Implementation Guidelines for the use of Audio-Visual Content in DVB services delivered over IP

DVB Document A084

July 2004

Contents

Introduction

This document presents guidelines for the use of H.264/AVC and MPEG-4 High Efficiency AAC, as defined in ITU-T Rec. H.264 and ISO/IEC 14496-10 [1] for H.264/AVC, and for High Efficiency AAC in ISO/IEC 14496-3:2001 [2] and in ISO/IEC 14496-3:2001/AMD-1 [3].

This document does not address guidelines for the use of audio and video formats defined in ETR 101 154 [9] in DVB services over IP, such as MPEG-1 audio and MPEG-2 video. Applications that wish to use these formats for the encoding of audio and video streams must transport such streams contained in a DVB compliant MPEG-2 Transport Stream over IP. Consequently, the guidelines provided in ETR 101 154 [9] apply. For the transport of an MPEG-2 TS in RTP packets over IP, RFC 2250 [8] shall be used.

For delivery of H.264/AVC and High Efficiency AAC encoded content over IP, the following hierarchical classification of IP-IRDs is defined:

- **Capability A IP-IRD**s are capable of decoding bitstreams conforming to Baseline Profile at Level 1[1] with constraint_set1_flag being equal to 1 as specified in [1].

- **Capability B IP-IRD**s are capable of decoding bitstreams conforming to Baseline Profile at Level 1.2 with constraint_set1_flag being equal to 1 as specified in [1].

- **Capability C IP-IRD**s are capable of decoding bitstreams conforming to Baseline Profile at Level 2 with constraint_set1_flag being equal to 1 as specified in [1].

- **Capability D IP-IRD**s are capable of decoding bitstreams conforming to Main profile at level 3 as specified in [1].

- **Capability E IP-IRD**s are capable of decoding bitstreams conforming to Main profile at level 4 as specified in [1].

An IP-IRD of one of the capability classes above shall meet the minimum functionality, as defined in this specification, for decoding of H.264/AVC and High Efficiency AAC delivered over an IP network. The specification of this minimum functionality in no way prohibits IP-IRD manufacturers from including additional features, and should not be interpreted as stipulating any form of upper limit to the performance.

Where an IP-IRD feature described in the present document is mandatory, the word "shall" is used and the text is in italic; all other features are optional. The guidelines presented for IP-IRDs observe the following principles:

- IP-IRDs allow for future compatible extensions to the bit-stream syntax;

- all "reserved", "unspecified", and "private" bits in H.264/AVC, High Efficiency AAC and IP protocols shall be ignored by IP-IRDs not designed to make use of them.

The rules of operation for the encoders are features and constraints which the encoding system should adhere to in order to ensure that the transmissions can be correctly decoded. These constraints may be mandatory or optional. Where a feature or constraint is mandatory, the word "shall" is used and the text is italic; all other features are optional.

Clauses 4 to 6 provide the Digital Video Broadcasting (DVB) guidelines for the systems, video, and audio layer, respectively. For information, some of the key features are summarized below, but Clauses 4 to 6 should be consulted for all definitions:

Systems:

- H.264/AVC and High Efficiency AAC encoded data is delivered over IP in RTP packets.

Video:

- Capability A, B, and C IP-IRDs support the H.264/AVC Baseline Profile with constraint_set1_flag equal to 1.

---

[1] with MaxBR value equal to 128 and with MaxCPB value equal to 350

- Capability D and E IP-IRDs support the H.264/AVC Main Profile.

- IP-IRDs labelled with a particular capability Y are also capable of decoding and displaying pictures that can be decoded by IP-IRDs labelled with a particular capability X with X being an earlier letter than Y in the alphabet. For instance, Capability D IP-IRDs are capable of decoding bitstreams conforming to Main Profile at level 3 of H.264/AVC and below. Additionally, Capability D IP-IRDs are capable of decoding bitstreams that are also decodable by IP-IRDs with capabilities A, B, or C.

Audio:

- Use of the MPEG-4 Audio High Efficiency AAC Profile;

- Sampling rates between 8 kHz and 48 kHz are supported by IP-IRDs;

- IP-IRDs support mono, 2-channel stereo; support of multi-channel is optional.

See Annex A for a description of these Implementation Guidelines.

1 Scope

The present document provides implementation guidelines for the use of H.264/AVC and High Efficiency AAC for DVB compliant delivery in RTP packets over IP networks. Guidelines are given for the decoding of H.264/AVC and High Efficiency AAC in IP-IRDs, as well as rules of operation that encoders should apply to ensure that transmissions can be correctly decoded. These guidelines and rules may be mandatory, recommended or optional.

2 References

For the purpose of this ETR the following references apply. References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific. For a specific reference, subsequent revisions do not apply. For a non-specific reference, the latest version applies. A non-specific reference to an ETS refers to the latest version published as an EN with the same number.

[1]	ITU-T Recommendation H.264 "Advanced Video Coding for Generic Audiovisual Services" & ISO/IEC 14496-10 (2003): "Information Technology - Generic Coding of moving pictures and associated audio - Part 10: Advanced Video Coding".
[2]	ISO/IEC 14496-3 (2001): "Information technology - Generic coding of moving picture and associated audio information - Part 3: Audio".
[3]	ISO/IEC 14496-3:2001/AMD-1: Bandwidth Extension
[4]	RFC 3550: RTP, A Transport Protocol for Real Time Applications
[5]	RFC 3016: RTP payload format for MPEG-4 Audio/Visual streams
[6]	RFC 3640: RTP payload for transport of generic MPEG-4 content
[7]	RFC yyyy: RTP payload for transport of H.264/AVC Content
[8]	RFC 2250: RTP payload format for MPEG1/MPEG2 Video
[9]	ETSI TR 101 154: "Digital Video Broadcasting (DVB); Implementation guidelines for the use of MPEG-2 Systems, Video and Audio in satellite, cable and terrestrial broadcasting applications".
[10]	EBU Recommendation R.68: "Alignment level in digital audio production equipment and in digital audio recorders".
[11]	3GPP TS 26.234: "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Transparent end-to-end packet switched streaming service (PSS); Protocols and codecs (Release 5)".

3 Definitions and abbreviations

3.1 Definitions

For the purposes of the present document, the following terms and definitions apply:

IP-IRD: an Integrated Receiver-Decoder for DVB services delivered over IP.

Capability A IP-IRD: an IP-IRD that is capable of decoding and displaying pictures using H.264/AVC encoded bitstreams conforming to Baseline Profile at processing and memory limits less than or equal to those of Level 1 with a modified MaxBR value being equal to 128 and with a modified MaxCPB value being equal to 350 and with constraint_set1_flag being equal to 1.

Capability B IP-IRD: an IP-IRD that is capable of decoding and displaying pictures using H.264/AVC encoded bitstreams conforming to Baseline Profile at processing and memory limits less than or equal to those of Levels 1 to 1.2 with constraint_set1_flag being equal to 1.

Capability C IP-IRD: an IP-IRD that is capable of decoding and displaying pictures from H.264/AVC encoded bitstreams conforming to Baseline Profile at processing and memory limits less than or equal to those of Levels 1 to 2 with constraint_set1_flag being equal to 1.

Capability D IP-IRD: an IP-IRD that is capable of decoding and displaying pictures from H.264/AVC encoded bitstreams conforming to Main Profile at processing and memory limits less than or equal to those of Levels 1 to 3.

