INTERNATIONAL STANDARD

ISO/IEC 23003-3

Second edition
2020-06

Information technology — MPEG audio technologies —

Part 3: Unified speech and audio coding

Technologies de l'information — Technologies audio MPEG — Partie 3: Codage unifié parole et audio

Reference number
ISO/IEC 23003-3:2020(E)

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

Published in Switzerland

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.

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) or the IEC list of patent declarations received (see http://patents.iec.ch).

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

For an explanation of the voluntary nature of standards, 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 www.iso.org/iso/foreword.html.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

This second edition cancels and replaces the first edition (ISO/IEC 23003-3:2012), which has been technically revised. It also incorporates ISO/IEC 23003-3:2012/Cor.1:2012, ISO/IEC 23003-3:2012/Cor.2:2013, ISO/IEC 23003-3:2012/Cor.3:2015, ISO/IEC 23003-3:2012/Cor.4:2015, ISO/IEC 23003-3:2012/Amd.1:2014, ISO/IEC 23003-3:2012/Amd.1:2014/Cor.1:2015, ISO/IEC 23003-3:2012/Amd.2:2015, ISO/IEC 23003-3:2012/Amd.2:2015/Cor.1:2015 and ISO/IEC 23003-3:2012/Amd.3:2016.

A list of all parts in the ISO/IEC 23003 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

As mobile appliances become multi-functional, multiple devices converge into a single device. Typically, a wide variety of multimedia content is required to be played on or streamed to these mobile devices, including audio data that consists of a mix of speech and music.

This document specifies unified speech and audio coding (USAC), which allows for coding of speech, audio or any mixture of speech and audio with a consistent audio quality for all sound material over a wide range of bitrates. It supports single and multi-channel coding at high bitrates and provides perceptually transparent quality. At the same time, it enables very efficient coding at very low bitrates while retaining the full audio bandwidth.

Where previous audio codecs had specific strengths in coding either speech or audio content, USAC is able to encode all content equally well, regardless of the content type.

In order to achieve equally good quality for coding audio and speech, the developers of USAC employed the proven MDCT-based transform coding techniques known from MPEG-4 audio and combined them with specialized speech coder elements like ACELP. Parametric coding tools such as MPEG-4 spectral band replication (SBR) and MPEG-D MPEG surround were enhanced and tightly integrated into the codec. The result delivers highly efficient coding and operates down to the lowest bit rates.

The main focus of this codec are applications in the field of typical broadcast scenarios, multimedia download to mobile devices, user-generated content such as podcasts, digital radio, mobile TV, audio books, etc.

The International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) draw attention to the fact that it is claimed that compliance with this document may involve the use of a patent.

ISO and IEC take no position concerning the evidence, validity and scope of this patent right.

The holder of this patent right has assured ISO and IEC that he/she is willing to negotiate licences under reasonable and non-discriminatory terms and conditions with applicants throughout the world. In this respect, the statement of the holder of this patent right is registered with ISO and IEC. Information may be obtained from the patent database available at www.iso.org/patents.

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

Information technology — MPEG audio technologies —

Part 3: **Unified speech and audio coding**

1 Scope

This document specifies a unified speech and audio codec which is capable of coding signals having an arbitrary mix of speech and audio content. The codec has a performance comparable to, or better than, the best known coding technology that might be tailored specifically to coding of either speech or general audio content. The codec supports single and multi-channel coding at high bitrates and provides perceptually transparent quality. At the same time, it enables very efficient coding at very low bitrates while retaining the full audio bandwidth.

This document incorporates several perceptually-based compression techniques developed in previous MPEG standards: perceptually shaped quantization noise, parametric coding of the upper spectrum region and parametric coding of the stereo sound stage. However, it combines these well-known perceptual techniques with a source coding technique: a model of sound production, specifically that of human speech.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 14496-3:2019, *Information technology — Coding of audio-visual objects — Part 3: Audio*

ISO/IEC 14496-26:2010, *Information technology — Coding of audio-visual objects — Part 26: Audio conformance*

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

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

3 Terms, definitions, symbols and abbreviated terms

3.1 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 14496-3, ISO/IEC 23003-1 and the following apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

3.1.1
algebraic codebook
fixed codebook where an algebraic code is used to populate the excitation vectors (innovation vectors)

Note 1 to entry: The excitation contains a small number of nonzero pulses with predefined interlaced sets of potential positions. The amplitudes and positions of the pulses of the k^{th} excitation codevector can be derived from its index k through a rule requiring no or minimal physical storage, in contrast with stochastic codebooks whereby the path from the index to the associated codevector involves look-up tables.

