INTERNATIONAL STANDARD

ISO/IEC 23003-3

First edition
2012-04-01

Information technology — MPEG audio technologies —

Part 3: Unified speech and audio coding

Technologies de l'information — Technologies audio MPEG —

Partie 3: Discours unifié et codage audio

Reference number
ISO/IEC 23003-3:2012(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword iv

Introduction v

1 Scope 1

2 Normative references 1

3 Terms, definitions, symbols and abbreviated terms 1
3.1 Terms and definitions 1
3.2 Symbols and abbreviated terms 2

4 Technical Overview 2
4.1 Decoder block diagram 2
4.2 Overview of the decoder tools 5
4.3 Combination of USAC with MPEG Surround and SAOC 8
4.4 Interface between USAC and Systems 9
4.5 USAC Profiles and Levels 9

5 Syntax 12
5.1 General 12
5.2 Decoder configuration (UsacConfig) 12
5.3 USAC bitstream payloads 17

6 Data Structure 50
6.1 USAC configuration 50
6.2 USAC payload 63

7 Tool Descriptions 81
7.1 Quantization 81
7.2 Noise Filling 82
7.3 Scalefactors 84
7.4 Spectral Noiseless coding 84
7.5 enhanced SBR Tool (eSBR) 90
7.6 Inter-subband-sample Temporal Envelope Shaping (inter-TES) 139
7.7 Joint Stereo Coding 142
7.8 TNS 149
7.9 Filterbank and block switching 151
7.10 Time-Warped Filterbank and Blockswitching 159
7.11 MPEG Surround for Mono to Stereo upmixing 167
7.12 AVQ decoding 180
7.13 LPC-filter 186
7.14 ACELP 193
7.15 MDCT based TCX 202
7.16 Forward Aliasing Cancellation (FAC) tool 206
7.17 Post-processing of the synthesis signal 208

Annex A (normative) **Tables** 211

Annex B (informative) **Encoder Tools** 216

Annex C (normative) **Tables for Arithmetic Decoder** 254

Annex D (normative) **Tables for Predictive Vector Coding** 260

Annex E (informative) **Adaptive Time / Frequency Post-Processing** 269

Annex F (informative) **Audio/Systems Interaction** 275

Annex G (informative) **Patent Statements** 277

Bibliography 278

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.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

ISO/IEC 23003-3 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

ISO/IEC 23003 consists of the following parts, under the general title *Information technology — MPEG audio technologies*:

— *Part 1: MPEG Surround*

— *Part 2: Spatial Audio Object Coding (SAOC)*

— *Part 3: Unified speech and audio coding*

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 part of ISO/IEC 23003 Unified Speech and Audio Coding (USAC) is a new audio coding standard that 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, multi-media 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 patents.

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

The holder of this patent right has assured ISO and the IEC that he 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 the IEC. Information may be obtained from the companies listed in Annex G.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights other than those identified in Annex G. ISO and the 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 part of ISO/IEC 23003 specifies a unfied 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 part of ISO/IEC 23003 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 referenced documents are indispensible for the application of this document. For undated references, the latest edition of the referenced document (including any amendments) applies.

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

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

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.

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

NOTE 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 kth 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
AVQ
Algebraic Vector Quantizer
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 The above 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 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 In the LPD USAC, a sub-sample resolution of $1/4^{th}$ or $1/2^{nd}$ of a sample is used.

3.1.5
ZIR
zero input response
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

3.2 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

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.

