International Telecommunication Union

ITU-R

Radiocommunication Sector of ITU

Recommendation ITU-R BS.1196-8

(10/2019)

Audio coding for digital broadcasting

BS Series

Broadcasting service (sound)

International Telecommunication Union

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 http://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 Recommendations

(Also available online at http://www.itu.int/publ/R-REC/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
SNG	Satellite news gathering
TF	Time signals and frequency standards emissions
V	Vocabulary and related subjects

Note*: This ITU-R Recommendation was approved in English under the procedure detailed in Resolution ITU-R 1.*

Electronic Publication
Geneva, 2019

RECOMMENDATION ITU-R BS.1196-8[*]

Audio coding for digital broadcasting

(Question ITU-R 19-1/6)

(1995-2001-2010-2012-02/2015-10/2015-2017-01/2019-10/2019)

Scope

This Recommendation specifies audio source coding systems applicable for digital sound and television broadcasting. It further specifies a system applicable for the backward compatible multichannel enhancement of digital sound and television broadcasting systems.

Keywords

Audio, audio coding, broadcast, digital, broadcasting, sound, television, codec

The ITU Radiocommunication Assembly,

considering

a) that user requirements for audio coding systems for digital broadcasting are specified in Recommendation ITU-R BS.1548;

b) that multi-channel sound system with and without accompanying picture is the subject of Recommendation ITU-R BS.775 and that a high-quality, multi-channel sound system using efficient bit rate reduction is essential in a digital broadcasting system;

c) that the advanced sound system specified in Recommendation ITU-R BS.2051 consists of three-dimensional channel configurations and uses either static or dynamic metadata to control object-based, scene-based, and channel-based signals;

d) that subjective assessment of audio systems with small impairments, including multi-channel sound systems is the subject of Recommendation ITU-R BS.1116;

e) that subjective assessment of audio systems of intermediate audio quality is subject of Recommendation ITU-R BS.1534 (MUSHRA);

f) that low bit-rate coding for high quality audio has been tested by the ITU Radiocommunication Sector;

g) that commonality in audio source coding methods among different services may provide increased system flexibility and lower receiver costs;

h) that many broadcast services already use or have specified the use of audio codecs from the families of MPEG-1, MPEG-2, MPEG-4, AC-3 and E-AC-3;

i) that Recommendation ITU-R BS.1548 lists codecs that have been shown to meet the broadcaster’s requirements for contribution, distribution and emission;

j) that those broadcasters which have not yet started services should be able to choose the system which is best suited to their application;

k) that broadcasters may need to consider compatibility with legacy broadcasting systems and equipment when selecting a system;

[*] This Recommendation should be brought to the attention of the International Standardization Organization (ISO) and the International Electrotechnical Commission (IEC).

l) that when introducing a multi-channel sound system existing mono and stereo receivers should be considered;

m) that a backward compatible multi-channel extension to an existing audio coding system can provide better bit rate efficiency than simulcast;

n) that an audio coding system should preferably be able to encode both speech and music with equally high fidelity,

recommends

1 that for new applications of digital sound or television broadcasting emission, where compatibility with legacy transmissions and equipment is not required, one of the following low bit-rate audio coding systems should be employed:

– Extended HE AAC as specified in ISO/IEC 23003-3:2012;

– E-AC-3 as specified in ETSI TS 102 366 (2014-08);

– AC-4 as specified in ETSI TS 103 190-1 v1.3.1 (2018-02) and ETSI TS 103 190-2 v1.2.1 (2018-02);

– MPEG-H 3D Audio LC Profile as specified in ISO/IEC 23008-3:2019;

– DTS-UHD as specified in ETSI TS 103 491 v1.1.1 (2017-04).

NOTE 1 – Extended HE AAC is a more flexible superset of MPEG-4 HE AAC v2, HE AAC and AAC LC, and includes MPEG-D Unified Speech and Audio Coding (USAC).

NOTE 2 – E-AC-3 is a more flexible superset of AC-3.

NOTE 3 – The AC-4, MPEG-H 3D Audio LC profile and DTS-UHD specifications include capabilities that are able to support the advanced sound system specified in Recommendation ITU-R BS.2051 and users should refer to Recommendation ITU-R BS.1548 for codec compliance;

2 that for applications of digital sound or television broadcasting emission, where compatibility with legacy transmissions and equipment is required, one of the following low bit-rate coding systems should be employed:

– MPEG-1 Layer II as specified in ISO/IEC 11172-3:1993;

– MPEG-2 Layer II half sample rate as specified in ISO/IEC 13818-3:1998;

– MPEG-2 AAC-LC or MPEG-2 AAC-LC with SBR as specified in ISO/IEC 13818-7:2006;

– MPEG-4 AAC-LC as specified in ISO/IEC 14496-3:2009;

– MPEG-4 HE AAC v2 as specified in ISO/IEC 14496-3:2009;

– AC-3 as specified in ETSI TS 102 366 (2014-08);

NOTE 4 – ISO/IEC 11172-3 may sometimes be referred to as 13818-3 as this specification includes 11172-3 by reference.

NOTE 5 – It is encouraged to support Extended HE AAC as specified in ISO/IEC 23003-3:2012. It includes all of the above mentioned AAC versions, thus guaranteeing compatibility with new future as well as legacy broadcast systems worldwide with the same single decoder implementation;

3 that for backward compatible multi-channel extension of digital television and sound broadcasting systems, the multichannel audio extensions described in ISO/IEC 23003-1:2007 should be used;

NOTE 6 – Since the MPEG Surround technology described in ISO/IEC 23003-1:2007 is independent of the compression technology (core coder) used for transmission of the backward compatible signal, the described multi-channel enhancement tools can be used in combination with any of the coding systems recommended under *recommends* 1 and 2;

4 that for distribution and contribution links, one of the following codings may be used at a bit rate of at least the following bit rate per audio signal (i.e. per mono signal or per component of an independently coded stereo signal) excluding ancillary data:

– MPEG-1 Layer II, as specified in ISO/IEC 11172-3, at a bit rate of at least 180 kbit/s per mono audio signal;

– MPEG-4 AAC, as specified in ISO/IEC 14496-3, at a bit rate of at least 144 kbit/s per mono audio signal in case of up to five cascades;

– AC-4 as specified in ETSI TS 103 190-1 v1.3.1 (2018-02) and ETSI TS 103 190-2 v1.2.1 (2018-02), at a bit rate of at least 128 kbit/s per mono audio signal in case of up to five cascades;

– MPEG-H 3D Audio, as specified in ISO/IEC 23008-3, at a bit rate of at least 144 kbit/s per mono audio signal in case of up to five cascades;

5 that for commentary links, ISO/IEC 11172-3 Layer III coding may be used at a bit rate of at least 60 kbit/s excluding ancillary data for mono signals, and at least 120 kbit/s excluding ancillary data for stereo signals, using joint stereo coding;

6 that for high quality applications the sampling frequency should be 48 kHz;

7 that the input signal to the low bit rate audio encoder should be emphasis-free and no emphasis should be applied by the encoder,

further recommends

1 that the latest version of Recommendation ITU-R BS.1548 should be referred to for information about coding system configurations that have been demonstrated to meet quality and other user requirements for contribution, distribution, and emission;

2 that further studies of the requirements for the advanced sound system specified in the latest version of Recommendation ITU-R BS.2051 are needed and that this Recommendation should be updated when these studies are completed.

NOTE – Information about the codecs included in this Recommendation may be found in Annexes 1 to 8.

Annex 1
(informative)

MPEG-1 and MPEG-2, layer II and III audio

1 Encoding

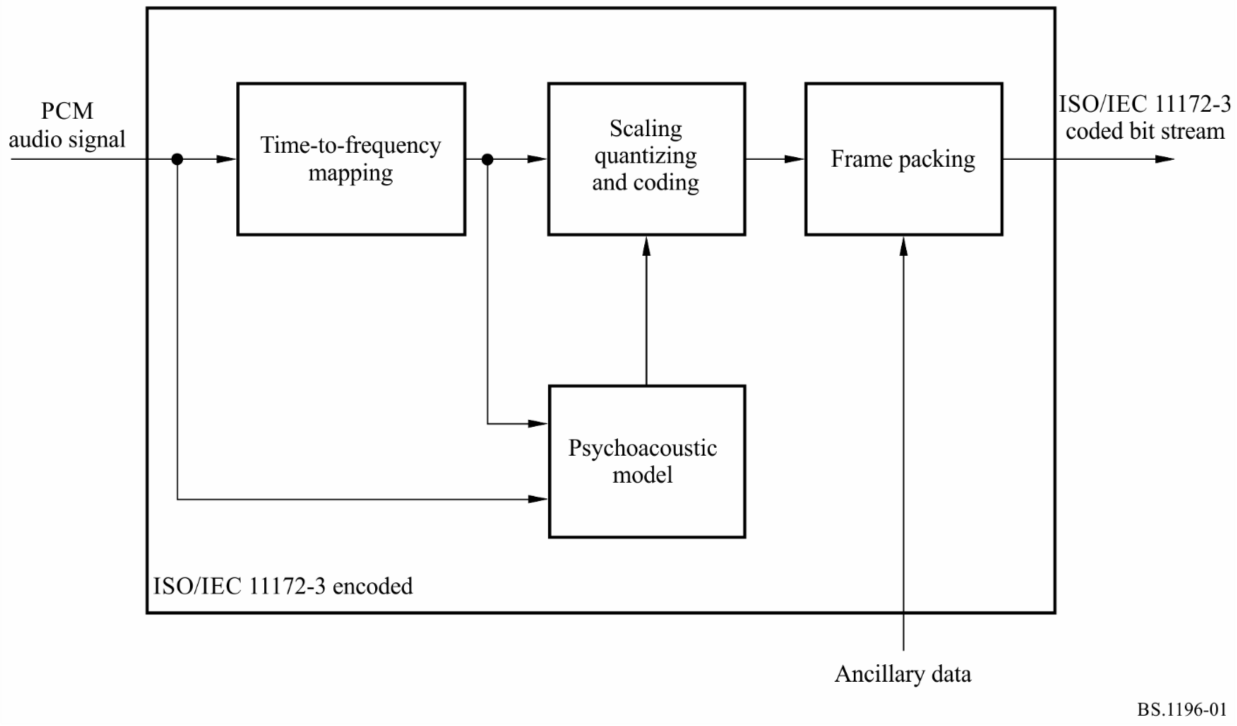
The encoder processes the digital audio signal and produces the compressed bit stream. The encoder algorithm is not standardized and may use various means for encoding, such as estimation of the auditory masking threshold, quantization, and scaling (following Note 1). However, the encoder output must be such that a decoder conforming to this Recommendation will produce an audio signal suitable for the intended application.

NOTE 1 – An encoder complying with the description given in Annexes C and D to ISO/IEC 11172-3, 1993 will give a satisfactory minimum standard of performance.

The following description is of a typical encoder, as shown in Fig. 1. Input audio samples are fed into the encoder. The time-to-frequency mapping creates a filtered and sub-sampled representation of the input audio stream. The mapped samples may be either sub-band samples (as in Layer I or II, see below) or transformed sub-band samples (as in Layer III). A psycho-acoustic model, using a fast Fourier transform, operating in parallel with the time-to-frequency mapping of the audio signal creates a set of data to control the quantizing and coding. These data are different depending on the actual coder implementation. One possibility is to use an estimation of the masking threshold to control the quantizer. The scaling, quantizing and coding block creates a set of coded symbols from the mapped input samples. Again, the transfer function of this block can depend on the implementation of the encoding system. The block "frame packing" assembles the actual bit stream for the chosen layer from the output data of the other blocks (e.g. bit allocation data, scale factors, coded sub-band samples) and adds other information in the ancillary data field (e.g. error protection), if necessary.

FIGURE 1

Block diagram of a typical encoder



2 Layers

Depending on the application, different layers of the coding system with increasing complexity and performance can be used.

Layer I: This layer contains the basic mapping of the digital audio input into 32 sub-bands, fixed segmentation to format the data into blocks, a psycho-acoustic model to determine the adaptive bit allocation, and quantization using block companding and formatting. One Layer I frame represents 384 samples per channel.

Layer II: This layer provides additional coding of bit allocation, scale factors, and samples. One Layer II frame represents 3 × 384 = 1 152 samples per channel.

Layer III: This layer introduces increased frequency resolution based on a hybrid filter bank (a 32 sub-band filter bank with variable length modified discrete cosine transform). It adds a non-uniform quantizer, adaptive segmentation, and entropy coding of the quantized values. One Layer III frame represents 1 152 samples per channel.

There are four different modes possible for any of the layers:

– single channel;

– dual channel (two independent audio signals coded within one bit stream, e.g. bilingual application);

– stereo (left and right signals of a stereo pair coded within one bit stream);

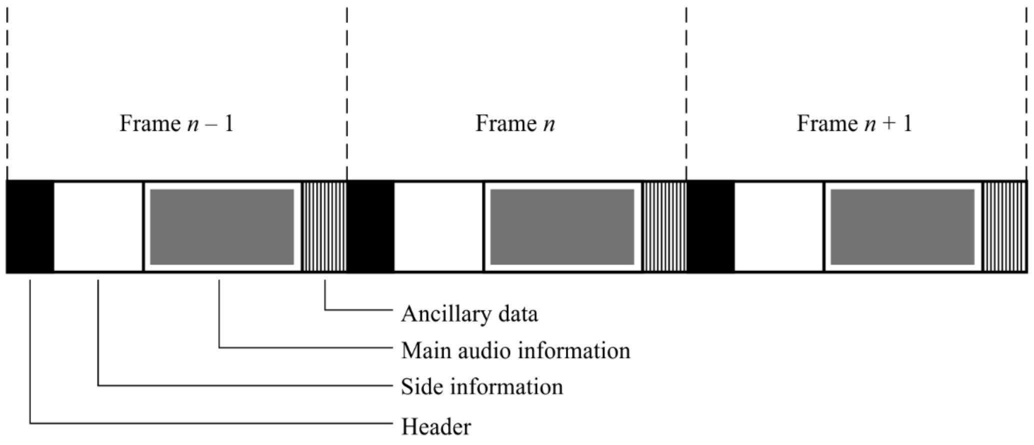
– joint stereo (left and right signals of a stereo pair coded within one bit stream with the stereo irrelevancy and redundancy exploited). The joint stereo mode can be used to increase the audio quality at low bit rates and/or to reduce the bit rate for stereophonic signals.