3.1.2
algebraic vector quantizer
AVQ
process associating, to an input block of 8 coefficients, the nearest neighbour from an 8-dimensional lattice and a set of binary indices to represent the selected lattice point

Note 1 to entry: This definition describes the encoder. At the decoder, AVQ describes the process to obtain, from the received set of binary indices, the 8-dimensional lattice point that was selected at the encoder.

3.1.3
closed-loop pitch
result of the adaptive codebook search, a process of estimating the pitch (lag) value from the weighted input speech and the long-term filter state

Note 1 to entry: In the closed-loop search, the lag is searched using error minimization loop (analysis-by-synthesis). In USAC, closed-loop pitch search is performed for every subframe.

3.1.4
fractional pitch
set of pitch lag values having sub-sample resolution

Note 1 to entry: In the LPD USAC, a sub-sample resolution of $1/4^{th}$ or $1/2^{nd}$ of a sample is used.

3.1.5
zero-input response
ZIR
output of a filter due to past inputs, i.e. due to the present state of the filter, given that an input of zeros is applied

3.1.6
immediate play-out frame
IPF
audio frame that contains an extension payload of type ID_EXT_ELE_AUDIOPREROLL

Note 1 to entry: The extension payload shall contain a Config() element as defined in 7.18, i.e., configLen > 0. It should also contain an adequate number of audio pre-roll frames, i.e., numPreRollFrames > 0, for pre-rolling the audio decoder.

3.1.7
independently decodable frame
IF
audio frame in which the bitstream element usacIndependencyFlag has a value of 1

3.1.8
bitstream
encoded audio data

3.1.9
conformance data
conformance test sequences and conformance tools

3.1.10
conformance tool
tool to check certain conformance criteria

3.1.11
conformance test sequence
conformance test bitstreams and corresponding reference waveforms

3.1.12
conformance test bitstream
USAC bitstream used for testing the conformance of a USAC decoder

3.1.13
conformance test condition
condition which applies to properties of a conformance test bitstream in order to test a certain functionality of the USAC decoder

3.1.14
conformance test case
combination of one or more conformance test conditions for which a set of conformance test sequences is provided

3.1.15
main audio channel
audio channel conveyed by means of a UsacSingleChannelElement or UsacChannelPairElement

3.1.16
reference waveform
decoded counterpart of a bitstream

3.1.17
USAC bitstream
data encoded according to this document

3.2 Symbols and abbreviated terms

For the purposes of this document, the symbols and abbreviated terms given in ISO/IEC 14496-3 and the following apply.

ACELP algebraic code-excited linear predictor

PVC predictive vector coding

uclbf unary code, left bit first

NOTE "left bit first" refers to the order in which the unary codes are received. The value is encoded using a conventional unary code, where any decimal value *d* is represented by *d* ‘1’ bits followed by one ‘0’ stop-bit.

USAC unified speech and audio coding

UsacCPE UsacChannelPairElement

UsacEXT UsacExtElement

UsacLFE UsacLfeElement

UsacSCE UsacSingleChannelElement

v[ ] = {a} This expression indicates that all elements of the array v shall be set to the value a.

4 Technical overview

4.1 Decoder block diagram

The block diagram of the USAC decoder as shown in Figure 1 reflects the general structure of MPEG-D USAC which can be described as follows (from bottom to top): There is a common pre/postprocessing stage consisting of an MPEG Surround functional unit to handle stereo processing (MPS212) and an enhanced SBR (eSBR) unit which handles the parametric representation of the higher audio frequencies in the input signal. Then there are two branches, one consisting of a modified advanced audio coding (AAC) tool path (frequency domain, "FD") and the other consisting of a linear prediction coding (LP or LPC domain, "LPD") based path. The latter can use either a frequency domain representation or a time domain representation of the LPC residual. All transmitted spectra for both FD and LPD path are represented in MDCT domain. The quantized spectral coefficients are coded using a context adaptive arithmetic coder. The time domain representation uses an ACELP excitation coding scheme.