Capability E IP-IRD: an IP-IRD that is capable of decoding and displaying pictures from H.264/AVC encoded bitstreams conforming to Main Profile at processing and memory limits less than or equal to those of Levels 1 to 4.

3.2 Abbreviations

For the purposes of the present document, the following abbreviations apply:

AAC LC	Advanced Audio Coding Low Complexity
H.264/AVC	H.264/Advanced Video Coding
CIF	Common Interchange Format
DVB	Digital Video Broadcasting
IRD	Integrated Receiver-Decoder
HDTV	High Definition Television
MPEG	Moving Pictures Experts Group (ISO/IEC JTC 1/SC 29/WG 11)
QCIF	Quarter Common Interchange Format
SBR	Spectral Band Replication
SDTV	Standard Definition Television
VCEG	Video Coding Experts Group (ITU-T SG16 Advanced Video Coding)

4 Systems layer

4.1 Transport over IP Networks

When H.264/AVC and High Efficiency AAC data are transported over IP networks, RTP, a Transport Protocol for Real-Time Applications as defined in RFC 3550 [4], shall be used. This clause describes the guidelines and requirements for transport of H.264/AVC and High Efficiency AAC in RTP packets for delivery over IP networks and for decoding of such RTP packets in the IP-IRD.

While the general RTP specification is defined in RFC 3550 [4], RTP payload formats are codec specific and defined in separate RFCs. In clause 4.2 guidelines and requirements for transport of H.264/AVC and High Efficiency AAC in RTP packets are defined. Clause 4.3 discusses some issues related to RTP usage.

The IP-IRD design should be made under the assumption that any legal structure as permitted RTP packets may occur, even if presently reserved or unused. *To allow full upward compatibility with future enhanced versions, a DVB IP-IRD shall be able to skip over data structures which are currently "reserved", or which correspond to functions not implemented by the IP-IRD. For example, an IP-IRD shall allow the presence of unknown MIME format parameters for RFC payloads, while ignoring its meaning.*

4.2 RTP payload formats

For transport over IP networks, H.264/AVC data and High Efficiency AAC data are contained in RTP packets as defined in RFC 3550 [4]. The specific formats of the RTP packets are defined in 4.2.1 for H.264/AVC and in 4.2.2 for High Efficiency AAC.

4.2.1 RTP packetization of H.264/AVC

For transport over IP, the H.264/AVC data is packetized in RTP packets using RFC yyyy [7].

Encoding: When transporting H.264/AVC video over IP, RFC yyyy [7] shall be used.

Decoding: Each IP-IRD shall be able to receive and decode RTP packets with H.264/AVC data as defined in RFC yyyy [7].

4.2.2 RTP packetization of High Efficiency AAC

Encoding: When transporting High Efficiency AAC audio over IP, either RFC 3016 [5] or RFC 3640 [[6]] shall be used.

Decoding: Each IP-IRD shall support both RFC 3016 [5] and RFC 3640 [[6]] to receive and decode High Efficiency AAC data contained in RTP packets.

5 Video

This Clause describes the guidelines for H.264/AVC video encoding and for decoding of H.264/AVC data in the IP-IRD. The bitstreams resulting from H.264/AVC encoding shall conform to the corresponding profile specification in [1]. The IP-IRD shall allow any legal structure as permitted by the specifications in [1] in the encoded video stream even if presently "reserved" or "unused".

To allow full compliance to the specifications in [1] and upward compatibility with future enhanced versions, an IP-IRD shall be able to skip over data structures which are currently "reserved", or which correspond to functions not implemented by the IP-IRD.

5.1 Profile and Level

Encoding: The H.264/AVC Baseline Profile with constraint_set1_flag being equal to 1 or the H.264/AVC Main Profile may be used to encode video conforming to H.264/AVC Levels 1 to 2. To encode video at a higher Level than Level 2, the H.264/AVC Main Profile may be used. For compatibility with Capability A, B and C IP-IRDs, it is recommended not to use Main Profile for Levels 1 to 2.

Decoding: *Each Capability A IP-IRD shall be capable of decoding and displaying pictures using H.264/AVC encoded bitstreams conforming to Baseline Profile at processing and memory limits less than or equal to those of Level 1 with a modified MaxBR value being equal to 128 and with a modified MaxCPB value being equal to 350 and with constraint_set1_flag being equal to 1.*

Each Capability B IP-IRD shall be capable of decoding and displaying pictures using H.264/AVC encoded bitstreams conforming to Baseline Profile at processing and memory limits less than or equal to those of Levels 1 to 1.2 with constraint_set1_flag being equal to 1.

Each Capability C IP-IRD shall be capable of decoding and displaying pictures from H.264/AVC encoded bitstreams conforming to Baseline Profile at processing and memory limits less than or equal to those of Levels 1 to 2 with constraint_set1_flag being equal to 1.

Each Capability D IP-IRD shall be capable of decoding and displaying pictures from H.264/AVC encoded bitstreams conforming to Main Profile at processing and memory limits less than or equal to those of Levels 1 to 3.

Each Capability E IP-IRD shall be capable of decoding and displaying pictures from H.264/AVC encoded bitstreams conforming to Main Profile at processing and memory limits less than or equal to those of Levels 1 to 4.

NOTE – Capability D and E IP-IRDs are also capable of decoding and displaying pictures from H.264/AVC encoded bitstreams conforming Capability A-C IP-IRDs.

5.2   Frame Rate

Encoding: To encode video, each frame rate allowed by the applied H.264/AVC Profile and Level may be used. The maximum time distance between two pictures should not exceed 0.7 s.

Decoding: *Each IP-IRD shall support each frame rate allowed by the H.264/AVC Profile and Level that is applied for decoding in the IP-IRD.* This includes variable frame rate.

5.3   Aspect Ratio

Encoding: To encode video, each sample and picture aspect ratio allowed by the applied H.264/AVC Profile and Level may be used. It is recommended to avoid very large or very small picture aspect ratios and that those picture aspect ratios specified in [9] are being used.

Decoding: *Each IP-IRD shall support each sample and picture aspect ratio permitted by the applied H.264/AVC Profile and Level.*

5.4   Luminance resolution

Encoding: To encode video, each luminance resolution allowed by the applied H.264/AVC Profile and Level may be used.

Decoding: *Each IP-IRD shall support each luminance resolution permitted by the relevant H.264/AVC Profile and Level.*

5.5   Chromaticity

Encoding: It is recommended to specify the chromaticity coordinates of the colour primaries of the source using the syntax elements colour_primaries, transfer_characteristics, and matrix_coefficients in the VUI. ITU-R Recommendation BT.709 is recommended as the preferred colour primaries and transfer characteristics.

Decoding: *Each IRD shall be capable to decode any allowed values of colour_primaries, transfer_characteristics, and matrix_coefficients.* It is recommended that appropriate processing be included for the display of pictures.

5.6   Chroma

Encoding: It is recommended to specify the chroma locations using the syntax elements chroma_sample_loc_type_top_field and chroma_sample_loc_type_bottom_field in the VUI. It is recommended to use chroma sample type 0.

Decoding: *Each IRD shall be capable to decode any allowed values of chroma_sample_loc_type_top_field and chroma_sample_loc_type_bottom_field.* It is recommended that appropriate processing be included for the display of pictures.