3 Coded bit stream format

An overview of the ISO/IEC 11172-3 bit stream is given in Fig. 2 for Layer II and Fig. 3 for Layer III. A coded bit stream consists of consecutive frames. Depending on the layer, a frame includes the following fields:

FIGURE 2

ISO/IEC 11172-3 Layer II bit stream format



Layer II:

Header: part of the bit stream containing synchronization and status information

Side information: part of the bit stream containing bit allocation and scale factor information

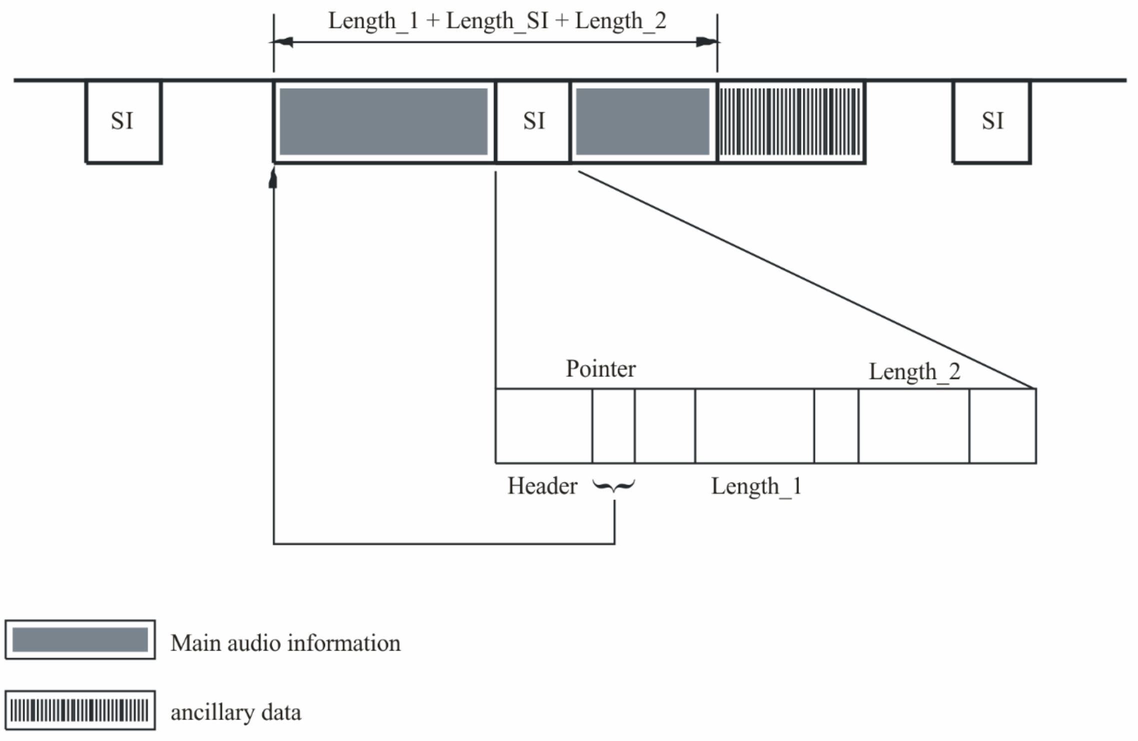
Main audio information: part of the bit stream containing encoded sub-band samples

Ancillary data: part of the bit stream containing user definable data

BS.1196-02

FIGURE 3

ISO/IEC 11172-3 Layer III bit stream format



Layer III:

Side information (SI): part of the bit stream containing header, pointer, length_1 and length_2, scale factor information, etc.;

Header: part of the bit stream containing synchronization and status information;

Pointer: pointing to beginning of main audio information;

Length_1: length of first part of main audio information;

Length_2: length of second part of main audio information;

Main audio information: part of the bit stream containing encoded audio;

Ancillary data: part of the bit stream containing user definable data.

BS.1196-03

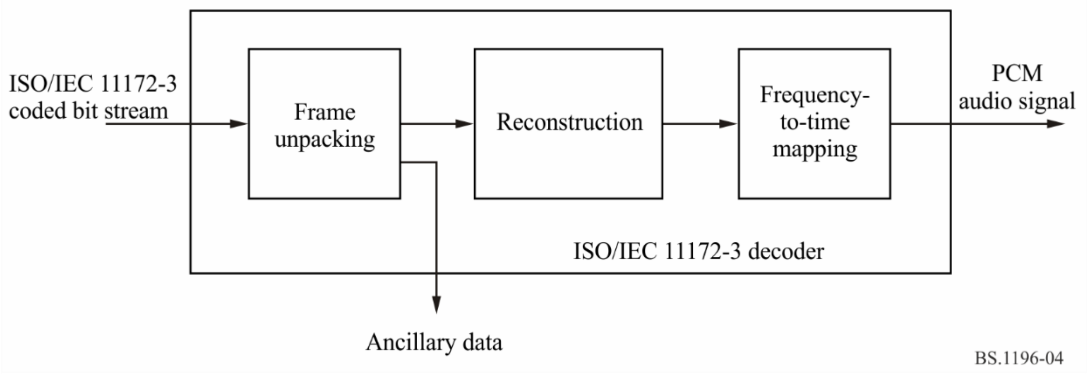
4 Decoding

The decoder accepts coded audio bit streams in the syntax defined in ISO/IEC 11172-3, decodes the data elements, and uses the information to produce digital audio output.

The coded audio bit stream is fed into the decoder. The bit stream unpacking and decoding process optionally performs error detection if error-check is applied in the encoder. The bit stream is unpacked to recover the various pieces of information, such as audio frame header, bit allocation, scale factors, mapped samples, and, optionally, ancillary data. The reconstruction process reconstructs the quantized version of the set of mapped samples. The frequency-to-time mapping transforms these mapped samples back into linear PCM audio samples.

FIGURE 4

Block diagram of the decoder



Annex 2
(informative)

MPEG-2 and MPEG-4 AAC audio

1 Introduction

ISO/IEC 13818-7 describes the MPEG-2 audio non-backwards compatible standards called MPEG-2 Advanced Audio Coding (AAC), a higher quality multichannel standard than achievable while requiring MPEG-1 backwards compatibility.

The AAC system consists of three profiles in order to allow a trade-off between the required memory and processing power, and audio quality:

– *Main profile*

Main profile provides the highest audio quality at any given data rate. All tools except the gain control may be used to provide high audio quality. The required memory and processing power are higher than the LC profile. A main profile decoder can decode an LC-profile encoded bit stream.

– *Low complexity (LC) profile*

The required processing power and memory of the LC profile are smaller than the main profile, while the quality performance keeps high. The LC profile is without predictor and the gain control tool, but with temporal noise shaping (TNS) order limited.

– *Scalable sampling rate (SSR) profile*

The SSR profile can provide a frequency scalable signal with gain control tool. It can choose frequency bands to decode, so the decoder requires less hardware. To decode the only lowest frequency band at the 48 kHz sampling frequency, for instance, the decoder can reproduce 6 kHz bandwidth audio signal with minimum decoding complexity.

AAC system supports 12 types of sampling frequencies ranging from 8 to 96 kHz, as shown in Table 1, and up to 48 audio channels. Table 2 shows default channel configurations, which include mono, two-channel, five-channel (three front/two rear channels) and five-channel plus low-frequency effects (LFE) channel (bandwidth < 200 Hz), etc. In addition to the default configurations, it is possible to specify the number of loudspeakers at each position (front, side, and back), allowing flexible multichannel loudspeaker arrangement. Down-mix capability is also supported. The user can designate a coefficient to down-mix multichannel audio signals into two-channel. Sound quality can therefore be controlled using a playback device with only two channels.

TABLE 1

Supported sampling frequencies

Sampling frequency (Hz)
96 000
88 200
64 000
48 000
44 100
32 000
24 000
22 050
16 000
12 000
11 025
8 000

TABLE 2

Default channel configurations [(1)]

Value [(2)]	**Number of speakers**	**Audio syntactic elements, listed in order received**	**Default element to speaker mapping [(3)]**	**Channel name specified in Recs ITU-R BS.775 or BS.2051 [(4)]**
1	1	single_channel_element	M+000	Mono
2	2	channel_pair_element	M+030, M-030	Left, right
3	3	single_channel_element()	M+000	Centre
		channel_pair_element()	M+030, M-030	Left, right
4	4	single_channel_element()	M+000	Centre
		channel_pair_element()	M+030, M-030	Left, right
		single_channel_element()	M+180	Mono surround
5	5	single_channel_element()	M+000	Centre
		channel_pair_element()	M+030, M-030	Left, right
		channel_pair_element()	M+110, M-110	Left surround, right surround
6	5 + 1	single_channel_element()	M+000	Centre
		channel_pair_element()	M+030, M-030	Left, right
		channel_pair_element()	M+110, M-110	Left surround, right surround
		lfe_element()	LFE1	Low frequency effects

TABLE 2 (*cont*)

Value [(2)]	Number of speakers	Audio syntactic elements, listed in order received	Default element to speaker mapping [(3)]	Channel name specified in Recs ITU-R BS.775 or BS.2051 [(4)]
7	7 + 1 Front	single_channel_element()	M+000	N/A[(5)]
		channel_pair_element()	M+030, M-030	
		channel_pair_element()	M+045, M-045	
		channel_pair_element()	M+110, M-110	
		lfe_element()	LFE1	
8-10	–	–	reserved	–
11	6 + 1	single_channel_element()	M+000	N/A
		channel_pair_element()	M+030, M-030	
		channel_pair_element()	M+110, M-110	
		single_channel_element()	M+180	
		lfe_element()	LFE1	
12	7 + 1 Back	single_channel_element()	M+000	Centre
		channel_pair_element()	M+030, M-030	Left,right
		channel_pair_element()	M+090, M-090	Left side surround, right side surround
		channel_pair_element()	M+135, M-135	Left rear surround, right rear surround
		lfe_element()	LFE1	Low frequency effect
13	22 + 2	single_channel_element()	M+000	Front centre
		channel_pair_element()	M+030, M-030	Front left centre, front right centre
		channel_pair_element()	M+060, M-060	Front left, front right
		channel_pair_element()	M+090, M-090	Side left, side right
		channel_pair_element()	M+135, M-135	Back left, back right
		single_channel_element()	M+180	Back centre
		lfe_element()	LFE1	Low frequency effects-1
		lfe_element()	LFE2	Low frequency effects-2
		single_channel_element()	U+000	Top front centre
		channel_pair_element()	U+045, U-045	Top front left, top front right
		channel_pair_element()	U+090, U-090	Top side left, top side right
		single_channel_element()	T+000	Top centre
		channel_pair_element()	U+135, U-135	Top back left, top back right
		single_channel_element()	U+180	Top back centre
		single_channel_element()	B+000	Bottom front centre
		channel_pair_element()	B+045, U-045	Bottom front left, bottom front right

TABLE 2 (*end*)

Value [(2)]	Number of speakers	Audio syntactic elements, listed in order received	Default element to speaker mapping [(3)]	Channel name specified in Recs ITU-R BS.775 or BS.2051 [(4)]
14	7 + 1 Top	single_channel_element()	M+000	Centre
		channel_pair_element()	M+030, M-030	Left, right
		channel_pair_element()	M+110, M-110	Left surround, right surround
		lfe_element()	LFE1	Low frequency effects
		channel_pair_element()	U+030, U-030	Left top front, right top front
15	–	–	reserved	–

[(1)] The list is quoted from Table 1.19 of ISO/IEC 14496-3:2009/Amd.4:2013.

[(2)] The audio output channel configuration is indicated by a four-bit field that carries a channel configuration value as defined in ISO/IEC 23001-8:2013, "Coding Independent Code Points". MPEG-2 is applicable to channel configuration values of up to 7. MPEG-4 AAC is applicable to channel configuration values of up to 15.

[(3)] Identification of the speakers by labels according to Recommendation ITU-R BS.2051.

[(4)] Please note that the channel labels and names depend on the actual channel configuration.

[(5)] N/A: not applicable; channel configuration not available in Recommendation ITU-R BS.2051 and Recommendation ITU-R BS.775.

2 Encoding

The basic structure of the MPEG-2 AAC encoder is shown in Fig. 5. The AAC system consists of the following coding tools:

– Gain control: A gain control splits the input signal into four equally spaced frequency bands. The gain control is used for SSR profile.

– Filter bank: A filter bank modified discrete cosine transform (MDCT) decomposes the input signal into sub-sampled spectral components with frequency resolution of 23 Hz and time resolution of 21.3 ms (128 spectral components) or with frequency resolution of 187 Hz and time resolution of 2.6 ms (1 024 spectral components) at 48 kHz sampling. The window shape is selected between two alternative window shapes.

– Temporal noise shaping (TNS): After the analysis filter bank, TNS operation is performed. The TNS technique permits the encoder to have control over the temporal fine structure of the quantization noise.