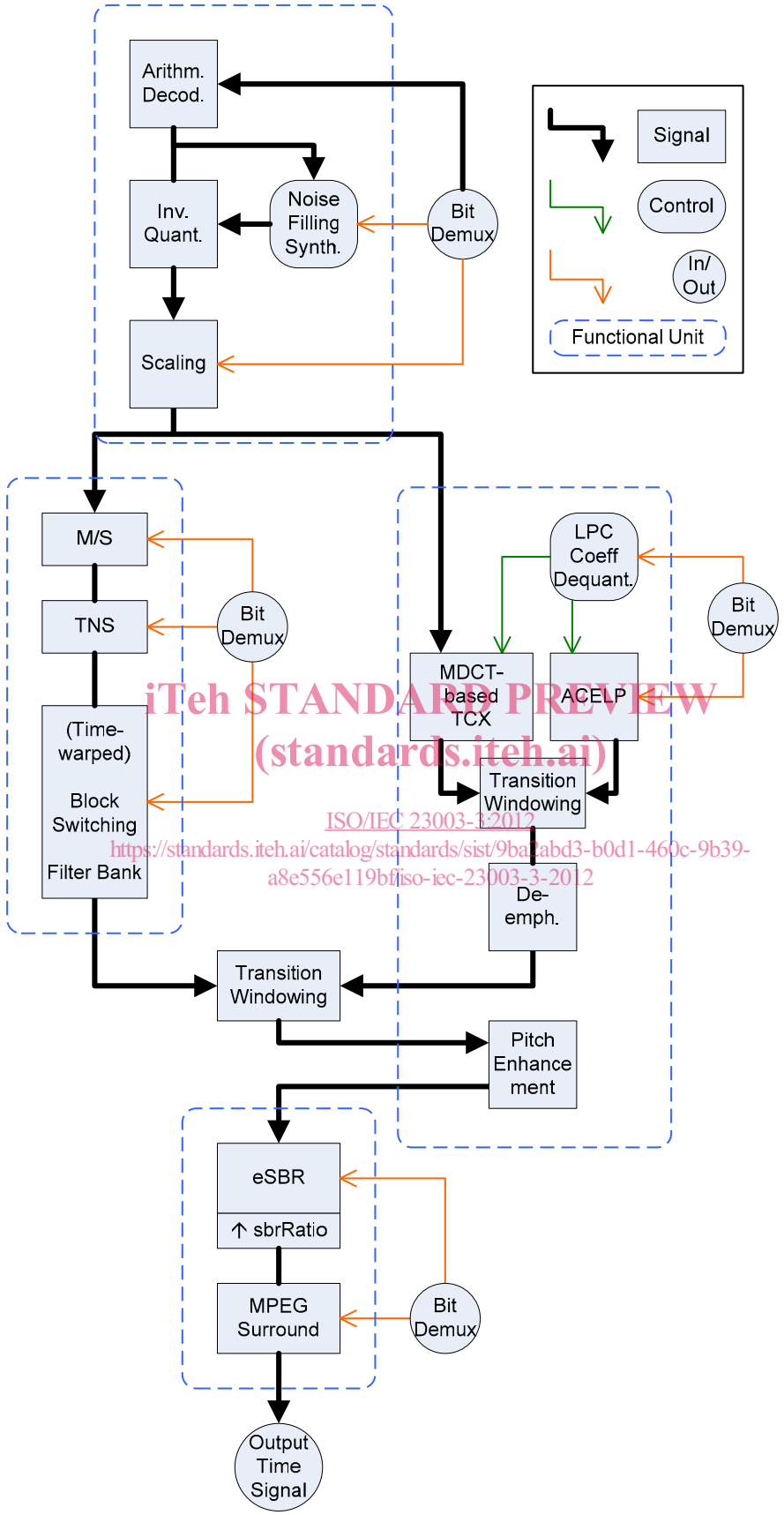

Figure 1 — Simplified block diagram of the typical USAC decoder configuration

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:
  - Tthe 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 spectral noiseless decoding tool takes information from the bitstream payload demultiplexer, parses that information, decodes the 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 use of the noise filling tool is optional.

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, please refer to ISO/IEC 14496-3:2009, 4.1.1.2.

For an overview over the temporal noise shaping (TNS) tool, please refer to ISO/IEC 14496-3:2009, 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 120, 128, 240, 256, 480, 512, 960 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).

The time-warped filterbank / block switching tool replaces the normal filterbank / block switching tool when the time warping mode is enabled. The filterbank is the same (IMDCT) as for the normal filterbank, but in addition the windowed time domain samples are mapped from the warped time domain to the linear time domain by time-varying resampling.

The inputs to the time-warped filterbank tools are:

— The inversely quantized spectra

— The filterbank control information

— The time-warping control information

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

— The linear time domain reconstructed audio signal(s).

The enhanced SBR (eSBR) tool regenerates the highband of the audio signal. It is based on replication of the sequences of harmonics, truncated during encoding. It adjusts the spectral envelope of the generated highband and applies inverse filtering, and adds noise and sinusoidal components in order to recreate the spectral characteristics of the original signal.

The input to the eSBR tool is:

— The quantized envelope data

— Control data

— A time domain signal from the frequency domain core decoder or the ACELP/TCX core decoder

The output of the eSBR tool is either:

— A time domain signal or

— A QMF-domain representation of a signal, e.g. in case MPS212 is used.

The MPEG Surround 2-1-2 (MPS212) tool produces multiple signals from one input signal by applying a sophisticated upmix procedure to the input signal controlled by appropriate spatial parameters. In the USAC context MPS212 is used for coding a stereo signal, by transmitting parametric side information alongside a transmitted downmixed signal.