5.7   Parameter Constraints

Encoding: For broadcast applications it is recommended that sequence and picture parameter sets are sent together with a random access point (e.g. an IDR picture) to be encoded at least once every 500 milliseconds. For multicast or streaming applications a maximum interval of 5 seconds between random access points should not be exceeded. When changing sequence or picture parameter sets, it is recommended to use different values for seq_parameter_set_id or pic_parameter_set_id as the previous active ones.

NOTE 1:   Increasing the frequency of sequence and picture parameter sets and IDR pictures will reduce channel hopping time but will reduce the efficiency of the video compression.

NOTE 2:   Having a regular interval between IDR pictures may improve trick mode performance, but may reduce the efficiency of the video compression.

6 Audio

This Clause describes the guidelines for encoding with the MPEG-4 AAC profile or MPEG-4 High Efficiency Audio Profile in DVB IP Network bit-streams, and for decoding this bit-stream in the IP-IRD.

The recommended level for reference tones for transmission is 18 dB below clipping level, in accordance with EBU Recommendation R.68 [10].

For High Efficiency AAC , the audio encoding shall conform to the requirements defined in ISO/IEC 14496-3:2001 [2] *and ISO/IEC14496-3:2001/AMD-1 [3] for the MPEG-4 Audio High Efficiency Profile.*

The IP-IRD design should be made under the assumption that any legal structure as permitted by ISO/IEC 14496-3:2001 [2] or ISO/IEC 14496-3:2001/AMD-1 [3] may occur in the broadcast stream even if presently reserved or unused. *To allow full compliance to ISO/IEC 14496-3:2001* [2] *and upward compatibility with future enhanced versions, a DVB IP-IRD shall be able to skip over data structures which are currently "reserved", or which correspond to functions not implemented by the IP-IRD. For example, an IP-IRD which is not designed to make use of the extension payload shall skip over that portion of the bit-stream.*

The following clauses are based on ISO/IEC 14496-3:2001 [2] (MPEG-4 audio) and ISO/IEC 14496-3:2001/AMD-1 [3] (Bandwidth Extension)].

6.1   Audio mode

Encoding: The audio shall be encoded in mono or 2-channel-stereo according to the functionality defined in the High Efficiency AAC Profile Level 2 or in multi-channel according to the functionality defined in the High Efficiency AAC Profile Level 4, as specified in ISO/IEC 14496-3:2001 [2] and ISO/IEC 14496-3:2001/AMD-1 [3]. A simulcast of a mono/stereo signal together with the multi-channel signal is optional.

Decoding: *Each IP-IRD shall be capable of decoding in mono or 2-channel-stereo of the functionality defined in the High Efficiency AAC Profile Level 2, as specified in ISO/IEC 14496-3:2001* [2] *and ISO/IEC 14496-3:2001/AMD-1 [3].* The support of multi-channel decoding in an IP-IRD is optional.

6.2   Profiles

Encoding: The encoder shall use either the AAC Profile or the High Efficiency AAC Profile. Use of the High Efficiency AAC Profile is recommended.

Decoding: *IP-IRDs shall be capable of decoding the High Efficiency AAC Profile.*

6.3 Bit rate

Encoding: Audio may be encoded at any bit rate allowed by the applied profile and selected Level.

Decoding: *Each IP-IRD shall support any bit rate allowed by the High Efficiency AAC Profile and selected Level.*

6.4 Sampling frequency

Encoding: Any of the audio sampling rates of the High Efficiency AAC Profile Level 2 may be used for mono and 2-channel stereo and of the High Efficiency AAC Profile Level 4 for multichannel audio.

Decoding: *Each IP-IRD shall support each audio sampling rate permitted by the High Efficiency AAC Profile Level 2 for mono and 2-channel stereo and of the High Efficiency AAC Profile Level 4 for multichannel audio.*

6.5 Dynamic Range Control

Encoding: The encoder may use the MPEG-4 AAC Dynamic Range Control (DRC) tool.

Decoding: *Each IP-IRD shall support the MPEG-4 AAC Dynamic Range Control (DRC) tool.*

6.6 Matrix Downmix

Decoding: *Each IP-IRD shall support the matrix downmix as defined in MPEG-4.*

Annex A: Description of the Implementation Guidelines Informative

A.1 Introduction

These guidelines specify how advanced audio and video compression algorithms may be used for DVB services delivered over IP. A wide range of potential applications are covered, ranging from low-resolution services delivered to small portable receivers all the way up to HDTV services.

These guidelines apply to all DVB services directly over IP without the use of an intermediate MPEG-2 Transport Stream. An example of this type of DVB service is DVB-H, using multi-protocol encapsulation. The corresponding guidelines for audio-visual coding for DVB services which use an MPEG-2 Transport Stream are given in TR 101 154 for distribution services and in TR 102 154 for contribution services. Examples of Transport Stream based DVB service are the familiar DVB-S, DVB-C and DVB-T transmissions.

The "systems layer" of these guidelines addresses issues related to transport and synchronization of advanced audio and video. The systems layer is based on the use of RTP, a generic Transport Protocol for Real-Time Applications as defined in RFC 3550. Use of RTP requires the definition of payload formats that are specific for each content format, and so the system layer defines which RTP payload formats to use for transport of advanced audio and video, as well as applicable constraints for that. Further information on the systems layer is given in section A.2.

The advanced video coding uses H.264/AVC, an algorithm of many names. The work began in ITU-T under the working name H.26L. At a similar time, ISO/IEC began considering Advanced Video Coding (AVC) within the MPEG-4 standard. The ITU-T and MPEG experts then agreed to form a Joint Video Team, referred to as JVT. Within ITU-T it is published as H.264, whilst ISO/IEC will publish it as Part 10 of the MPEG-4 specification, 14496-10. As with all ITU-T and ISO/IEC algorithms since H.261 and MPEG-1, the architecture is based on a motion-compensated block transform. Like MPEG-1 and MPEG-2, H.264/AVC has intra-coded pictures, predicatively coded pictures and bi-directionally coded pictures (known as I-, P- and B-frames in MPEG-1 and MPEG-2). However, H.264/AVC has smaller, dynamically selected block sizes to allow the encoder to represent both large and small moving objects more efficiently. It also provides multiple reference frames to allow the encoder to find the best match over several frames and it supports greater precision in the representation of motion vectors. The variable-length coding used to compress the picture and motion information is context-adaptive to give greater efficiency. For further information on the video codec see section A.3.

The advanced audio coding uses the MPEG-4 High Efficiency AAC Profile. This is derived from the MPEG-2 Advanced Audio Coding (AAC), first published in 1997 MPEG-4 AAC is closely based on MPEG-2 AAC but includes some further enhancements such as perceptual noise substitution to give better performance at low bit rates. The new MPEG-4 High Efficiency AAC Profile adds spectral band replication, to allow more efficient representation of high-frequency information by using the lower harmonic as a reference. For further information on the audio codec see section A.4.

A.2 Systems

Protocol Stack

For delivery of DVB Services over IP-based networks a protocol stack is defined in a suite of DVB specifications. The systems part of the guidelines defined in this Specification addresses only the part of the protocol stack that is related to the transport and synchronization of HE AAC audio and H.264/AVC video. This part of the DVB-IP protocol stack is given in figure A-1. For completeness, RTCP and RTSP are also included, as they are relevant for RTP usage, though there are no specific guidelines for RTCP and RTSP defined in this Specification.