– Mid/side (M/S) stereo coding and intensity stereo coding: For multichannel audio signals, intensity stereo coding and M/S stereo coding may be applied. In intensity stereo coding only the energy envelope is transmitted to reduce the transmitted directional information. In M/S stereo coding, the normalized sum (M as in middle) and difference signals (S as in side) may be transmitted instead of transmitting the original left and right signals.

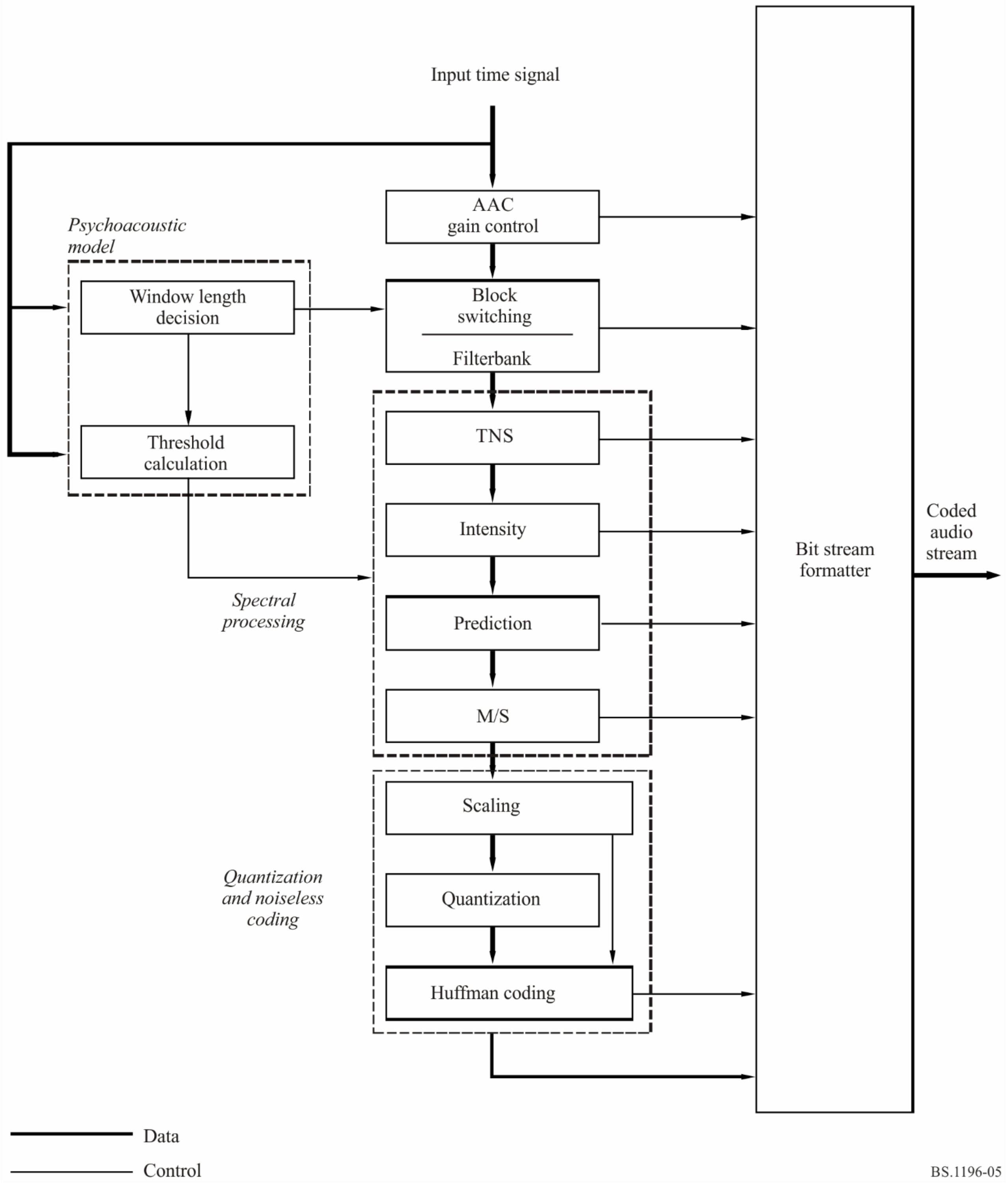
– Prediction: To reduce the redundancy for stationary signals, the time-domain prediction between sub-sampled spectral components of subsequent frames is performed.

– Quantization and noiseless coding: In the quantization tool, a non-uniform quantizer is used with a step size of 1.5 dB. Huffman coding is applied for quantized spectrum, the different scale factors, and directional information.

- Bit-stream formatter: Finally a bit-stream formatter is used to multiplex the bit stream, which consists of the quantized and coded spectral coefficients and some additional information from each tool.
- Psychoacoustic model: The current masking threshold is computed using a psychoacoustic model from the input signal. A psychoacoustic model similar to ISO/IEC 11172-3 psychoacoustic model 2 is employed. A signal-to-mask ratio, which is derived from the masking threshold and input signal level, is used during the quantization process in order to minimize the audible quantization noise and additionally for the selection of adequate coding tool.

FIGURE 5

MPEG-2 AAC encoder block diagram

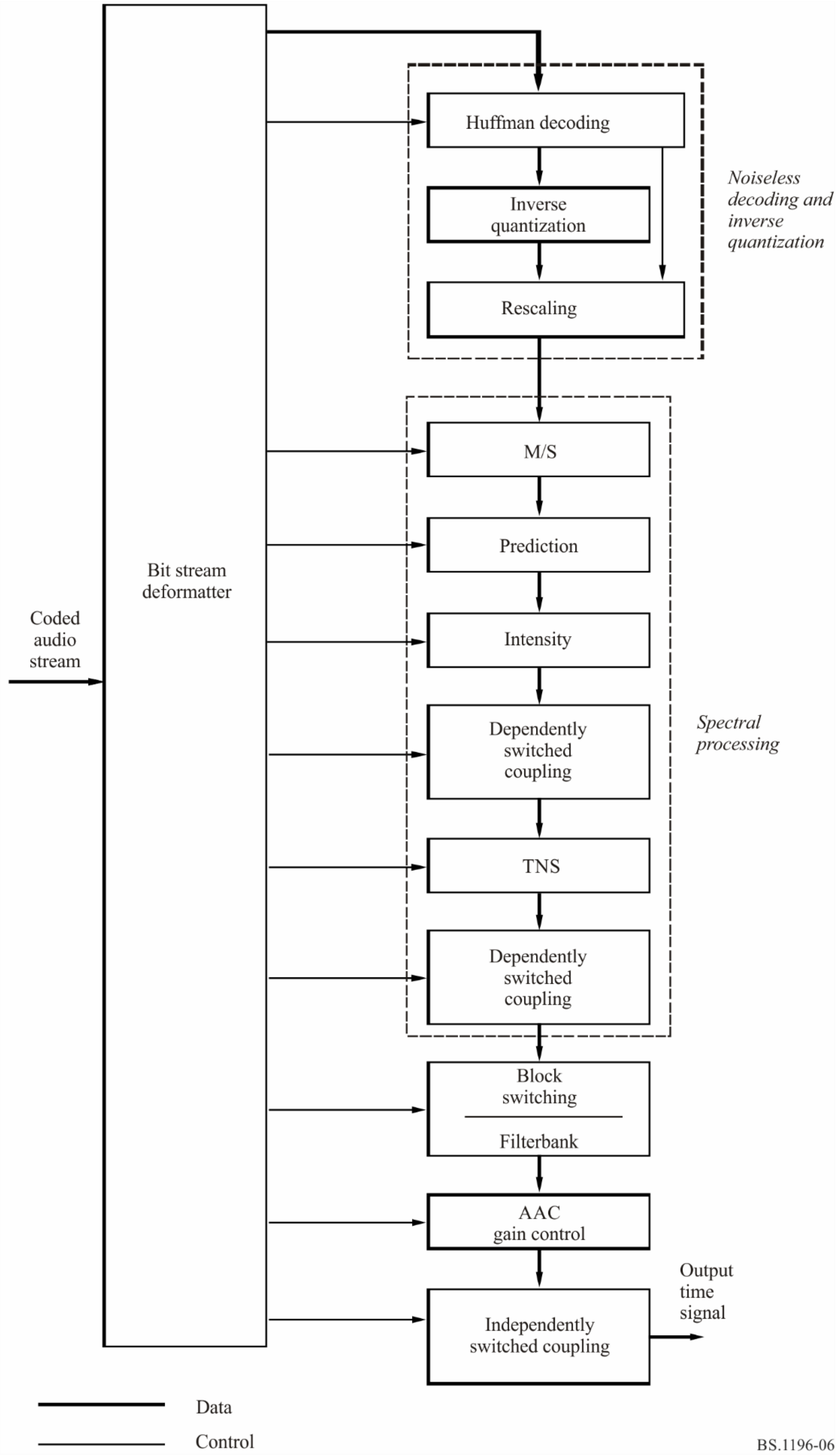


3 Decoding

The basic structure of the MPEG-2 AAC decoder is shown in Fig. 6. The decoding process is basically the inverse of the encoding process.

FIGURE 6

MPEG-2 AAC decoder block diagram

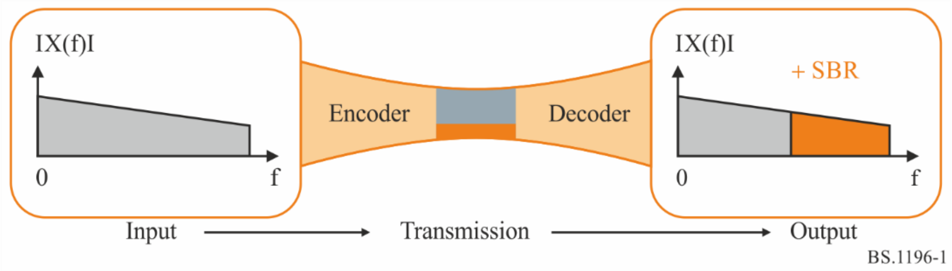


The functions of the decoder are to find the description of the quantized audio spectra in the bit stream, decode the quantized values and other reconstruction information, reconstruct the quantized spectra, process the reconstructed spectra through whatever tools are active in the bit stream in order to arrive at the actual signal spectra as described by the input bit stream, and finally convert the frequency domain spectra to the time domain, with or without an optional gain control tool. Following the initial reconstruction and scaling of the spectrum reconstruction, there are many optional tools that modify one or more of the spectra in order to provide more efficient coding. For each of the optional tools that operate in the spectral domain, the option to "pass through" is retained, and in all cases where a spectral operation is omitted, the spectra at its input are passed directly through the tool without modification.

4 High efficiency AAC and spectral band replication

High Efficiency AAC (HE AAC) introduces spectral band replication (SBR). SBR is a method for highly efficient coding of high frequencies in audio compression algorithms. It offers improved performance of low bit rate audio and speech codecs by either increasing the audio bandwidth at a given bit rate or by improving coding efficiency at a given quality level.

Only the lower part of the spectrum is encoded and transmitted. This is the part of the spectrum to which the human ear is most sensitive. Instead of transmitting the higher part of the spectrum, SBR is used as a post-decoding process to reconstruct the higher frequencies based on an analysis of the transmitted lower frequencies. Accurate reconstruction is ensured by transmitting SBR-related parameters in the encoded bit stream at a very low data rate.

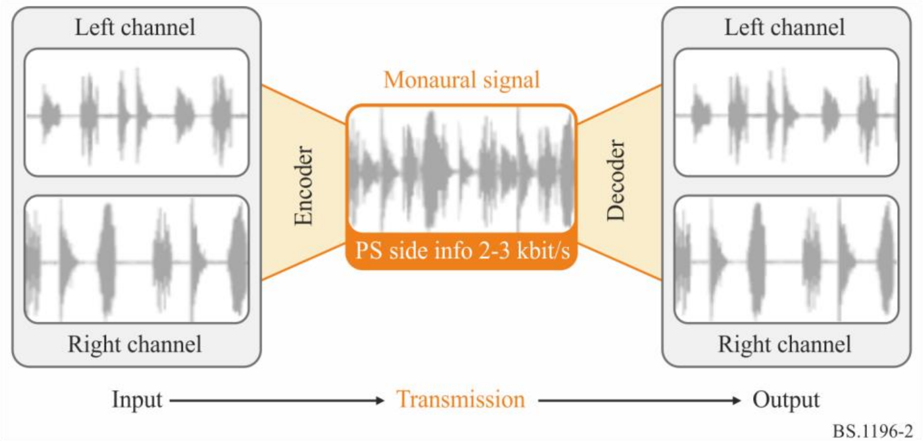


The HE AAC bit stream is an enhancement of the AAC audio bit stream. The additional SBR data is embedded in the AAC fill element, thus guaranteeing compatibility with the AAC standard. The HE AAC technology is a dual-rate system. The backward compatible plain AAC audio bit stream is run at half the sample rate of the SBR enhancement, thus an AAC decoder, which is not capable of decoding the SBR enhancement data, will produce an output time-signal at half the sampling rate than the one produced by an HE AAC decoder.

5 High efficiency AAC version 2 and parametric stereo

HE AAC v2 is an extension to HE AAC and introduces parametric stereo (PS) to enhance the efficiency of audio compression for low bit rate stereo signals.

The encoder analyses the stereo audio signal and constructs a parametric representation of the stereo image. There is now no need to transmit both channels and only a mono-aural representation of the original stereo signal is encoded. This signal is transmitted together with parameters required for the reconstruction of the stereo image.



As a result, the perceived audio quality of a low bit rate audio bit stream (for example, 24 kbit/s) incorporating parametric stereo is significantly higher compared to the quality of a similar bit stream without parametric stereo.

