	5	9	16	28	28

The following restrictions apply to coding of audio objects and the associated OAM data:

Table P10 — Restrictions applying to Object Processing According to the Levels of the Low Complexity Profile

Restriction applies to	Maximum allowed value n depending on MpeghProfileLevelIndication				
	Lvl 1	Lvl 2	Lvl 3	Lvl 4	Lvl 5
(number of objects without divergence) + 3·(number of objects with divergence > 0) ≤ n	5	9	16	28	28

— Efficient Object Metadata Decoding is not permitted, i.e. lowDelayMetadataCoding shall be 1.

— Furthermore the OAM frame length shall comply to:

OAMFrameLength = outputFrameLength / n,
with n being a positive integer in the range of $\{1,...,4\}$

— Objects shall not employ divergence and spread at the same time.

  — If an object is defined with a spatial extent (spread α > 0.0° for uniform spread, spread_width α_{width} > 0.0° for non-uniform spread) it shall have a divergence value equal to zero.

  — If an object is defined with a divergence value > 0, it shall not have a spatial extent (spread α shall be equal to 0.0° for uniform spread, spread_width α_{width} shall be equal to 0.0° for non-uniform spread)

The following restrictions apply to binaural rendering:

The value of bsBinauralDataFormatID in BinauralRendering() should be set to 1 (if the FD Binaural renderer is implemented) or to 2 (if the TD Binaural renderer is implemented). The value of bsBinauralDataFormatID can be set to 0 if the Parameterization of Binaural Room Impulse Responses according to 13.2.3 or 13.3.3 is implemented.

The number of BRIR sets is restricted to a maximum number of 3.

In case of H2B filters, the number of BRIR filter pairs to be provided shall correspond to 'Maximum H2B filter order' column in Table P8. In the other cases, the following applies:

The number of BRIR pairs in each BRIR set shall correspond to the number indicated in the relevant level-dependent row of Table P8. The measured BRIR positions shall correspond to all nominal geometric positions corresponding to the list of LoudspeakerGeometry indices in Table P8. The correspondence between LoudspeakerGeometry index and nominal geometric position is defined in ISO/IEC 23001-8. Thereby, it is ensured that one BRIR pair is available for each possible regular input channel configuration that can be used within the indicated level.

An input channel configuration is regular if it is defined by means of an ISO/IEC 23001-8 ChannelConfiguration or a list of ISO/IEC 23001-8 LoudspeakerGeometry (CICPspeakerIdx).

If binaural rendering is activated, the measured BRIR positions shall be passed to the mpegh3daLocalSetupInformation(). Thus, all renderer stages are set to the target layout that is equal to the transmitted channel configuration. As one BRIR is available per regular input channel, the Format Converter can be passed through in case regular input channel positions are used.

Table P8 — The binaural restrictions for the LC profile

Level	Number of BRIR pairs	Maximum H2B filter order	BRIR positions by means of Loudspeaker Position Abbreviation	BRIR positions by means of LoudspeakerGeometry according to ISO/IEC 23001-8
1	3	1	L, R, C	0, 1, 2
2	10	2	L, R, C, LS, Rs, Lc, Rc, Lsr, Rsr, Cs	0, 1, 2, 4, 5, 6, 7, 8, 9, 10,
3	21	3	L, R, C, Ls, Rs, Lc, Rc, Lsr, Rsr, Cs, Lss, Rss, Lv, Rc, Cv, Lvr, Rvr, Cvr, Rs, Lvs, Rvs	0, 1, 2, 4, 5, 6, 7, 8, 9, 10, 13, 14, 17, 18, 19, 20, 21, 22, 25, 30, 31
4	28	5	L, R, C, Ls, Rs, Lc, Rc, Lsr, Rsr, Cs, Lss, Rss, Lv, Rc, Cv, Lvr, Rvr, Cvr, Lvss, Rvss, Ts, Lb, Rb, Cb, Lvs, Rvs, Lbs, Rbs,	0, 1, 2, 4, 5, 6, 7, 8, 9, 10, 13, 14, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29, 30, 31, 37, 38,
5	28	6	L, R, C, Ls, Rs, Lc, Rc, Lsr, Rsr, Cs, Lss, Rss, Lv, Rc, Cv, Lvr, Rvr, Cvr, Lvss, Rvss, Ts, Lb, Rb, Cb, Lvs, Rvs, Lbs, Rbs,	0, 1, 2, 4, 5, 6, 7, 8, 9, 10, 13, 14, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29, 30, 31, 37, 38,

The following additional parameter values restrictions apply:

— The value of kMax in FdBinauralRendererParam() shall be equal to or less than 48 (bands).

— The value of kConv in FdBinauralRendererParam() shall be equal to 32.

— The values of rt60[k] in SfrBrirParam() shall be less than or equal to 1.0 (sec).

— The average of the values of nFilter[k] shall be less than or equal to 64.

— The values of nFilter[k] in VoffBrirParam() should be less than or equal to 256.

The following coding tools, modules, or features shall not be employed

— Time warped filterbank

— 768 sample outputFrameLength, i.e. coreSbrFrameLengthIndex shall not be 0

The following text describes restrictions dependent on the length of the arithmetic coder codeword, arith_data(). For this text the following definitions apply:

F_{sOut} core coder sampling rate as indicated by means of usacSamplingFrequencyIndex or usacSamplingFrequency in mpegh3daConfig()

F_{sMax} maximum allowed sampling rate of a given level in this profile

N_{chMax} maximum number of decoder processed core channels of a given level in this profile according to Table P2.

N_{chLtpf} number of core coder channels in which the long term post filter (LTPF) is applied

N_{chInf} number of core coder channels in which the independent noise filling (INF) is applied