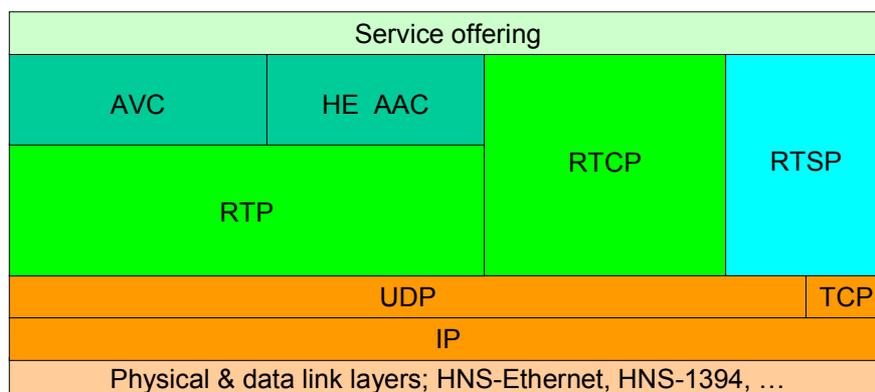| Service offering			
AVC	HE AAC	RTCP	RTSP
RTP			
UDP			TCP
IP			
Physical & data link layers; HNS-Ethernet, HNS-1394, …			

Figure A-1: The part of the DVB-IP protocol stack relevant for the transport of advanced audio and video (Note: Guidelines for RTCP and RTSP usage are beyond the scope of this specification)

Transport of High Efficiency AAC audio

The transport of HE AAC audio and H.264/AVC video is based on RTP, a generic Transport Protocol for Real-Time Applications as defined in RFC 3550 [4]. RFC 3550 defines the elements of the RTP transport protocol that are independent of the data that is transported, while separate RFCs define how to use RTP for transport of specific data such as audio and video coded. Both for HE AAC audio and for H.264/AVC video RTP payload formats are (being) defined.

To transport HE AAC, both RFC 3016 [5] and RFC 3640 [[6]] can be used. RFC 3016 offers compatibility with AAC services that comply with Release 5 of 3GPP specification for Packet Switched Streaming Services [11]. Note that these 3GPP services use AAC only and not the High Efficiency extension with SBR. RFC 3016 allows for the carriage of multiple AAC frames in one RTP packet by applying the Low overhead Audio Transport Multiplex (LATM) framing structure within the RTP payload, as defined in IEC/ISO 14496-3 (2001) [2]. RFC 3016 does also allow for carriage of HE AAC, but the presence of SBR data can only be signalled implicitly, that is a decoder can only detect carriage of HE AAC data by identifying SBR data when parsing the extension payload of AAC frames. Higher level signalling is not possible with RFC 3016.

Next to RFC 3016, also RFC 3640 can be used to transport HE AAC. RFC 3640 provides a "generic" solution for transport of MPEG-4 data. RFC 3640 supports both, implicit signalling (similarly as with RFC 3016) as well as explicit signalling by means of conveying the AudioSpecificConfig() as the required MIME parameter 'config', as defined in RFC 3640. The framing structure defined in RFC 3640 does support carriage of multiple AAC frames in one RTP packet with optional interleaving to improve error resiliency in packet loss. For example, if each RTP packet carries three AAC frames, then with interleaving the RTP packets may carry the AAC frames as given in figure A-2.
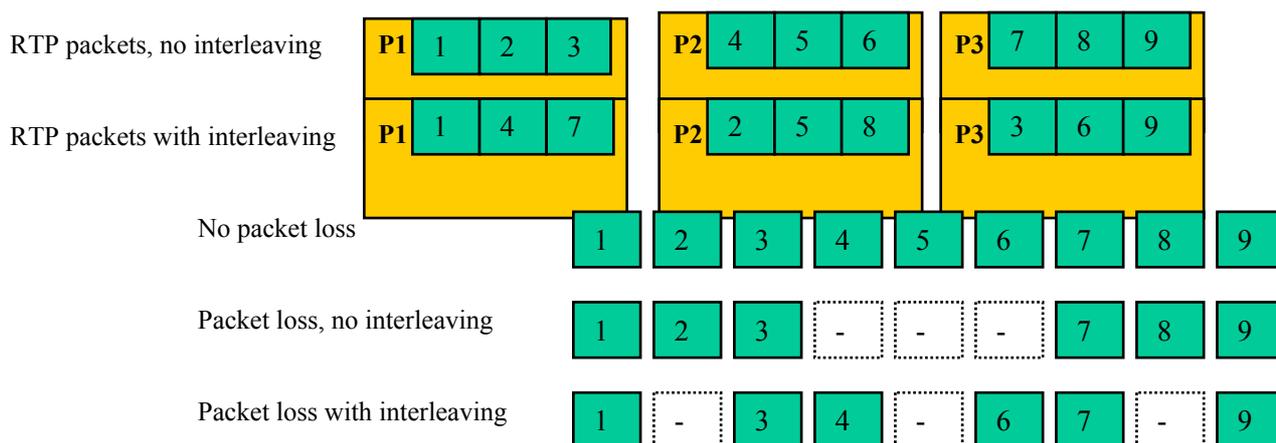
RTP packets, no interleaving   **P1** 1 2 3   **P2** 4 5 6   **P3** 7 8 9

RTP packets with interleaving   **P1** 1 4 7   **P2** 2 5 8   **P3** 3 6 9

No packet loss   1 2 3 4 5 6 7 8 9

Packet loss, no interleaving   1 2 3 - - - 7 8 9

Packet loss with interleaving   1 - 3 4 - 6 7 - 9

Figure A-2: Interleaving of AAC frames

Without interleaving, then RTP packet P1 carries the AAC frames 1, 2 and 3, while packet P2 and P3 carry the frames 4, 5 and 6 and the frames 7, 8 and 9, respectively. When P2 gets lost, then AAC frames 4, 5 and 6 get lost, and hence the decoder needs to reconstruct three missing AAC frames that are contiguous. In this example, interleaving is applied so that P1 carries 1, 4 and 7, P2 carries 2, 5 and 8, and P3 carries 3, 6 and 9. When P2 gets lost in this case, again three frames get lost, but due to the interleaving, the frames that are immediately adjacent to each lost frame are received and can be used by the decoder to reconstruct the lost frames, thereby exploiting the typical temporal redundancy between adjacent frames to improve the perceptual performance of the receiver.

Transport of H.264 / AVC video

To transport H.264/VC video data, RFC yyyy [7] is used. The H.264/AVC specification [1] distinguishes conceptually between a Video Coding Layer (VCL), and a Network Abstraction Layer (NAL). The VCL contains the video features of the codec (transform, quantization, motion compensation, loop filter, etc.). The NAL layer formats the VCL data into Network Abstraction Layer units (NAL units) suitable for transport across the applied network or storage medium. A NAL unit consists of a one-byte header and the payload; the header indicates the type of the NAL unit and other information, such as the (potential) presence of bit errors or syntax violations in the NAL unit payload, and information regarding the relative importance of the NAL unit for the decoding process. RFC yyyy specifies how to carry NAL units in RTP packets.