In case of transmitted spectral information, the decoder shall reconstruct the quantized spectra, process the reconstructed spectra through whatever tools are active in the bitstream payload in order to arrive at the actual signal spectra as described by the input bitstream payload, and finally convert the frequency domain spectra to the time domain. Following the initial reconstruction and scaling of the spectrum, there are optional tools that modify one or more of the spectra in order to provide more efficient coding.

In case of transmitted time domain signal representation, the decoder shall reconstruct the quantized time signal, process the reconstructed time signal through whatever tools are active in the bitstream payload in order to arrive at the actual time domain signal as described by the input bitstream payload.

For each of the optional tools that operate on the signal data, the option to "pass through" is retained, and in all cases where the processing is omitted, the spectra or time samples at its input are passed directly through the tool without modification.

In places where the bitstream changes its signal representation from time domain to frequency domain representation or from LP domain to non-LP domain or vice versa, the decoder shall facilitate the transition from one domain to the other by means of an appropriate transition mechanism.

eSBR and MPS212 processing is applied in the same manner to both coding paths after transition handling.

The USAC specification offers in some instances multiple decoding options that serve to provide different quality/complexity trade-offs.

NOTE An informative overview of encoder tools is given by Annex B.

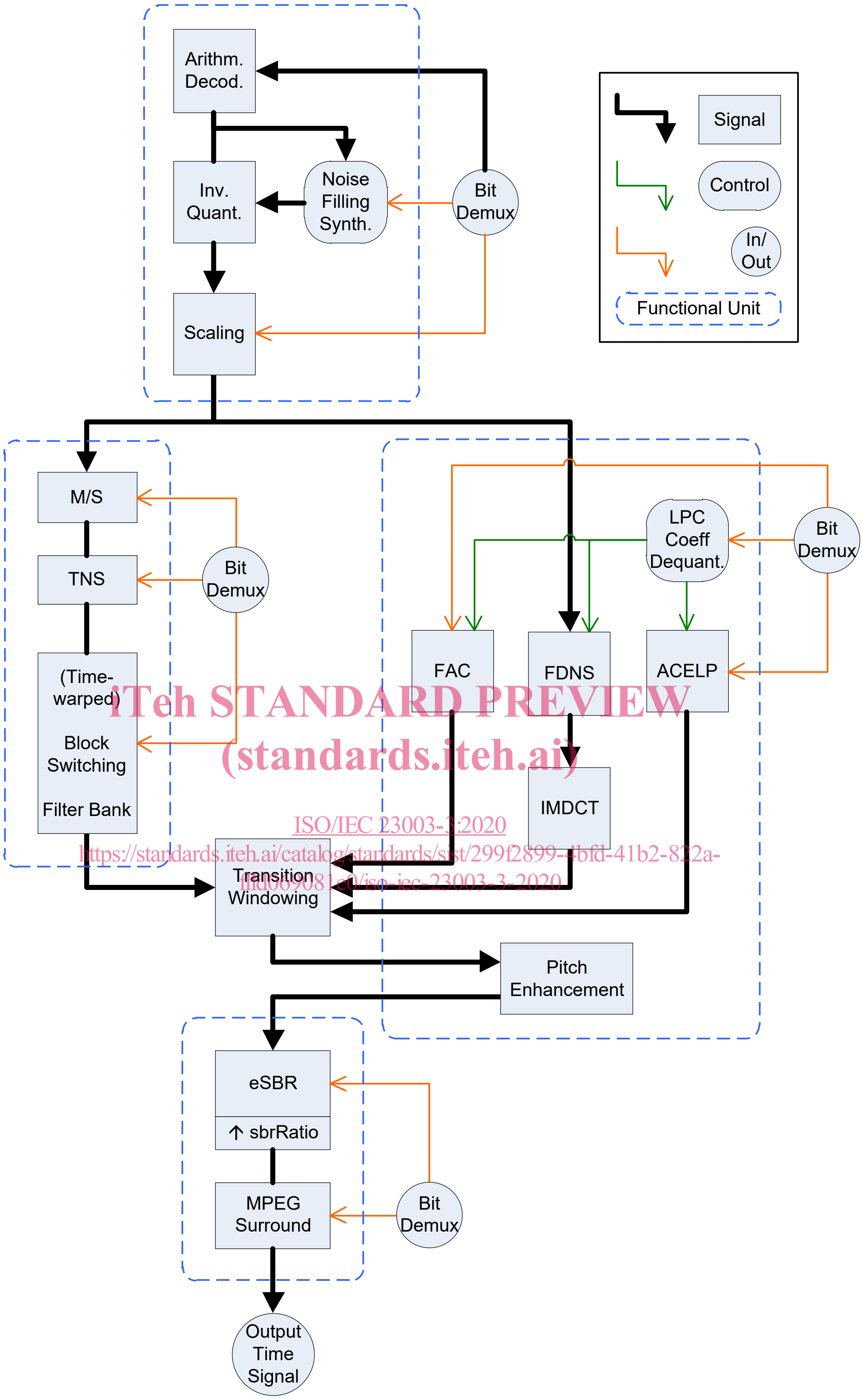

Figure 1 — Simplified block diagram of the typical USAC decoder configuration

4.2 Overview of the decoder tools

The input to the bitstream payload demultiplexer tool is the MPEG-D USAC bitstream payload. The demultiplexer separates the bitstream payload into the parts for each tool, and provides each of the tools with the bitstream payload information related to that tool.

The outputs from the bitstream payload demultiplexer tool are:

- depending on the core coding type in the current frame, either:
  - the quantized and noiselessly coded spectra represented by:
    - scalefactor information;
    - arithmetically coded spectral lines;
  - or: linear prediction (LP) parameters together with an excitation signal represented by either:
    - quantized and arithmetically coded spectral lines (transform coded excitation, TCX) or;
    - ACELP coded time domain excitation;
- the spectral noise filling information (optional);
- the M/S decision information (optional);
- the temporal noise shaping (TNS) information (optional);
- the filterbank control information;
- the time unwarping (TW) control information (optional);
- the enhanced spectral bandwidth replication (eSBR) control information (optional);
- the MPEG Surround 2-1-2 (MPS212) control information (optional).