The HE AAC v2 bit stream is built on the HE AAC bit stream. The additional parametric stereo data is embedded in the SBR extension element of a mono HE AAC stream, thus guaranteeing compatibility with HE AAC as well as with AAC.

An HE AAC decoder, which is not capable of decoding the parametric stereo enhancement, produces a mono output signal at the full bandwidth. A plain AAC decoder, which is not capable of decoding the SBR enhancement data, produces a mono output time-signal at half the sampling rate.

Annex 3
(informative)

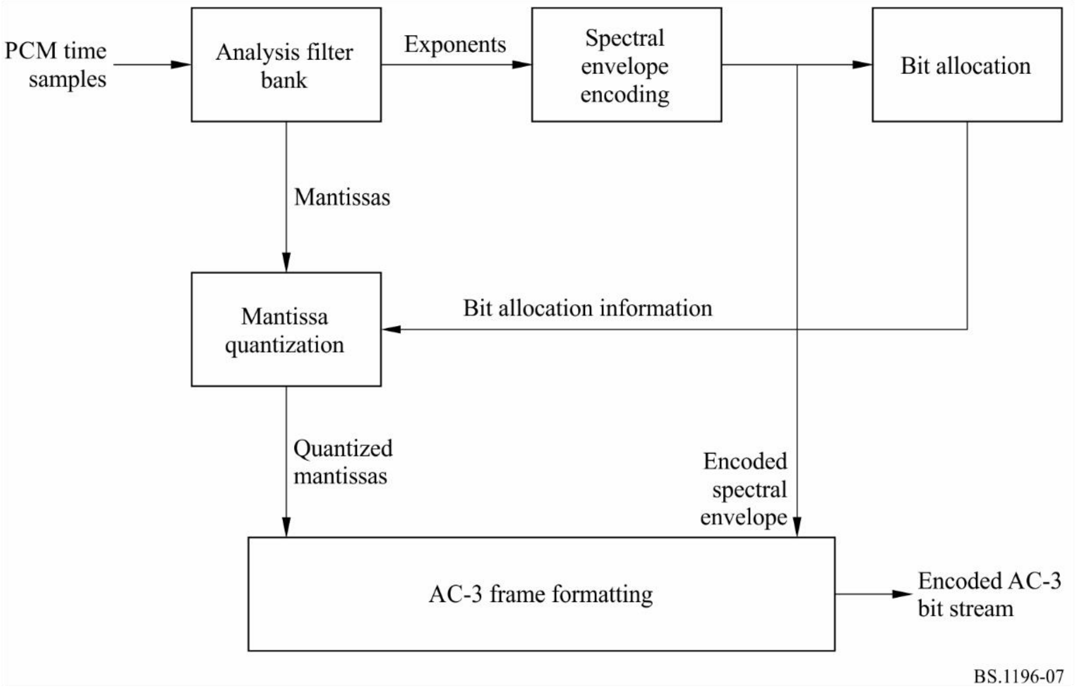
AC-3 and E-AC-3 audio

1 Encoding

The AC-3 digital compression algorithm can encode from 1 to 5.1 channels of source audio from a PCM representation into a serial bit stream at data rates ranging from 32 kbit/s to 640 kbit/s. The AC-3 algorithm achieves high coding gain (the ratio of the input bit rate to the output bit rate) by coarsely quantizing a frequency domain representation of the audio signal. A block diagram of this process is shown in Fig. 7. The first step in the encoding process is to transform the representation of audio from a sequence of PCM time samples into a sequence of blocks of frequency coefficients. This is done in the analysis filter bank. Overlapping blocks of 512 time-samples are multiplied by a time window and transformed into the frequency domain. Due to the overlapping blocks, each PCM input sample is represented in two sequential transformed blocks. The frequency domain representation may then be decimated by a factor of two so that each block contains 256 frequency coefficients. The individual frequency coefficients are represented in binary exponential notation as a binary exponent and a mantissa. The set of exponents is encoded into a coarse representation of the signal spectrum which is referred to as the spectral envelope. This spectral envelope is used by the core bit allocation routine which determines how many bits to use to encode each individual mantissa. The spectral envelope and the coarsely quantized mantissas for 6 audio blocks (1 536 audio samples) are formatted into an AC-3 frame. The AC-3 bit stream is a sequence of AC-3 frames.

FIGURE 7

The AC- encoder



The actual AC-3 encoder is more complex than indicated in Fig. 7. The following functions not shown above are also included:

– a frame header is attached which contains information (bit rate, sample rate, number of encoded channels, etc.) required to synchronize and decode the encoded bit stream;

– error detection codes are inserted in order to allow the decoder to verify that a received frame of data is error free;

– the analysis filter bank spectral resolution may be dynamically altered so as to better match the time/frequency characteristic of each audio block;

– the spectral envelope may be encoded with variable time/frequency resolution;

– a more complex bit allocation may be performed, and parameters of the core bit allocation routine modified so as to produce a more optimum bit allocation;

– the channels may be coupled together at high frequencies in order to achieve higher coding gain for operation at lower bit rates;

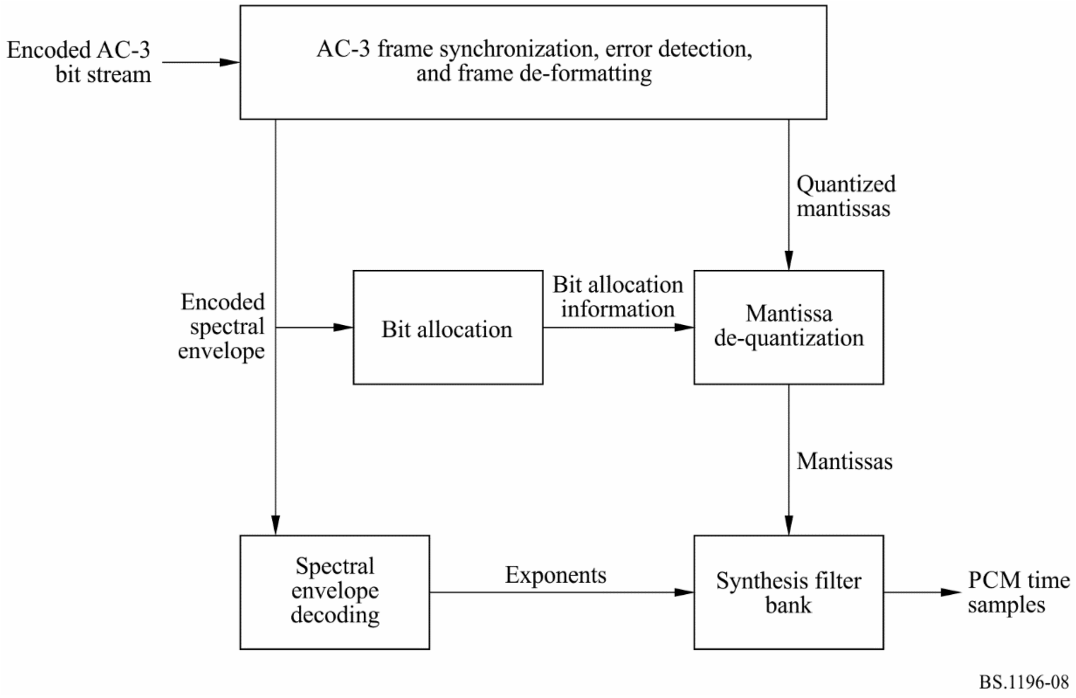
– in the two-channel mode a rematrixing process may be selectively performed in order to provide additional coding gain, and to allow improved results to be obtained in the event that the two-channel signal is decoded with a matrix surround decoder.

2 Decoding

The decoding process is basically the inverse of the encoding process. The decoder, shown in Fig. 8, must synchronize to the encoded bit stream, check for errors, and de-format the various types of data such as the encoded spectral envelope and the quantized mantissas. The bit allocation routine is run and the results used to unpack and de-quantize the mantissas. The spectral envelope is decoded to produce the exponents. The exponents and mantissas are transformed back into the time domain to produce the decoded PCM time samples.

FIGURE 8

The AC-3 decoder


Encoded AC-3 bit stream
AC-3 frame synchronization, error detection, and frame de-formatting
Quantized mantissas
Encoded spectral envelope
Bit allocation
Bit allocation information
Mantissa de-quantization
Mantissas
Spectral envelope decoding
Exponents
Synthesis filter bank
PCM time samples
BS.1196-08

The actual AC-3 decoder is more complex than indicated in Fig. 8. The following functions not shown above are included:

- error concealment or muting may be applied in case a data error is detected;
- channels which have had their high-frequency content coupled together must be de-coupled;
- de-matrixing must be applied (in the 2-channel mode) whenever the channels have been re-matrixed;
- the synthesis filter bank resolution must be dynamically altered in the same manner as the encoder analysis filter bank had been during the encoding process.

3 E-AC-3

Enhanced AC-3 (E-AC-3) adds several additional coding tools and features to the basic AC-3 codec described above. The additional coding tools provide improved coding efficiency allowing operation at lower bit rates, while the additional features provide additional application flexibility.

Additional coding tools:

- Adaptive Hybrid Transform – Additional layer applied in the analysis/synthesis filter bank to provide finer (1/6 of AC-3) spectral resolution.
- Transient pre-noise processing – Additional tool to reduce transient pre-noise.
- Spectral extension – Decoder synthesis of highest frequency components based on side information created by encoder.
- Enhanced coupling – Treats phase as well as amplitude in channel coupling.

Additional features:

- Finer data rate granularity.
- Higher maximum data rate (3 Mbit/s).
- Sub-streams can carry additional audio channels, e.g. 7.1 chs, or commentary tracks.

Annex 4
(informative)

MPEG Surround

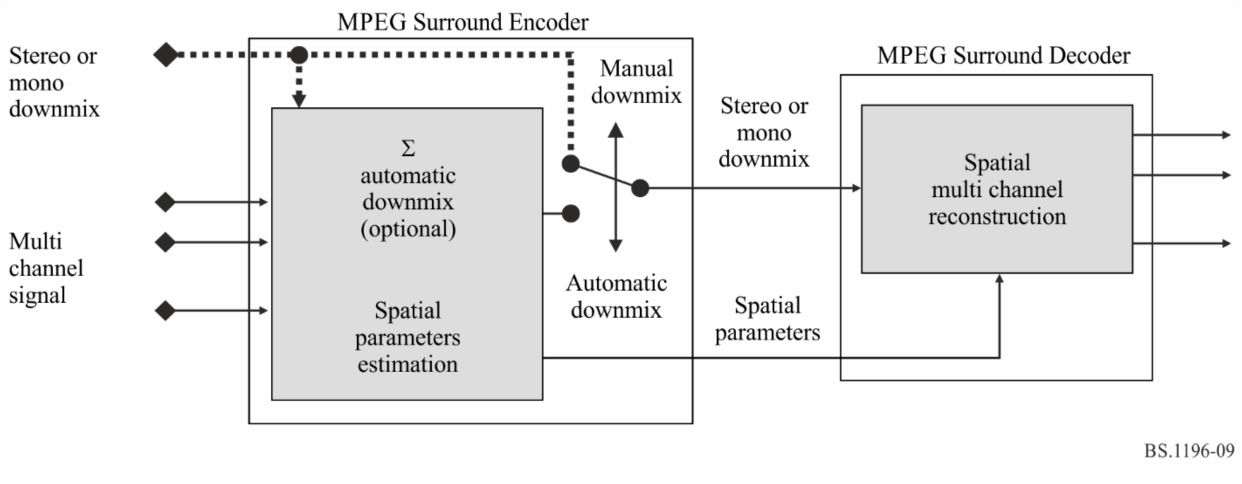
1 Introduction

ISO/IEC 23003-1 or MPEG Surround technology provides an extremely efficient method for coding of multi-channel sound and allows the transmission of surround sound at bit-rates that have been commonly used for coding of mono or stereo sound. It is capable of representing an N channel multi-channel audio signal based on an M<N channel downmix and additional control data. In the preferred operating modes, an MPEG Surround encoder creates either a mono or stereo downmix from the multi-channel audio input signal. This downmix is encoded using a standard core audio codec, e.g. one of the coding systems recommended under *recommends* 1 and 2. In addition to the downmix, MPEG Surround generates a spatial image parameter description of the multi-channel audio that is added as an ancillary data stream to the core audio codec in a backwards compatible fashion. Legacy mono or stereo decoders will ignore the ancillary data and playback the stereo or mono downmix audio signal. MPEG Surround capable decoders will first decode the mono or stereo downmix and then use the spatial image parameters extracted from the ancillary data stream to generate a high quality multi-channel audio signal.

Figure 9 illustrates the principle of MPEG Surround.

FIGURE 9

Principle of MPEG Surround, the downmix is coded using a core audio codec



By using MPEG Surround, existing services can easily be upgraded to provide for surround sound in a backward compatible fashion. While a stereo decoder in an existing legacy consumer device ignores the MPEG Surround data and plays back the stereo signal without any quality degradation, an MPEG Surround-enabled decoder will deliver high quality multi-channel audio.

2 Encoding

The aim of the MPEG Surround encoder is to represent a multi-channel input signal as a backward compatible mono or stereo signal, combined with spatial parameters that enable reconstruction of a multi-channel output that resembles the original multi-channel input signals from a perceptual point of view. Other than the automatically generated downmix an externally created downmix (“artistic downmix”) can be used. The downmix shall preserve the spatial characteristics of the input sound.