Synchronization of content delivered over IP

RTP also provides tools for synchronization. For that purpose, an RTP time stamp is present in the RTP header; the RTP time stamps are used to determine the presentation time of the audio and video access units. The method to synchronize content transported in RTP packets is described RFC 3550 [4]. By means of figure A-3 a simplified summary is given below:

a) RTP time stamps convey the sampling instant of access units at the encoder. The RTP time stamp is expressed in units of a clock, which is required to increase monotonically and linearly. The frequency of this clock is specified for each payload format, either explicitly or by default. Often, but not necessarily, this clock is the sampling clock. In figure A-1, TSa(i) and TSv(j) are RTP time stamps that are used to present the access units at the correct timing at the receiver; this requires that the receiver reconstructs the video clock and audio clock with the same mutual offset in time as at the sender.

b) When transporting RTP packets, the RTCP Control Protocol, also defined in RFC 3550 [4], is used for purposes such as monitoring and control. RTCP data is carried in RTCP packets. There are several RTCP packet types, one of which is the Sender Report (SR) RTCP packet type. Each RTCP SR packet contains an RTP time stamp and an NTP time stamp; both time stamps correspond to the same instant in time. However, the RTP time stamp is expressed in the same units as RTP time stamps in data packets, while the NTP time stamp is expressed in "wallclock" time; see section 4 of RFC 3550 [4]. In figure A-3, NTPa(k) and NTPv(n) are the NTP time stamps of the audio and video RTCP packets. At(k) and Vt(n) are the values of the audio and video clock at the same instant in time as NTPa(k) and NTPv(n), respectively. Each SR(k) for audio provides NTPa(k) as NTP time stamp and At(k) as RTP time stamp. Similarly, each SR(n) for video provides NTPv(n) as the NTP time stamps and Vt(n) as RTP time stamp.
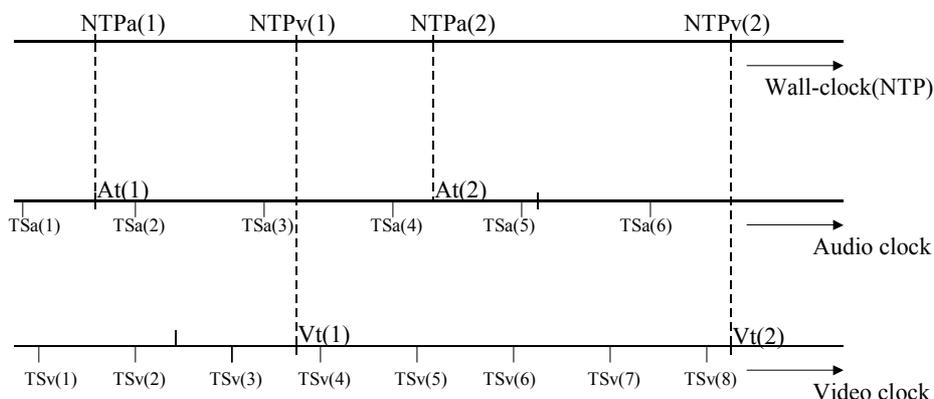
Figure A-1 RTP tools for synchronization

c) Synchronized playback of streams is only possible if the streams use the same wall-clock to encode NTP values in SR packets. If the same wall-clock is used, receivers can achieve synchronization by using the correspondence between RTP and NTP time stamps. To synchronize an audio and a video stream, one needs to receive an RTCP SR packet relating to the audio stream, and an RTCP SR packet relating to the video stream. These SR packets provide a pair of NTP timestamps and their corresponding RTP timestamps that is used to align the media. For example, in figure A-3, [NTPv(k) – NTPa(n)] represents the offset in time between Vt(k) and At(n), expressed in wallclock time.

d) The time between sending subsequent RTCP SR packets may vary; the default RTCP timing rules suggest to send an RTCP SR packet every 5 seconds. This means that upon entering a streaming session there may be an initial delay - on average a 2.5 seconds duration if the default RTCP timing rules are used - when the receiver does not yet have the necessary information to perform inter-stream synchronization.

Synchronization with content delivered over MPEG-2 TS

Applications may require synchronization of audiovisual content delivered over IP with content delivered over an MPEG-2 TS. For example, a broadcaster may wish to provide audio in another language as part of a broadcast program, but using transport over IP instead of transporting this additional audio stream over the same MPEG-2 TS as the broadcast program.

Synchronization of a stream delivered over IP with a broadcast program requires that the receiver knows the timing relationship between the RTP time stamps of the stream that is delivered over IP and the MPEG-2 time stamps of the broadcast program. It is beyond the scope of this document how to convey such timing relationship.

Service discovery

For discovery of DVB services over IP it is referred to the IPI specification for low and mid level (PSI / SI equivalent) functionality and to the GBS specification for higher level (SI / metadata related, except structures and containers) functionality.

Linking to applications

Audio and video delivered over IP can be presented in an MHP application by means of including appropriate URLs.

Capability exchange

By means of capability exchange protocols the sender and receiver can communicate whether the receiver has A, B, C, D or E IP-IRD capabilities for H.264/AVC decoding. In addition, it can also be communicated whether the receiver has multi-channel or only mono/stereo capabilities for High Efficiency AAC decoding. For capability exchange protocols it is referred to the IPI specification.

A.3 Video

Video overview

The H.264/AVC design supports the coding of video (in 4:2:0 chroma format) that contains either progressive or interlaced frames, which may be mixed together in the same sequence. Generally, a frame of video contains two interleaved fields, the top and the bottom field. The two fields of an interlaced frame, which are separated in time by a field period (half the time of a frame period), may be coded separately as two fields or together as a frame. A progressive frame should always be coded as a single frame; however, it is still considered to consist of two fields at the same instant of time. H.264/AVC covers a Video Coding Layer (VCL), which is designed to efficiently represent the video content, and a Network Abstraction Layer (NAL), which formats the VCL representation of the video and provides header information in a manner appropriate for conveyance by a variety of transport layers or storage media The structure of H.264/AVC video encoder is shown in Fig. A-4.
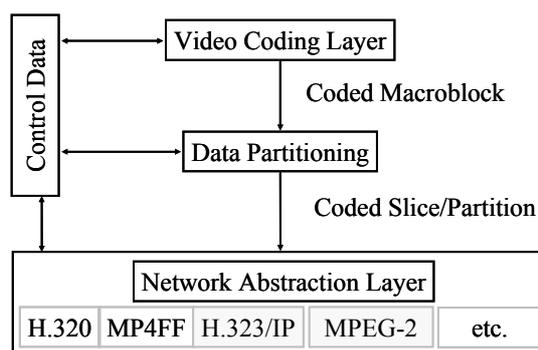

Figure A-4: Structure of H.264/AVC video encoder.

Network Abstraction Layer

The Video Coding Layer (VCL), which is described below, is specified to efficiently represent the content of the video data. The Network Abstraction Layer (NAL) is specified to format that data and provide header information in a manner appropriate for conveyance by the transport layers or storage media. All data are contained in NAL units, each of which contains an integer number of bytes. A NAL unit specifies a generic format for use in both packet-oriented and bitstream systems. The format of NAL units for both packet-oriented transport and bitstream is identical except that each NAL unit can be preceded by a start code prefix in a bitstream-oriented transport layer. The NAL facilitates the ability to map H.264/AVC VCL data to transport layers such as

- RTP/IP for any kind of real-time wire-line and wireless Internet services (conversational and streaming)
- File formats, e.g. ISO MP4 for storage and MMS
- H.32X for wireline and wireless conversational services
- MPEG-2 systems for broadcasting services, etc.