The scalefactor noiseless decoding tool takes information from the bitstream payload demultiplexer, parses that information, and decodes the Huffman and DPCM coded scalefactors.

The input to the scalefactor noiseless decoding tool is:

- the scalefactor information for the noiselessly coded spectra.

The output of the scalefactor noiseless decoding tool is:

- the decoded integer representation of the scalefactors.

The context adaptive arithmetic decoding tool performs the spectral noiseless decoding step. It takes information from the bitstream payload demultiplexer, parses that information, decodes the context adaptive arithmetically coded data, and reconstructs the quantized spectra.

The input to this noiseless decoding tool is:

- the noiselessly coded spectra.

The output of this noiseless decoding tool is:

- the quantized values of the spectra.

The inverse quantizer tool takes the quantized values for the spectra, and converts the integer values to the non-scaled, reconstructed spectra. This quantizer is a companding quantizer, whose companding factor depends on the chosen core coding mode.

The input to the inverse quantizer tool is:

— the quantized values for the spectra.

The output of the inverse quantizer tool is:

— the un-scaled, inversely quantized spectra.

The noise filling tool is used to fill spectral gaps in the decoded spectra, which occur when spectral value are quantized to zero, e.g., due to a strong restriction on bit demand in the encoder.

The inputs to the noise filling tool are:

— the un-scaled, inversely quantized spectra;

— noise filling parameters;

— the decoded integer representation of the scalefactors.

The outputs to the noise filling tool are:

— the un-scaled, inversely quantized spectral values for spectral lines which were previously quantized to zero;

— modified integer representation of the scalefactors.

The rescaling tool converts the integer representation of the scalefactors to the actual values, and multiplies the un-scaled inversely quantized spectra by the relevant scalefactors.

The inputs to the scalefactors tool are:

— the decoded integer representation of the scalefactors;

— the un-scaled, inversely quantized spectra.

The output from the scalefactors tool is:

— the scaled, inversely quantized spectra.

For an overview over the M/S tool, refer to ISO/IEC 14496-3:2019, 4.1.1.2.

For an overview over the temporal noise shaping (TNS) tool, refer to ISO/IEC 14496-3:2019, 4.1.1.2.

The filterbank/block switching tool applies the inverse of the frequency mapping that was carried out in the encoder. An inverse modified discrete cosine transform (IMDCT) is used for the filterbank tool. The IMDCT can be configured to support 96, 128, 192, 256, 384, 512, 768, or 1024 spectral coefficients.

The inputs to the filterbank tool are:

— the (inversely quantized) spectra;

— the filterbank control information.

The output(s) from the filterbank tool is (are):

— the time domain reconstructed audio signal(s).