MPEG Surround builds upon the parametric stereo technology that has been combined with HE- AAC, resulting in the HE AAC v2 standard specification. By combining multiple parametric stereo modules and other newly developed modules, various structures supporting different combinations of number of output and downmix channels have been defined. As an example, for a 5.1 multi-channel input signal, three different configurations are available; one configuration for stereo downmix based systems (525 configuration.), and two different configurations for the mono downmix based systems (a 515_1 and 515_2 configuration that employ a different concatenation of boxes).

MPEG Surround incorporates a number of tools enabling features that allow for broad application of the standard. A key feature of MPEG Surround is the ability to scale the spatial image quality gradually from very low spatial overhead towards transparency. Another key feature is that the decoder input can be made compatible to existing matrixed surround technologies.

These and other features are realized by the following prominent encoding tools:

– Residual coding: In addition to the spatial parameters, also residual signals can be conveyed using a hybrid coding technique. These signals substitute part of the decorrelated signals (that are part of the Parametric stereo boxes). Residual signals are coded by transforming the QMF domain signals to the MDCT domain after which the MDCT coefficients are coded using AAC.

– Matrix compatibility: Optionally, the stereo downmix can be pre-processed to be compatible to legacy matrix surround technologies to ensure backward-compatibility with decoders that can only decode the stereo bit-stream but are equipped with a matrix-surround decoder.

– Arbitrary downmix signals: The MPEG Surround system is capable of handling not just encoder-generated downmixes but also artistic downmixes supplied to the encoder in addition to the multi-channel original signal.

– MPEG Surround over PCM: Typically, the MPEG Surround spatial parameters are carried in the ancillary data portion of the underlying audio compression scheme. For applications where the downmix is transmitted as PCM, MPEG Surround also supports a method that allows the spatial parameters to be carried over uncompressed audio channels. The underlying technology is referred to as buried data.

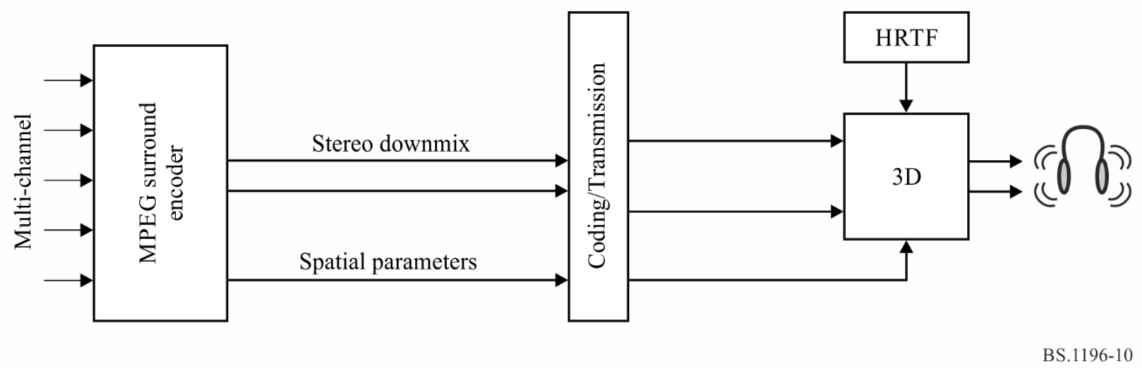
3 Decoding

Next to rendering to a multi-channel output, an MPEG Surround decoder also supports rendering to alternative output configurations:

– Virtual Surround: The MPEG Surround system can exploit the spatial parameters to render the downmix to a stereo virtual surround output for playback over legacy headphones. The standard does not specify the Head Related Transfer Function (HRTF), but merely the interface to these HRTF allowing freedom in implementation depending on the use case. The virtual surround processing can be applied in both the decoder as well as in the encoder, the latter providing the possibility for a virtual surround experience on the downmix, not requiring an MPEG Surround decoder. An MPEG Surround decoder can however undo the virtual surround processing on the downmix and reapply an alternative virtual surround. The basic principle is outlined in Fig. 10.

FIGURE 10

Virtual Surround decoding of MPEG Surround



– Enhanced Matrix Mode: In the case of legacy stereo content, where no spatial side information is present, the MPEG Surround is capable of estimating the spatial side information from the downmix and thus creates the multi-channel output yet offering a quality which is beyond conventional matrix-surround systems.

– Pruning: As a result of the underlying structure, an MPEG Surround decoder can render its output to channel configurations where the number of channels is lower than the number of channels in the multi-channel input of the encoder.

4 Profiles and levels

The MPEG Surround decoder can be implemented as a high quality version and a low power version. Both versions operate on the same data stream, albeit with different output signals.

The MPEG Surround Baseline Profile defines six different hierarchical levels which allow for different numbers of input and output channels, for different ranges of sampling rates, and for a different bandwidth of the residual signal decoding. The level of the decoder must be equal to, or larger than the level of the bit stream in order to ensure proper decoding. In addition, decoders of Level 1, 2 and 3 are capable of decoding all bit streams of Level 2, 3 and 4, though at a possibly slightly reduced quality due to the limitations of the decoder. The quality and format of the output of an MPEG Surround decoder furthermore depends on the specific decoder configuration. However, decoder configuration aspects are completely orthogonal to the different levels of this profile.

5 Interconnection with audio codecs

MPEG Surround operates as a pre- and post-processing extension on top of legacy audio coding schemes. It is therefore equipped with means to accommodate virtually any core audio coder. The framing in MPEG Surround is highly flexible to ensure synchrony with a wide range of coders and means to optimize the connection with coders that already use parametric tools (e.g. spectral band replication) are provided.

Annex 5
(informative)

Extended high efficiency AAC (Extended HE AAC)

1 Introduction

The Extended HE AAC profile is specified within ISO/IEC 23003-3 MPEG-D Unified Speech and Audio Coding (USAC). USAC is an audio coding standard that allows for the 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 where it provides perceptually transparent quality. At the same time, it allows very efficient coding at very low bitrates while retaining the full audio bandwidth.

Where previous audio codecs had specific strengths and weaknesses when coding either speech or audio content, USAC is able to encode all content with equally high fidelity regardless of the content type.

In order to achieve equally good quality for coding audio and speech, USAC employs the proven modified discrete cosine transform (MDCT) based coding techniques known from MPEG-4 Audio (MPEG-4 AAC, HE AAC, HE AAC v2) and combines them with specialized speech coder elements like algebraic code-excited linear prediction (ACELP). Parametric coding tools such as MPEG-4 Spectral Band Replication (SBR) and MPEG-D MPEG Surround are enhanced and tightly integrated into the codec. The result delivers highly efficient coding and operates down to the lowest bit rates.

Currently the USAC standard specifies two profiles:

– *Baseline USAC profile*

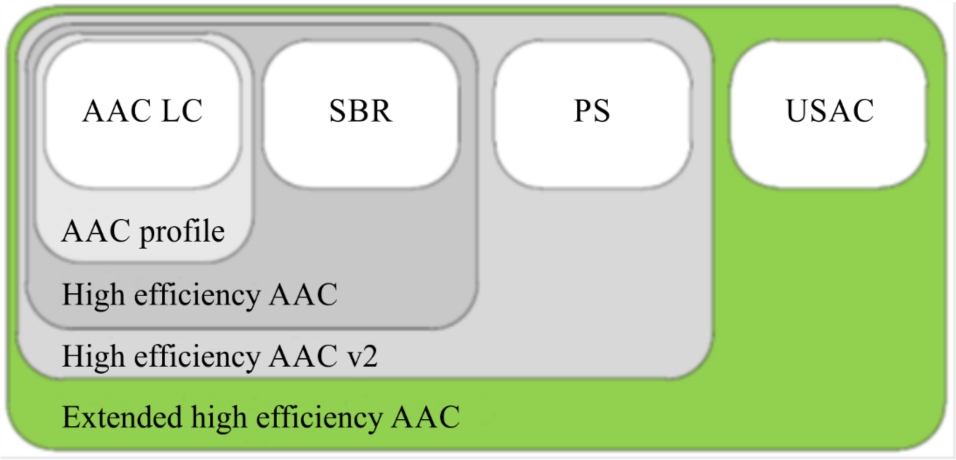
The Baseline USAC profile provides the full functionality of the USAC standard while keeping the overall computational complexity low. Tools with excessive demand for memory or processing power are excluded.

– *Extended HE AAC profile*

Specifically aimed at applications which need to retain compatibility with the existing AAC family of profiles (AAC, HE AAC and HE AAC v2), this profile extends the existing HE AAC v2 profile by adding USAC capabilities. This profile includes level 2 of the *Baseline USAC profile*. Consequently, Extended HE AAC profile decoders can decode all HE AAC v2 bit streams as well as USAC bit streams (up to two channels).

FIGURE 11

Structure of extended high efficiency AAC



BS.1196-11

USAC supports sampling frequencies from 7.35 kHz up to 96 kHz and has shown to deliver good audio quality for a bit rate range starting from 8 kbit/s up to bit rates where perceptual transparency is achieved. This was proven in the verification test (document MPEG2011/N12232) from ISO/IEC JTC 1/SC 29/WG 11 which is attached to Document 6B/286(Rev.2).

The channel configuration can be freely chosen. 13 different default channel configurations can be efficiently signaled for the most common application scenarios. These default configurations include all MPEG-4 channel configurations, such as mono, stereo, 5.0 and 5.1 Surround, or even 7.1 or 22.2 speaker set ups.

2 Encoding

As common use in MPEG standardization, the ISO/IEC 23003-3 standard only specifies the decoding process for MPEG-D USAC files and data streams. It does not normatively specify the encoding process.

A typical, possible encoder structure is shown in Fig. 12.

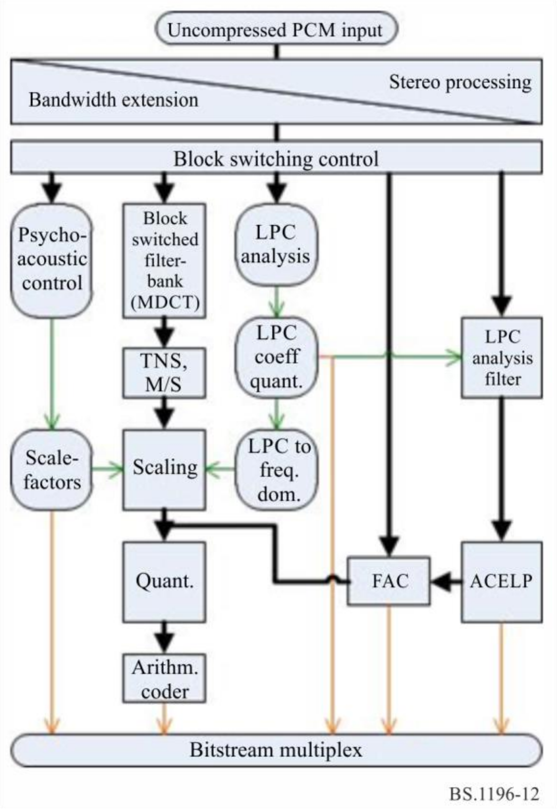
The encoder consists of the following coding tools:

– Stereo processing: At low/intermediate bit rates, USAC employs parametric stereo coding technologies. These are similar in principle to the PS tool as described in ISO/IEC 23003-3 Appendix 2.5 but instead based on MPEG Surround as described in Annex 4 and hence called MPEG Surround 2-1-2 (MPS 2-1-2). The encoder extracts a highly efficient parametric representation of the stereo image from the input audio signal. These parameters are transmitted in the bitstream together with a monaural downmix signal. Optionally the encoder can choose to transmit a residual signal which amends the stereo signal reconstruction process at the decoder. The residual coding mechanism allows a smooth scaling from full parametric to full discrete channel stereo coding. The MPS 2-1-2 tool is intrinsic part of the USAC codec. At higher bitrates, where parametric coding and ACELP are typically not active, stereo coding can be performed exclusively in the MDCT domain by means of a complex-valued stereo prediction. Thus, this method is called complex prediction stereo coding. It can be seen as a generalization of the traditional M/S stereo coding.

- Bandwidth extension: The parametric bandwidth extension is a multiple enhanced version of the MPEG-4 spectral band replication (SBR), which is described in ISO/IEC 23003-3 Appendix 2.4. The encoder estimates spectral envelope and tonality of the higher audio frequency bands and transmits corresponding parameters to the decoder. The encoder can choose from two different transposer types (harmony or copy-up) and from three transposition factors (1:2, 3:8, 1:4). The enhanced SBR tool is an intrinsic part of the USAC codec.
- Filter bank, block switching: An MDCT based filterbank forms the basis for the core coder. Depending on the applied quantization noise shaping mechanism, the transform resolution can be chosen from one out of 1024, 512, 256, or 128 spectral lines. In combination with the 3:8 SBR transposition factor the resolution can be changed to ¾ of the above listed alternatives, providing better temporal granularity even at lower sampling rates.
- Temporal noise shaping (TNS), M/S stereo coding, quantization: These tools have been adopted from AAC and are employed in similar fashion as described in ISO/IEC 23003-3 Appendix 2.2.
- Context adaptive arithmetic coder: noiseless (i.e. entropy) coding of the MDCT spectral coefficients is handled by an arithmetic coder which selects its probability tables based on previously encoded spectral lines.
- Psychoacoustic control, scalefactor scaling: The scalefactor based psychoacoustic model is similar to the one used in AAC, see ISO/IEC 23003-3 Appendix 2.2.
- Scaling based on linear predictive coding (LPC) parameters: This spectral noise shaping tool can be used as an alternative to the above mentioned scalefactor scaling. The weighted version of a frequency representation of an LPC filter coefficient set is applied to the MDCT spectral coefficients prior to quantization and coding.
- ACELP: The algebraic code excited linear prediction (ACELP) coder tool employs the proven adaptive/innovation codebook excitation representation as known from state-of-the-art speech codecs.
- Bitstream multiplex: The final bit stream is composed of the various elements which the encoder tools produce.
- FAC: The forward aliasing correction (FAC) tool provides a mechanism to seamlessly transition from aliasing afflicted MDCT based coding to time-domain based ACELP coding.