The full degree of customization of the video content to fit the needs of each particular application was outside the scope of the H.264/AVC standardization effort, but the design of the NAL anticipates a variety of such mappings.

One key concept of the NAL is parameter sets. A parameter set is supposed to contain information that is expected to rarely change over time. There are two types of parameter sets:

- sequence parameter sets, which apply to a series of consecutive coded video pictures and

- picture parameter sets, which apply to the decoding of one or more individual pictures

The sequence and picture parameter set mechanism decouples the transmission of infrequently changing information from the transmission of coded representations of the values of the samples in the video pictures. Each VCL NAL unit contains an identifier that refers to the content of the relevant picture parameter set, and each picture parameter set contains an identifier that refers to the content of the relevant sequence parameter set. In this manner, a small amount of data (the identifier) can be used to refer to a larger amount of information (the parameter set) without repeating that information within each VCL NAL unit.

Video Coding Layer

The video coding layer of H.264/AVC is similar in spirit to other standards such as MPEG-2 Video. It consists of a hybrid of temporal and spatial prediction in conjunction with transform coding. Figure A-5 shows a block diagram of the video coding layer for a macroblock, which consists of a 16x16 luma block and two 8x8 chroma blocks.
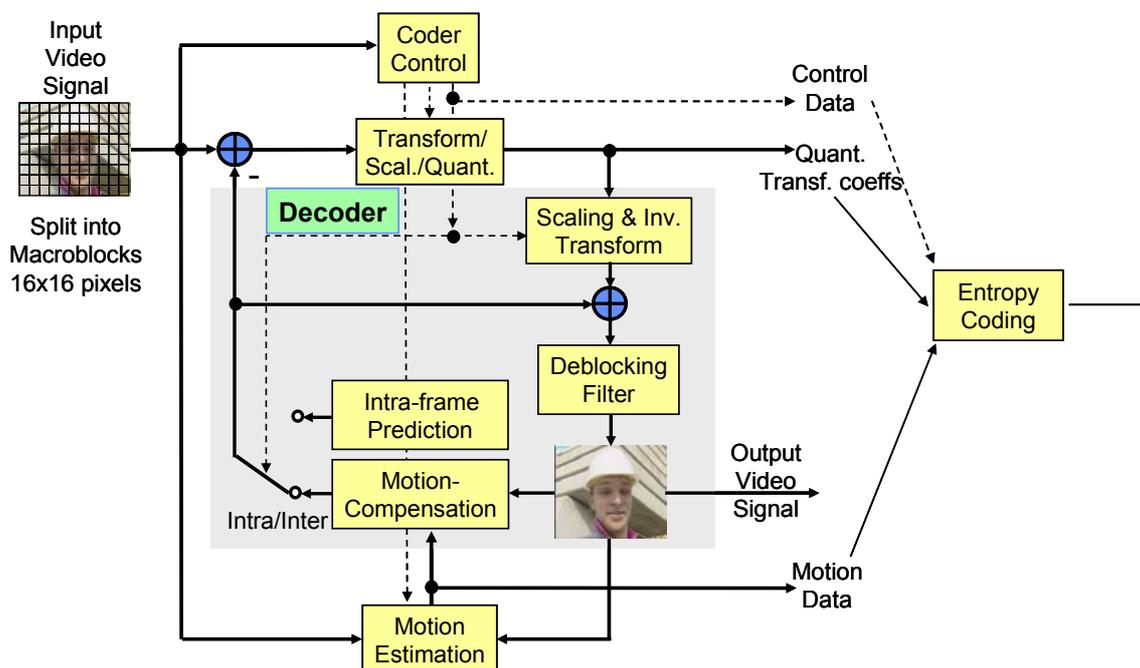

Figure A-5: Basic coding structure for H.264/AVC for a macroblock.

In summary, the picture is split into macroblocks. The first picture of a sequence or a random access point is typically coded in Intra, i.e., without using other information than the information contained in the picture itself. Each sample of a luma or chroma block of a macroblock in such an Intra frame is predicted using spatially neighbouring samples of previously coded blocks. The encoding process is to choose which and how neighbouring samples are used for Intra prediction which is simultaneously conducted at encoder and decoder using the transmitted Intra prediction side information.

For all remaining pictures of a sequence or between random access points, typically Inter coding is utilized. Inter coding employs prediction (motion compensation) from other previously decoded pictures. The encoding process for Inter prediction (motion estimation) consists of choosing motion data comprising the reference picture and a spatial displacement that is applied to all samples of the macroblock. The motion data which are transmitted as side information are used by encoder and decoder to simultaneously provide the inter prediction signal.

The residual of the prediction (either Intra or Inter) which is the difference between the original and the predicted macroblock is transformed. The transform coefficients are scaled and quantized. The quantized transform coefficients are entropy coded and transmitted together with the side information for either Intra-frame or Inter-frame prediction.

The encoder contains the decoder to conduct prediction for the next blocks or next picture. Therefore, the quantized transform coefficients are inverse scaled and inverse transformed in the same way as at the decoder side resulting in the decoded prediction residual. The decoded prediction residual is added to the prediction. The result of that addition is fed into a deblocking filter which provides the decoded video as its output.

The new features of H.264/AVC compared to MPEG-2 Video are listed as follows: variable block-size motion compensation with small block sizes from 16x16 luma samples down to 4x4 luma samples per block, quarter-sample-accurate motion compensation, motion vectors pointing over picture boundaries, multiple reference picture motion compensation, decoupling of referencing order from display order, decoupling of picture representation methods from picture referencing capability, weighted prediction, improved "skipped" and "direct" motion inference, directional spatial prediction for intra coding, in-the-loop deblocking filtering, 4x4 block-size transform, hierarchical block transform, short word-length/exact-match inverse transform, context-adaptive binary arithmetic entropy coding, flexible slice size, flexible macroblock ordering (FMO), arbitrary slice ordering (ASO), redundant pictures, data partitioning, SP/SI synchronization/switching pictures.

Explanation of H.264/AVC Profiles and Levels

Profiles and levels specify conformance points. These conformance points are designed to facilitate interoperability between various applications of the standard that have similar functional requirements. A *profile* defines a set of coding tools or algorithms that can be used in generating a conforming bit-stream, whereas a *level* places constraints on certain key parameters of the bitstream.

All decoders conforming to a specific profile must support all features in that profile. Encoders are not required to make use of any particular set of features supported in a profile but have to provide conforming bitstreams, i.e. bitstreams that can be decoded by conforming decoders. In H.264/AVC, three profiles are defined, which are the *Baseline*, *Main*, and *Extended Profile*.

The *Baseline* profile supports all features in H.264/AVC except the following two feature sets:

- **Set 1**: B slices, weighted prediction, context-adaptive binary arithmetic coding (CABAC), field coding, and picture or macroblock adaptive switching between frame and field coding.
- **Set 2**: SP/SI slices and slice data partitioning.