The input to the MPS212 tool is:

— A downmixed time domain signal or

— A QMF-domain representation of a downmixed signal from the eSBR tool

The output of the MPS212 tool is:

— A two-channel time domain signal

The ACELP tool provides a way to efficiently represent a time domain excitation signal by combining a long term predictor (adaptive codebook codeword) with a pulse-like sequence (innovation codebook codeword). The reconstructed excitation is sent through an LP synthesis filter to form a time domain signal.

The input to the ACELP tool is:

— Adaptive and innovation codebook indices

— Adaptive and innovation codes gain values

— Other control data

— Inversely quantized and interpolated LPC filter coefficients

The output of the ACELP tool is:

— The time domain reconstructed audio signal

The MDCT based TCX decoding tool is used to turn the weighted LP residual representation from an MDCT-domain back to the time domain and outputs a time domain signal in which weighted LP synthesis filtering has been applied. The IMDCT can be configured to support 256, 512, or 1024 spectral coefficients.

The input to the TCX tool is:

— The (inversely quantized) MDCT spectra

— Inversely quantized and interpolated LPC filter coefficients

The output of the TCX tool is:

— The time domain reconstructed audio signal

4.3 Combination of USAC with MPEG Surround and SAOC

The output of the USAC decoder can be further processed by MPEG Surround (MPS) (ISO/IEC 23003-1) or Spatial Audio Object Coding (SAOC) (ISO/IEC 23003-2). If the SBR tool in USAC is active, a USAC decoder can typically be efficiently combined with a subsequent MPS/SAOC decoder by connecting them in the QMF domain in the same way as it is described for HE-AAC in ISO/IEC 23003-1:2007, 4.4. If a connection in the QMF domain is not possible, they need to be connected in the time domain.

If MPS/SAOC side information is embedded into a USAC bitstream by means of the usacExtElement mechanism (with usacExtElementType being ID_EXT_ELE_MPEGS or ID_EXT_ELE_SAOC), the time-alignment between the USAC data and the MPS/SAOC data assumes the most efficient connection between the USAC decoder and the MPS/SAOC decoder. If the SBR tool in USAC is active and if MPS/SAOC employs a 64 band QMF domain representation (see ISO/IEC 23003-1:2007, 6.6.3), the most efficient connection is in the QMF domain. Otherwise, the most efficient connection is in the time domain. This corresponds to the time-alignment for the combination of HE-AAC and MPS as defined in ISO/IEC 23003-1:2007, 4.4, 4.5, and 7.2.1.

The additional delay introduced by adding MPS decoding after USAC decoding is given by ISO/IEC 23003-1:2007, 4.5 and depends on whether HQ MPS or LP MPS is used, and whether MPS is connected to USAC in the QMF domain or in the time domain.

4.4 Interface between USAC and Systems

This subclause clarifies the interface between USAC and MPEG Systems. Every access unit delivered to the audio decoder from the systems interface shall result in a corresponding composition unit delivered from the audio decoder to the systems interface, i.e., the compositor. This shall include start-up and shut-down conditions, i.e., when the access unit is the first or the last in a finite sequence of access units.

For an audio composition unit, ISO/IEC 14496-1:2010, 7.1.3.5 Composition Time Stamp (CTS) specifies that the composition time applies to the *n*-th audio sample within the composition unit. For USAC, the value of *n* is always 1. Note that this applies to the output of the USAC decoder itself. In the case that a USAC decoder is, for example, being combined with an MPS decoder as described in 4.3, the additional delay caused by the MPS decoding process (see 4.3 and ISO/IEC 23003-1:2007, 4.5) needs to be taken into account for the composition units delivered at the output of the MPS decoder.

4.5 USAC Profiles and Levels

4.5.1 Introduction

This subclause defines profiles and their levels for Unified Speech and Audio Coding.

Complexity units are defined to give an approximation of the decoder complexity in terms of processing power and RAM usage required for the decoding process. The approximated processing power is given in "Processor Complexity Units" (PCU), specified in MOPS. The approximated RAM usage is given in "RAM Complexity Units" (RCU), specified in kWords (1000 words).

4.5.2 MPEG-4 HE AACv2 Compatibility

Large parts of the USAC codec are inherited from the codec tools and structure subsumed in the MPEG-4 HE AAC v2 profile. A few of these tools have been adopted into USAC as is. Many more have been adopted into USAC and greatly enhanced in terms of performance, capability and flexibility. Others were substituted with tools which provide a range of advantages over their MPEG-4 counterparts. As a result, USAC retains all *functionalities and performance features* that the AAC family of technologies – AAC, HE AAC, HE AAC v2 – can provide. However, it does not adopt all *tools*.

If a decoder is intended to provide full AAC family functionality, including the legacy MPEG-4 AAC tools, all coding tools listed in Table 1 shall be considered.