FIGURE 12

MPEG-D USAC encoder block diagram

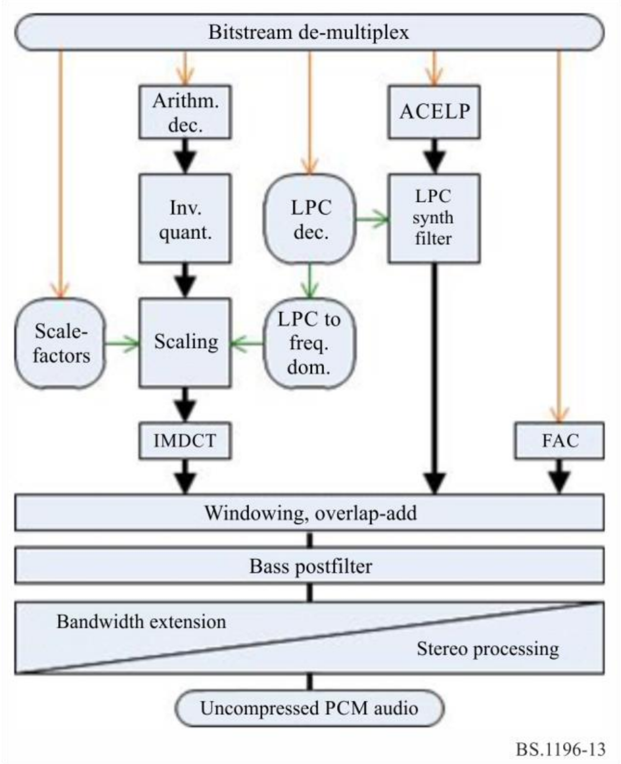


3 Decoding

The basic structure of the MPEG-D USAC decoder is shown in Fig. 13. The decoding process generally follows the inverse path of the encoding process.

FIGURE 13

MPEG-D USAC decoder block diagram



The process of decoding can be coarsely outlined as follows:

– Bitstream de-multiplex: The decoder finds all tool related information in the bitstream and forwards them to the respective decoder modules.

– Core decoding: Depending on the bitstream content, the decoder either:

  – decodes and inverse quantizes the MDCT spectral coefficients, applies scaling either based on scalefactor information or based on LPC coefficient information, and applies further (optional) MDCT based tools if present and applicable. Finally the inverse MDCT is applied to obtain the corresponding time domain signal.

  – or decodes ACELP related information, produces an excitation signal and synthesizes an output signal with the help of an LPC filter.

– Windowing, overlap-add: The subsequent frames of the core coder are concatenated or merged in the usual overlap-add process as known from AAC. Transitions between ACELP and MDCT based coding is accomplished by merging the decoded FAC data.

– Bass postfilter: An optional pitch enhancement filter can be applied to enhance speech quality.

– Bandwidth extension, stereo processing: At the end the parametric coding tools for bandwidth extension and stereo coding tools are applied to reconstruct the full bandwidth, discrete stereo signal.

For each of the optional tools, the option to “pass through” is retained, and in all cases where an operation is omitted, the data at its input is passed directly through the tool without modification.

4 Profiles and levels

MPEG currently defines two profiles which employ the USAC codec.

– *Baseline USAC profile*

The baseline USAC profile contains the complete USAC codec except for a small number of tools which exhibit excessive worst-case computational complexity. These tools are not described above. This profile provides a clear stand-alone profile for applications and use cases where the capability of supporting the AAC family of profiles (AAC profile, HE AAC profiles, HE AAC v2 profile) is not relevant.

– *Extended HE AAC profile*

The extended high efficiency AAC profile contains all of the tools of the high efficiency AAC v2 profile and is as such capable of decoding all AAC family profile streams. In addition, the profile incorporates mono/stereo capability of the baseline USAC profile. Consequently, this profile provides a natural evolution of the HE AAC v2 profile because the mono/stereo part of USAC (when operated at low rates) provides the additional value of consistent performance across content types at low bitrates.

Annex 6
(informative)

Coding Independent Code Points (CICP) for MPEG coding

1 Introduction

ISO/IEC23001-8:2013 describes the coding aspects of audio programmes that are independent of the coded representation including the position and layout of loudspeaker systems. Default channel configurations include the channel configurations specified in Recommendations ITU-R BS.775 or ITU-R BS.2051. All channel configurations are shown in Table 3.

TABLE 3

Channel configurations and loudspeaker layouts (Note 1)

Channel configuration Value[*1] (Note 1)	**Number of speakers (Note 2)**	**Default element to loudspeaker mapping (Note 3)**	**Channel name specified in Recs ITU-R BS.775 or BS.2051 (Note 4)**
0	Any setup		
1	1/0.0 (0+1+0)	M+000	Mono
2	2/0.0 (0+2+0)	M+030	Left
		M-030	Right
3	3/0.0 (0+3+0)	M+000	Centre
		M+030	Left
		M-030	Right

TABLE 3 (*cont.*)

Channel configuration Value[*1] **(Note 1)**	**Number of speakers (Note 2)**	**Default element to loudspeaker mapping (Note 3)**	**Channel name specified in Recs ITU-R BS.775 or BS.2051 (Note 4)**
4	3/1.0 (0+4+0)	M+000	Centre
		M+030	Left
		M-030	Right
		M+180	Mono surround
5	3/2.0 (0+5+0)	M+000	Centre
		M+030	Left
		M-030	Right
		M+110	Left surround
		M-110	Right surround
6	3/2.1 (0+5+0)	M+000	Centre
		M+030	Left
		M-030	Right
		M+110	Left surround
		M-110	Right surround
		LFE1	Low frequency effects
7	5/2.1 (0+7+0)	M+000	N/A[*2]
		M+030	
		M-030	
		M+045	
		M-045	
		M+110	
		M-110	
		LFE1	
8	1+1	Channel 1	N/A
		Channel 2	
9	2/1.0 (0+3+0)	M+030	Left
		M-030	Right
		M+180	Mono surround
10	2/2.0 (0+4+0)	M+030	Left
		M-030	Right
		M+110	Left surround
		M-110	Right surround

TABLE 3 (*cont.*)

Channel configuration Value[*1] **(Note 1)**	**Number of speakers (Note 2)**	**Default element to loudspeaker mapping (Note 3)**	**Channel name specified in Recs ITU-R BS.775 or BS.2051 (Note 4)**
11	3/3.1 (0+6+0)	M+000	N/A
		M+030	
		M-030	
		M+110	
		M-110	
		M+180	
		LFE1	
12	3/4.1 (0+7+0)	M+000	Centre
		M+030	Left
		M-030	Right
		M+090	Left side surround
		M-090	Right side surround
		M+135	Left rear surround
		M-135	Right rear surround
		LFE1	Low frequency effects
13	11/11.2 (9+10+3)	M+000	Front centre
		M+030	Front left centre
		M-030	Front right centre
		M+060	Front left
		M-060	Front right
		M+090	Side left
		M-090	Side right
		M+135	Back left
		M-135	Back right
		M+180	Back centre
		LFE1	Low frequency effects-1
		LFE2	Low frequency effects-2
		U+000	Top front centre
		U+045	Top front left
		U-045	Top front right
		U+090	Top side left
		U-090	Top side right
		T+000	Top centre
		U+135	Top back left
		U-135	Top back right
		U+180	Top back centre
		B+000	Bottom front centre
		B+045	Bottom front left
		U-045	Bottom front right

TABLE 3 (*cont.*)

Channel configuration Value^{*1} **(Note 1)**	**Number of speakers (Note 2)**	**Default element to loudspeaker mapping (Note 3)**	**Channel name specified in Recs ITU-R BS.775 or BS.2051 (Note 4)**
14	5/2.1 (2+5+0)	M+000	Centre
		M+030	Left
		M-030	Right
		M+110	Left surround
		M-110	Right surround
		LFE1	Low frequency effects
		U+030	Left top front
		U-030	Right top front
15	5/5.2 (3+7+0)	M+000	Centre
		M+030	Left
		M-030	Right
		M+090	Left side
		M-090	Right side
		M+135	Left back
		M-135	Right back
		U+045	Left height
		U-045	Right height
		UH+180	Centre height
		LFE1	Left low frequency effects
		LFE2	Right low frequency effects
16	5/4.1 (4+5+0)	M+000	Centre
		M+030	Left
		M-030	Right
		M+110	Left surround
		M-110	Right surround
		LFE1	Low frequency effects
		U+030	Left top front
		U-030	Right top front
		U+110	Left top rear
		U-110	Right top rear
17	6/5.1 (6+5+0)	M+000	N/A
		M+030	
		M-030	
		M+110	
		M-110	
		LFE1	
		U+000	
		U+030	
		U-030	
		U+110	
		U-110	
		T+000	

TABLE 3 (*cont.*)

Channel configuration Value^{*1} **(Note 1)**	**Number of speakers (Note 2)**	**Default element to loudspeaker mapping (Note 3)**	**Channel name specified in Recs ITU-R BS.775 or BS.2051 (Note 4)**
18	6/7.1 (6+7+0)	M+000	N/A
		M+030	
		M-030	
		M+110	
		M-110	
		M+150	
		M-150	
		LFE1	
		U+000	
		U+030	
		U-030	
		U+110	
		U-110	
		T+000	
19	5/6.1 (4+7+0)	M+000	Centre
		M+030	Left
		M-030	Right
		M+090	Left side surround
		M-090	Right side surround
		M+135	Left rear surround
		M-135	Right rear surround
		LFE	Low frequency effects
		U+045	Left top front
		U-045	Right top front
		U+135	Left top back
		U-135	Right top back
20	7/6.1 (4+9+0)	M+000	Centre
		M+SC	Left screen
		M-SC	Right screen
		M+030	Left
		M-030	Right
		M+090	Left side surround
		M-090	Right side surround
		M+135	Left rear surround
		M-135	Right rear surround
		LFE	Low frequency effects
		U+045	Left top front
		U-045	Right top front
		U+135	Left top back
		U-135	Right top back

TABLE 3 (*end*)

Channel configuration Value[*1] **(Note 1)**	**Number of speakers (Note 2)**	**Default element to loudspeaker mapping (Note 3)**	**Channel name specified in Recs ITU-R BS.775 or BS.2051 (Note 4)**
21-63	Reserved		

[*1] The audio output channel configuration is indicated by a six-bit field that carries a channel configuration value as defined in ISO/IEC 23001-8:2013, "Coding Independent Code Points".

[*2] N/A: not applicable; channel configuration not available in Recommendation ITU-R BS.2051 or in Recommendation ITU-R BS.775.

NOTE 1 – The list is derived from Table 8 of ISO/IEC 23001-8:2013 / Amd.1:2015.

NOTE 2 – The notion of the number of speaker is given in the convention of “Front speakers/ surround speakers. LFE speakers” and in brackets as “Upper layer speakers + middle layer speakers + bottom layer speakers” with LFE speakers excluded.

NOTE 3 – Identification of the speakers by labels in accordance with Recommendation ITU-R BS.2051.

NOTE 4 – The channel labels and names depend on the actual channel configuration.

Annex 7
(informative)

AC-4

1 Introduction

AC-4 is an advanced coding system for digital broadcasting that uses a digital compression algorithm and various parametric coding tools to improve efficiency and functionality. AC-4 natively supports both channel-based and object-based input and output formats.

Table 4 provides a list of the channel formats that AC-4 supports natively, which covers the channel configurations required for emission defined Recommendation in ITU-R BS.1548. AC-4 also supports the coding of higher-channel count formats that may be used to support advanced audio systems in Recommendation ITU-R BS.2051.

TABLE 4

Summary of native-supported channel formats

Format	**Number of channels**	**Note**
Mono (1/0 format*)	1	
Stereo (2/0 format*)	2	

TABLE 4 (*end*)

Format	Number of channels	Note
3.0 (3/0 format) [(1)]	3	
5.0/5.1 (3/2 format) [(1)]	5/6	
7.0/7.1 (System I) [(2)]	7/8	Signalling of three different speaker configurations
7.0.4/7.1+4 (System J) [(2)]	11/12	Signalling of subsets with fewer channels
9.0.4/9.1.4	13/14	Signalling of subsets with fewer channels
22.2 (System H) [(2)]	24	

[(1)] Specified in Recommendation ITU-R BS.775.

[(2)] Specified in Recommendation ITU-R BS.2051.

AC-4 is able to encode 1 to 22.2 channels of source PCM audio into a serial bitstream at data rates from 24 kbit/s to 1 536 kbit/s. In addition to supporting channel-based representations, AC-4 also supports coding of dynamic audio objects. A full description of the AC-4 bitstream syntax can be found in document in ETSI TS 103 190-2.

Table 5 provides suggested bit rates for the various channel configurations to meet requirements outlined in Recommendation ITU-R BS.1548.