The first set of additional features is supported by the *Main* profile. However, the *Main* profile does not support the FMO, ASO, and redundant pictures features which are supported by the *Baseline* profile. Thus, only a subset of the coded video sequences that are decodable by a *Baseline* profile decoder can be decoded by a *Main* profile decoder. Flags are used to indicate which profiles of decoder can decode the coded video sequence. These flags are called constrained_set0_flag indicating a bitstream that can be decoded by a *Baseline* profile decoder, constrained_set1_flag indicating a bitstream that can be decoded by a *Main* profile decoder, and constrained_set2_flag indicating a bitstream that can be decoded by an *Extended* profile decoder. Therefore, a stream labelled as conforming to Baseline profile with constrained_set1_flag enabled may contain all *Baseline* features but shall not contain FMO, ASO, and redundant pictures since they are not supported in *Main* profile.

The *Extended* Profile supports all features of the *Baseline* profile, and both sets of features on top of *Baseline* profile, except for CABAC.

In H.264/AVC, the same set of level definitions is used with all profiles, but individual implementations may support a different level for each supported profile. Fifteen levels are defined specifying upper limits for the picture size (in units of macroblocks) ranging from QCIF to all the way to above 4k x 2k, decoder-processing rate (in macroblocks per second) ranging from 250k pixels per sec to 250M pixels per sec, size of the multi-picture buffers, video bit rate ranging from 64 kbps to 240 Mbps, and video buffer size.

Summary of key tools and parameter ranges for Capability A to E IRDs

The following table summarizes the assignment of profiles and levels to the five IP-IRDs that are specified in this report.

Capability	Profile	Additional Constraint on Profile	Level	Max frame size [macroblocks]	Max frame @ max frame size [f/s]	Max bit rate [kbit/s]
A	Baseline	constraint_set1_flag = 1	1[1]	99 (QCIF=176x144)	15	128
B	Baseline	constraint_set1_flag = 1	1.2	396 (CIF=352x288)	15.2	384
C	Baseline	constraint_set1_flag = 1	2	396 (CIF=352x288)	30	2000
D	Main	none	3	1,620 (625 SD =720x576)	25	10,000
E	Main	none	4	8,192 (2x1KHD=2048x1024)	15	20,000

The following should be noted.

IP-IRDs with Capability A, B, and C specify the Baseline profile with the additional constraint that constraint_set1_flag shall be set equal to 1 making these bitstreams also decodable by Main profile decoders. The reason for this additional constraint is that our investigations have shown that the features that are contained in Baseline but are not contained in Main profile (FMO, ASO, and redundant pictures) and are disabled by setting constraint_set1_flag equal to 1 do not provide any benefit at the packet error rates envisioned to be typical for the applications in which this report will be used. IP-IRDs with capability D and E shall be conforming to Main profile without any additional constraints. Because of the additional constraint and the requirements in H.264/AVC, IP-IRDs labelled with a particular capability Y are capable of decoding and displaying pictures that can be decoded by IP-IRDs labelled with a particular capability X with X being an earlier letter than Y in the alphabet. For instance, Capability D IP-IRDs are capable of decoding bitstreams conforming to Main Profile at level 3 of H.264/AVC and below. Additionally, Capability D IP-IRDs are capable of decoding bitstreams that are also decodable by IP-IRDs with capabilities A, B, or C.

For the Capability A IP-IRD, the values assigned to the MaxBR variable which corresponds the maximum bit rate and the value assigned to the MaxCPB variable which corresponds to the maximum bitstream buffer size are modified in that they have been doubled to accommodate larger bit rates and buffer sizes for better video quality.

Each level specifies a maximum number of macroblocks per second that can be processed by a corresponding decoder (not explicitly listed in the table). Additionally, the maximum number of macroblocks per frame is restricted as well. For example, for the Capability D IP-IRD, the maximum number of macroblocks per frame is given as 1,620 corresponding to a 625SD picture (level 3 of H.264/AVC). Together with the maximum number of macroblocks per second that can be processed which are given as 40,500, the maximum frame rate is given as 25 frames per second. Please note that this also permits the processing of 525SD pictures at 30 frames per second.

Other Video Parameters

This report is supposed to cover a large variety of applications. Therefore, we do not specify parameters such as frame rate, aspect ratio, chromaticity, chroma, and random access points as restrictively as they are specified in TR 101 154.

For parameters such as frame rate and aspect ratio, the constraints as specified in H.264/AVC are sufficient and need no further adjustment. It is only recommended to avoid extreme values.

For parameters such as chromaticity and chroma, it is recommended to utilize the parameters that are specified in the VUI of H.264/AVC which is part of the sequence parameter set.

Random access points are provided through so-called instantaneous decoding refresh (IDR) pictures. In our recommendations, we distinguish broadcast and other applications. For broadcast applications it is recommended that sequence and picture parameter sets are sent together with a random access point (e.g. an IDR picture) to be encoded at

---

[1] with MaxBR (maximum bit rate) value equal to 128 and with MaxCPB (maximum bitstream buffer size) value equal to 350

least once every 500 milliseconds. For multicast or streaming applications a maximum interval of 5 seconds between random access points should not be exceeded.

A.4 Audio

MPEG-4 High Efficiency AAC (HE AAC)

The principle problem of traditional perceptual audio codecs at low bit rates is, that they would need more bits to encode the whole spectrum accurately than available. The results are either coding artefacts or the transmission of a reduced bandwidth audio signal. To resolve this problem, MPEG decided to add a bandwidth extension technology as a new tool to the MPEG-4 audio toolbox. With SBR the higher frequency components of the audio signal are reconstructed at the decoder based on transposition and additional helper information. This method allows an accurate reproduction of the higher frequency components with a much higher coding efficiency compared to a traditional perceptual audio codec. Within MPEG the resulting audio codec is called MPEG-4 High Efficiency AAC (HE AAC) and is the combination of the MPEG-4 Audio Object Types AAC-Low Complexity (LC) and Spectral Band Replication (SBR). It is not a replacement for AAC, but rather a superset which extends the reach of high-quality MPEG-4 Audio to much lower bitrates. High Efficiency AAC decoders will decode both, plain AAC and the enhanced AAC plus SBR. The result is a backward compatible extension of the standard.

The basic idea behind SBR is the observation that usually a strong correlation between the characteristics of the high frequency range of a signal (further referred to as "highband") and the characteristics of the low frequency range (further referred to as "lowband") of the same signal is present. Thus, a good approximation of the representation of the original input signal highband can be achieved by a transposition from the lowband to the highband. In addition to the transposition, the reconstruction of the highband incorporates shaping of the spectral envelope. This process is controlled by transmission of the highband spectral envelope of the original input signal. Additional guidance information for the transposing process is sent from the encoder, which controls means, such as inverse filtering, noise and sine addition. This transmitted side information is further referred to as SBR data.