TABLE 5

Summary of bit rates for specific channel configurations

Mode	Bit rates
2.0 Stereo	96 kbit/s
5.1 Surround	192 kbit/s
22.2 Surround	1536 kbit/s

AC-4 also natively supports the following system features:

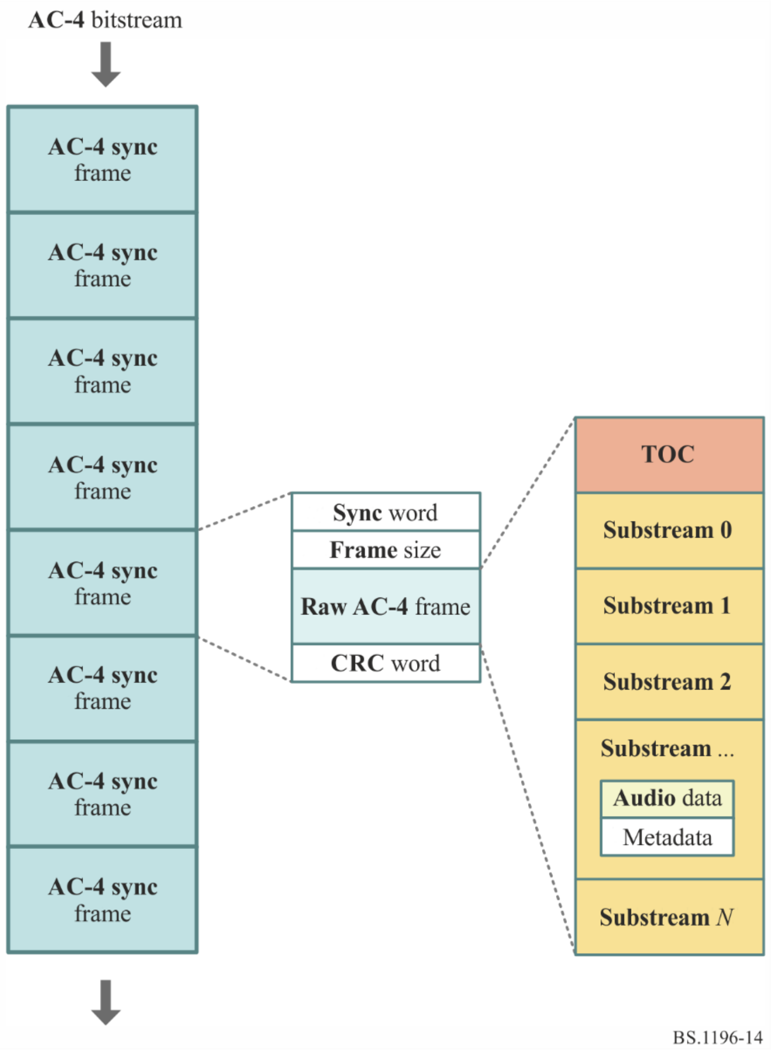
- Intelligent loudness management compliant with Recommendations ITU-R BS.1770 and ITU-R BS.1771, including signalling to indicate compliance with international and several regional loudness regulations currently in force.
- Support for encoding and decoding of channel and object-based audio representations
- Support for the listening environments as required in Recommendation ITU-R BS.1909, namely “Home environments” and “Mobile environments”. Including advanced dynamic range control applicable to a wide range of device types for both “home” and “mobile” environments.
- Dialogue enhancement.
- Video frame synchronous coding, enabling audio frames to be aligned with video frames.

- Native support for the carriage and signalling of enhanced ancillary data or metadata.

The AC-4 bitstream, as shown in Fig. 14, consists of AC-4 synchronization frames, which begins with a sync word and ends with a cyclic redundancy check (CRC) word. The sync word allows the decoder to easily identify the AC-4 frame begin decoding while the CRC word allows the decoder to detect the occurrence of bitstream errors and perform any required error concealment. The actual codec frame or "Raw AC-4" frame consists of the TOC (Table of Contents) and at least one substream.

FIGURE 14

High-Level Bitstream Syntax



Each substream includes the coded audio data as well as the associated metadata (ancillary data). The TOC contains the necessary information about how the substream, or multiple substreams should be decoded.

2 Encoding

The AC-4 encoder is not normatively specified, but a variety of coding tools are supported so that a compliant encoded bitstream can be generated.

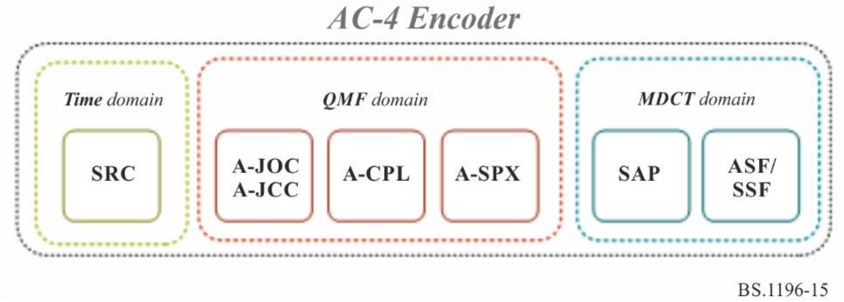
AC-4 uses transform domain quantization and coding using a modified discrete cosine transform (MDCT) with flexible window switching and parametric coding in a pseudo quadrature mirror filter bank. AC-4 supports encoding of 1 to 22.2 channels of source PCM audio into a serial bitstream at data rates from 24 kbit/s to 1536 kbit/s. For stereo 2.0 and 5.1 channel encoding, datarates of 96 kbit/s and 192 kbit/s meet the performance requirements specified in Recommendation ITU-R BS.1548. AC-4 also supports three bit allocation modes: Constant bit rate, Average bit rate and Variable bit rate.

The AC-4 encoder may be implemented with a variety of coding tools, shown in Fig. 15, to improve efficiency and flexibility/functionality depending on the mode of operations and/or application. The order of how the tools are implemented follows from left to right, meaning that PCM audio would be input to the tools on the left first and an AC-4 encoded bitstream would be output on the right. A description of the coding tools are provided below:

- SRC: The Sample Rate Converter tool is required in the AC-4 encoder to enable a frame duration that matches a video frame for frame synchronous coding modes. Depending on the frame rate, the input signal is converted to one of the internal sample rates of 46,034 Hz, 46,080 Hz, 48,000 Hz or 51,200 Hz used by the subsequent QMF and MDCT-based encoder tools.
- A-JCC: The Advanced Joint Channel Coding tool performs a downmix of the immersive channel (greater than 5 channel) input to a fewer number of channels and encoded the associated parameters. The parameters enable the reconstruction of all the input channel by the decoder.
- A-JOC: The Advanced Joint Object Coding tool takes audio objects as input and spatially encodes those objects to produce a smaller number of output objects, therefore reducing the number of MDCT-coded signals. The parameters that are encoded with the reduced number of output objects enable the reconstruction of the objects in the decoder.
- A-CPL: The Advanced Coupling tool performs a downmix from 2-channels to 1-channel and encodes associated parameters that enables the reconstruction of the original 2-channel input.
- A-SPX: The Advanced Spectral Extension tools encodes parameters associated with the high-frequency content of the input signal and then is spectrally extended in the decoder. The parameters comprise of envelope, tonality and noise measures. The spectral and temporal resolution of the parameters can be adapted to the characteristics of the input signal.
- SAP: The Stereo Audio Processing tool performs joint channel coding in the MDCT domain between two or more input channels.
- ASF: The Audio Spectral Front-end tool is an MDCT based quantization and coding tool that utilizes transform window switching. The window switch control module selects the optimum transform length for a frame depending on the type of input signal. MDCT coefficients and additional control information are stored after nonlinear quantization and noiseless coding to the bitstream. The bit distribution over time and the frequency are both controlled by a bit buffer and a perceptual model. The bit buffer model also takes into account bits used by the other encoder tools and general metadata.
- SSF: The Speech Spectral Front-end tool is an alternative MDCT based quantization and coding tool, specific for speech coding operating on short transforms. It performs quantization and coding in the MDCT domain with the use of a sub-band predictor. The SSF and ASF tools are mutually exclusive, therefore either ASF or SSF is used to encode the MDCT.

FIGURE 15

Encoding tools that are available in the AC-4 system

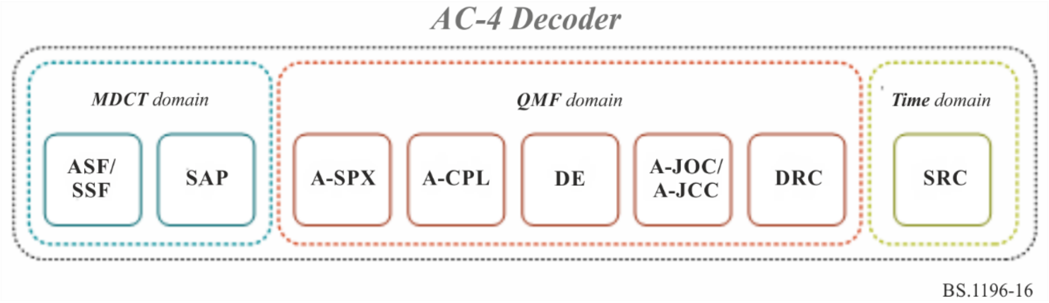


3 Decoding

The decoding process is basically the inverse of the encoding process and is shown in Fig. 16.

FIGURE 16

Decoding tools



The AC-4 bistream and decoder design natively support implementing lower complexity decoders for supporting devices with limited capabilities (e.g. mobile phones/tablets).

Two methods are utilized to support this capability:

1 Core decoding: AC-4 supports two decoding modes, Full and Core. The AC-4 decoder is able to operate in a core decoding mode, in which a core subset of channels (representing the contents of all the input channels) of the encoded program are decoded enabling a compatible reproduction of the program with reduced computational complexity.

2 Scalable decoding: AC-4 also supports sample rate scalability, where AC-4 is able to support higher sample rates, specifically 96 kHz and 192 kHz, in a scalable manner. Devices that support only 48 kHz outputs only need to decode the base layer.

In addition to supporting the decoding of the natively supported channel format, as shown in Table 4, AC-4 supports the use of ancillary data or metadata. This allows down mixing of the decoded channel output from a higher channel count to a lower output channel count as required for the device in a predictable manner.

Annex 8
(informative)

MPEG-H 3D Audio Low Complexity Profile

1 Introduction

MPEG-H 3D Audio is an audio coding standard developed to support coding audio as audio channels, audio objects, or Higher Order Ambisonics (HOA) and provides solutions for loudness normalization and dynamic range control. Each content type (channels, objects, or HOA) can be used alone or in combination with the other ones. The use of audio channel groups, objects or HOA allows for interactivity or personalization of a program, e.g. by selecting different language tracks or adjusting the gain or position of the objects during rendering in the MPEG-H decoder.

The MPEG-H 3D Audio specification is published as ISO/IEC 23008-3:2015. Amendment 3, specifying the Low Complexity Profile (LC Profile) of MPEG-H 3D Audio and additional technology was published in early 2017.

MPEG-H 3D Audio LC Profile can support up to 24 output loudspeakers and 56 codec core channels (out of which 28 channels can be decoded at a time).

Examples for possible target loudspeaker layouts:

- 2.0 Stereo (2/0 format specified in Recommendation ITU-R BS.775).
- 5.1 Multi-channel audio (3/2 format specified in Recommendation ITU-R BS.775).
- 10.2 Immersive audio (system F specified in Recommendation ITU-R BS.2051).
- 22.2 Immersive audio (system H specified in Recommendation ITU-R BS.2051).

The standard may be used in a wide variety of applications including stereo and surround sound storage and transmission. Its support for interactivity and immersive sound is important to satisfy the requirements of next-generation media delivery, particularly new television broadcast systems and entertainment streaming services as well as for virtual reality content and services.

For example, in TV broadcasting, commentary or dialogue may be sent as audio objects and combined with an immersive channel bed in the MPEG-H 3D Audio decoder. This allows efficient transmission of dialogue in multiple languages and also allows the listener to adjust the balance between dialogue and other sound elements to their preference. This concept can be extended to other elements not normally present in a broadcast, such as audio description for the visually impaired, director's commentary, or to dialogue from participants in sporting events.

MPEG-H 3D Audio LC Profile supports loudness management compliant with Recommendations ITU-R BS.1770 and ITU-R BS.1771, including compliance signalling to indicate compliance with international and several regional loudness regulations. It also supports advanced dynamic range control (DRC) for a wide range of device types for both “home” and “mobile” environments.

2 Encoding

The MPEG-H 3D Audio codec architecture is built upon a perceptual codec for compression of the different input signal classes, based on MPEG Unified Speech and Audio Coding (USAC). USAC allows for compression of mono to multi-channel audio signals at rates of 8 kbit/s per channel and higher.

For the new requirements that arose in the context of 3D audio, this technology has been extended by tools that especially exploit the perceptual effects of 3D reproduction and thereby further enhance the coding efficiency, such as:

– An enhanced noise filling is provided through Intelligent Gap Filling (IGF). IGF is a tool that parametrically restores portions of the transmitted spectrum using suitable information from spectral tiles that are adjacent in frequency and time. The assignment and the processing of these spectral tiles are controlled by the encoder based on an input signal analysis. Hereby, spectral gaps can be filled with spectral coefficients that perceptually have a better match than pseudo random noise sequences of conventional noise filling would provide.

– Apart from the enhancements in coding efficiency, the USAC-3D core is equipped with new signalling mechanisms for 3D content/loudspeaker layouts and for the type of signals in the compressed stream (audio channel vs. audio object vs. HOA signal).

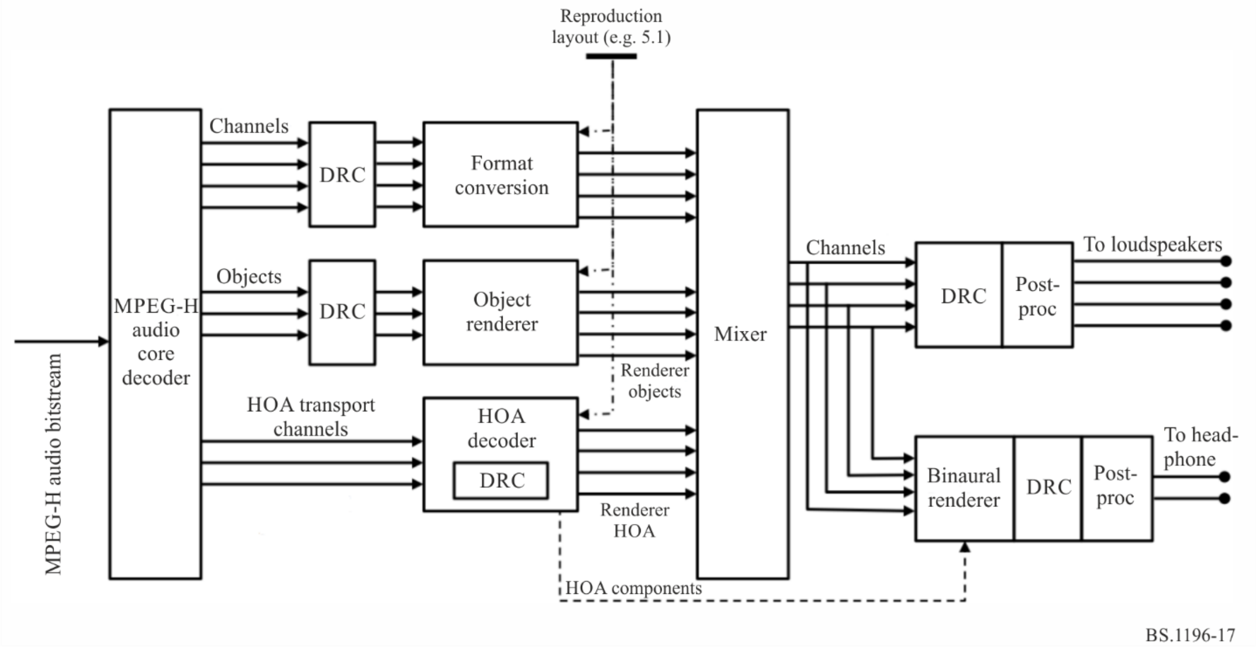
Another new aspect in the design of the compressed audio payload is an improved behaviour for instantaneous rate switching or fast cue-in as it appears in the context of MPEG Dynamic Adaptive Streaming (DASH). For this purpose, so-called ‘immediate playout frames’ have been added to the syntax that enable gapless transitions from one stream to the other. This is particularly advantageous for adaptive streaming over IP networks.

3 Decoding

A top-level block diagram of the overall MPEG-H 3D Audio LC Profile audio decoder architecture is depicted in Fig. 17.

FIGURE 17

Top level block diagram of the MPEG-H 3D LC Profile audio decoder architecture



The main components are a so-called USAC-3D core decoder, a set of renderers for the different signal classes and mixer. In a first stage, the different base signals are converted from their data-compressed representation by means of a so-called USAC-3D decoder.

The different signal classes (waveforms for channel signals and object signals or HOA coefficient signals) are then fed to their associated renderers that map those signals to loudspeaker feeds for the particular reproduction setup that is available at the receiver side. As soon as all rendered signals are available in the reproduction format, they are combined in a mixing stage to form a loudspeaker feed. In case a binaural representation is requested, the signal is converted to a virtual 3D scene for headphone reproduction. It is possible to transmit any combination of the different signal types in a single MPEG-H stream, for instance a combination of channel signals with object signals or an HOA scene with objects.

The renderers are:

– A format converter for converting channel signals from their production speaker format to the reproduction speaker layout.

– An object renderer to place static or dynamic object tracks into the reproduction layout.

– An HOA renderer to convert from the scene based HOA representation to the actual reproduction layout.

– Binaural rendering to convert from a virtual loudspeaker layout to headphone output

In addition, playback and rendering of the different signal classes can be controlled by a user interface, if the corresponding metadata marks these signals as enabled for interactivity.

Annex 9 (informative)

DTS-UHD

1 Introduction

The DTS-UHD coding system includes Audio Coding Engine (ACE) as an efficient method to compress audio waveforms. DTS-UHD supports encoding of Channel-based Audio (CBA), Ambisonic sound fields and Object-based audio (OBA). DTS-UHD has been standardized in ETSI TS 104 491, and forms part of the DVB MPEG 2 transport stream specification ETSI TS 101 154. The system provides a fully immersive experience, and through the use of objects, channels, soundfields, or a combination of all three, delivers interactivity and personalisation to enhance the experience. It also provides accessibility services support for visually and hearing-impaired users.

DTS-UHD supports all loudspeaker layouts specified in Recommendation ITU-R BS.2051.

While fully supporting channel-based audio, DTS-UHD is also capable of encoding up to 224 discrete audio objects, which can be further organized into as many as 32 object groups and 32 presentations within one stream.

A DTS-UHD Stream consists of a sequence of DTS-UHD frames containing three major elements:

– Frame Table of Contents (FTOC) – This element allows a decoder to navigate directly to elements of interest with the frame;

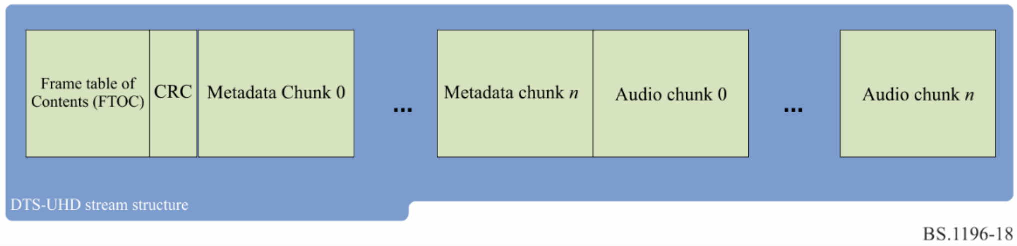
– Metadata chunk elements;

– Audio chunk elements.

The structure is outlined below:

FIGURE 18

DTS-UHD stream structure



The FTOC is present in every DTS-UHD Frame and indicates whether the frame is a sync frame or non-sync frame, with sync frames providing all parameters necessary to initiate a decoding session.

The audio chunks carry the compressed waveforms which are optimally compressed and carried and can be either sync frames or non-sync frames. The audio chunk is organized into a number of substreams that can represent mono waveforms, stereo waveforms, and LFE waveforms.

The Metadata chunk fully describes an audio component, including the type of component, which audio chunks needed to render the component, and information about loudness and dynamic range compression. A metadata chunk for an access point (a DTS-UHD sync frame) references an audio chunk that is also a sync frame and contains a full set of metadata needed to start decoding.

2 Encoder

The DTS-UHD ACE encoder is not normatively specified within ETSI. The encoder implements a full tool set able to supply multiple bitstreams in support of a number of different speaker configurations.

DTS-UHD encoding technology maintains the quality of audio and can produce bitstreams that fulfil high quality emission as specified in Recommendation ITU-R BS.1548 at the following bitrates:

Loudspeaker configuration	Bitrate (kbit/s)
0+2+0 (stereo)	128
0+5+0 (5.1)	192
4+7+0 (7.1.4)	288

The DTS-UHD encoder allows full control of the placement of sync frames to allow a service to have control of the entry points into the audio service.

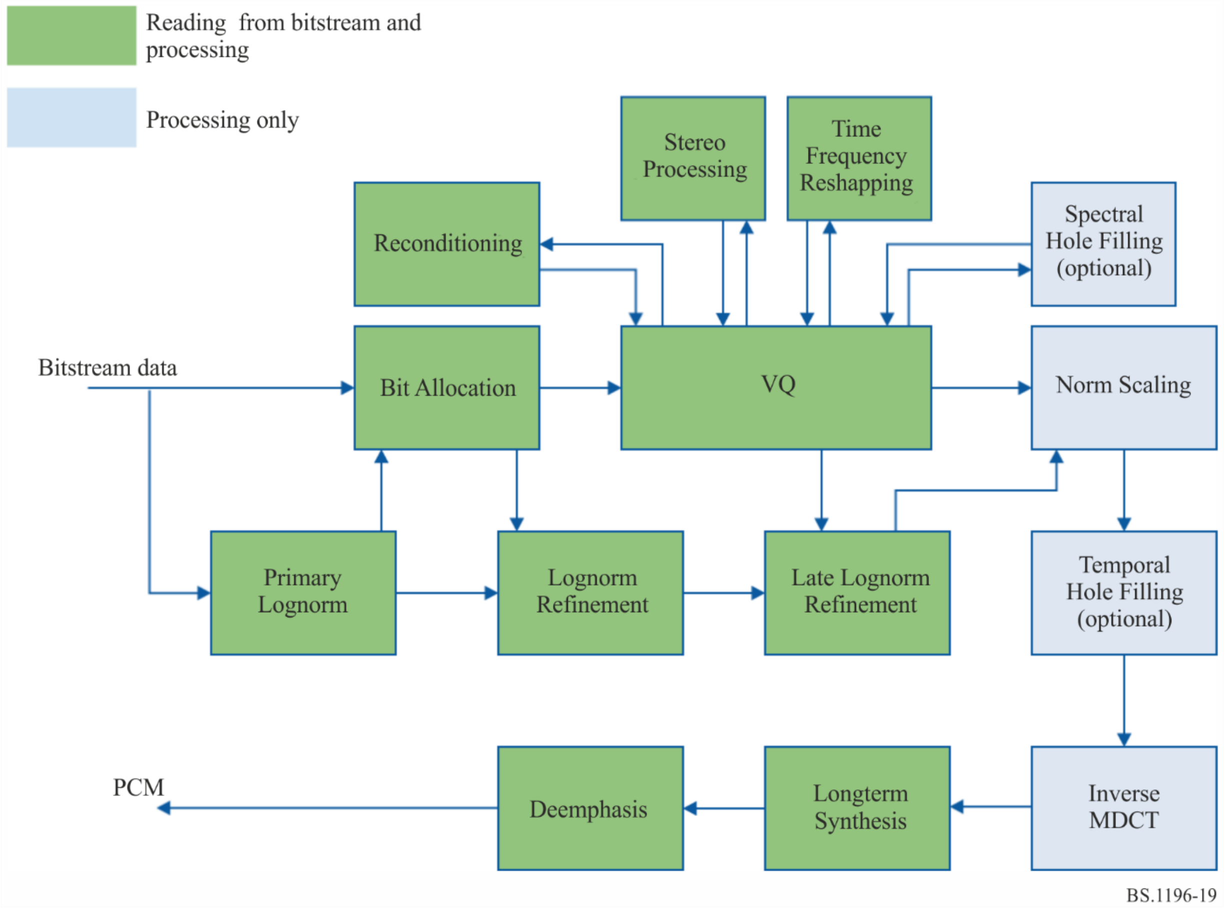
The DTS-UHD Variable Bit Rate (VBR) functionality guarantees Constant Bit Rate (CBR) over each GoF (Group of Frames – interval between two consecutive sync frames). The sync frames are set by the Encoder, and the sync interval can be dynamically changed during the encoding process as needed. The DTS-UHD Encoder can also set the maximum instantaneous VBR peak rate as required. The DTS-UHD Bit-Rate Control (BRC) model dynamically allocates bits within each audio frame across objects, channels and channel pairs within a bitstream. BRC directs more bits to encode the most perceptually relevant content in each frame while aiming for equal audio quality across all objects and channels within a frame. This allows for improved coding efficiency with an increased number of audio elements by taking advantage of varying distribution of perceptually relevant content across objects and channels. The Encoder also dynamically allocates bits to audio frames aiming for constant perceived audio quality along the time axis (within the constraints of GoF-wise CBR and specified max peak bitrate). Note that BRC functionality requires no additional information to be transmitted and that the Decoder is unaware of BRC functionality. Each object, channel or channel pair remains independently decodable and ‘fully exercisable’ in BRC mode.

3 Decoder

The ACE Decoder operates on ACE stereo streams, ACE mono streams and ACE LFE streams. Within an ACE frame, the streams have been separated into one of these three classifications. Stereo streams have an additional optimization that takes advantage of the information common to both channels.

FIGURE 19

ACE Decoder



Decoding a mono or stereo ACE stream involves the following stages:

Primary Lognorm: a first approximation of spectral band power is retrieved from the bitstream. These power values are represented in the logarithmic domain and referred to as lognorm values.

Bit Allocation: for each spectral band, the number of bits available for normalized spectral band decoding is computed. This computation depends on information in the bitstream and (potentially) primary lognorm values.

Lognorm Refinement: a small part of the bit allocation is used to refine the primary lognorm values.

VQ: (Vector Quantization), the majority of the bit allocation is used to reconstruct normalized spectral frequency values, one spectral band at a time.

Reconditioning: for originally “full” spectral bands and small bit allocation values, the initially reconstructed sparse bands are reconditioned to better approximate a full spectrum.

Stereo processing: for stereo streams, the two reconstructed channels are transformed to an original (unmixed) representation.

Time-Frequency Reshaping: the spectral bands are pre-processed by a simple Haar transform, bringing all bands to the same time-frequency representation.

Spectral Hole Filling: bands with original energy that are reconstructed to all-zeros are then filled with “noise”, generated from earlier reconstructed band values.

Late Lognorm Refinement: any bits that have been left over after VQ are used to do a final lognorm refinement.

Norm Scaling: spectral bands are restored to the power values corresponding to the final lognorm values.

Temporal Hole Filling: all-zero spectral bands are potentially filled with synthetic noise depending on the spectral band power in previous frames.

Inverse MDCT: the spectral representation is transformed in a time-domain representation (PCM).

Longterm Synthesis: a long-term synthesis filter is used to reconstruct signal components that have been removed at encoding time.