Figure A-6 shows a block diagram of a HE AAC Encoder. The AAC encoder is operated at half the input sampling frequency of the input audio signal, while the SBR encoder operates on the full sampling frequency. SBR data is embedded into the AAC bitstream by means of the extension_payload() element Two types of SBR extension data can be signalled through the extension_type field of the extension_payload(). For compatibility reasons with existing AAC only decoders, two different methods for signalling the existence of an SBR payload can be selected. Both methods are described below.
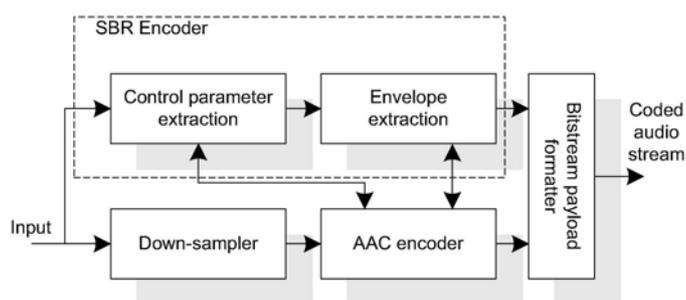

Figure A-6: HE AAC Encoder

The HE AAC decoder is depicted in Figure A-7. The coded audio stream is fed into a demultiplexing unit prior to the AAC decoder and the SBR decoder. The AAC decoder reproduces the lower frequency part of the audio spectrum. The time domain output signal from the underlying AAC decoder at the sampling rate fs_{AAC} is first fed into a 32 channel Quadrature Mirror Filter (QMF) analysis filterbank. Secondly, the high frequency generator module recreates the highband by patching QMF subbands from the existing low band to the high band. Furthermore, inverse filtering is applied on a per QMF subband basis, based on the control data obtained from the bitstream. The envelope adjuster modifies the spectral envelope of the regenerated highband, and adds additional components such as noise and sinusoids, all according to the control data in the bitstream. Finally a 64 channel QMF synthesis filterbank is applied to retain a time-domain output signal at twice the sampling rate, i.e. $fs_{out} = fs_{SBR} = 2 \cdot fs_{AAC}$.
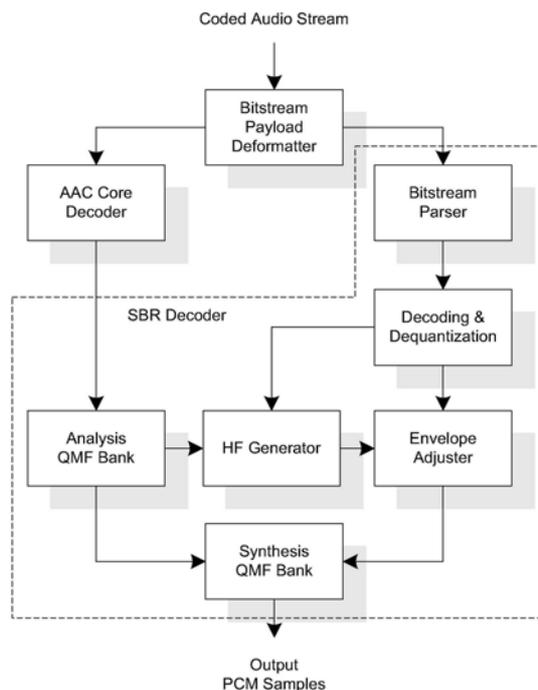
Figure A-7: HE AAC Decoder

HE AAC Levels and Main Parameters for DVB

MPEG-4 provides a huge toolset for the coding of audio objects. In order to allow effective implementations of the standard, subsets of this toolset have been identified that can be used for specific applications. The function of these subsets, called ``Profiles,'' is to limit the toolset a conforming decoder must implement. For each of these Profiles, one or more Levels have been specified, thus restricting the computational complexity.

The High Efficiency AAC Profile is introduced as a superset of the AAC Profile. Besides the Audio Object Type (AOT) AAC LC (which is present in the AAC Profile), it includes the AOT SBR. Levels are introduced within these Profiles in such a way, that a decoder supporting the High Efficiency AAC Profile at a given level can decode an AAC Profile stream at the same or lower level.

Table: Levels within the HE AAC Profile				
Level	Max. channels /object	Max. AAC sampling rate, SBR not present [kHz]	Max. AAC sampling rate, SBR present [kHz]	Max. SBR sampling rate, [kHz] (in/out)
1	NA	NA	NA	NA
2	2	48	24	24/48
3	2	48	48 [Note1]	48/48
4	5	48 [Note2]	24/48 [Note1]	48/48
5	5	96	48	48/96

Note 1: For level 3 and level 4 decoders, it is mandatory to operate SBR in a downsampled mode if the sampling rate of the AAC core is higher than 24kHz. Hence, if SBR operates on a 48kHz AAC signal, the internal sampling rate of SBR will be 96kHz, however, the output signal will be downsampled by SBR to 48kHz.

Note 2: For one or two channels the maximum AAC sampling rate, with SBR present, is 48kHz. For more than two channels the maximum AAC sampling rate, with SBR present, is 24kHz.

For DVB the level 2 for mono and stereo as well as the level 4 multichannel audio signals are supported. The Low Frequency Enhancement channel of a 5.1 audio signal is included in the level 4 definition of the number of channels.

Methods for signalling of SBR

Several ways how to signal the presence of SBR data are possible:

1. **implicit signaling:** if SBR extension elements (EXT_SBR_DATA or EXT_SBR_DATA_CRC) are detected in the bitstream, this implicitly means that SBR data is present. This mode provides backward compatibility with AAC-only decoders since a non-SBR aware AAC only decoder would simply skip the SBR data. On the other hand this signalling method may introduce challenges when operating the decoder in a complex system such as an embedded device, since in order to determine the output sampling rate the decoder would need to parse the payload at least partially in order to detect SBR (as explained above the output sampling rate, resp. the sampling rate of SBR is twice the sampling rate of AAC, i.e. the sampling rate associated with the AAC LC AOT) .

2. **explicit signaling:** the presence of SBR data is signalled by means of the AOT SBR in the AudioSpecificConfig(). This permits to convey configuration data specific to the SBR decoder, which includes separate specifications of the sampling rates for the SBR and AAC decoders. These specifications are also used to implicitly signal the down sampling mode. If the sampling rates for the SBR and AAC decoders are identical, the down-sampled SBR tool is used. Two types of explicit signalling are available:

    - **hierarchical signalling:** if the first AOT is signalled as SBR, a second AOT is signalled which indicates the underlying AOT, e.g. AAC LC. This is a non backward compatible signalling method.

    - **backward compatible signaling:** the extensionAOT is signalled at the end of the AudioSpecificConfig(). This signalling method can only be used in systems that convey the length of the AudioSpecificConfig(). Because of this restriction, backward compatible explicit signalling can for example not be used with LATM configurations.

Since backward compatible signalling of SBR is usually not required in the context of DVB services over IP, it is recommended to use explicit hierarchical signalling of SBR.

Which signalling options are available depends on the applied tool for transport of HE AAC audio:
    - With RFC 3016 only implicit signalling is possible
    - With RFC 3640 both explicit and implicit signalling are possible.

A.5 Future Work

In common with TR 101 154 and TR 102 154, these guidelines are a living document, subject to periodic revision. The intention is to develop revisions in a largely backwards compatible manner, so that no changes to the mandatory functionality of a previously defined IRD are made between one edition and the next.

One specific issue that is currently under consideration is the possibility of extending the guidelines to include even higher resolution content, such as 1080p 60Hz. If this is done, it is likely that Level 4.2 of H.264/AVC would be chosen.

A084

Bibliography

The following material, though not specifically referenced in the body of the present document (or not publicly available), gives supporting information.

To be determined

History

Document history		
V1.0.03	January 2003	Submitted for information to January Technical Module
V1.0.04	March 2003	Submitted for information to March Technical Module
V1.0.05	June 2003	Submitted for information to June Technical Module
V1.0.06	September 2003	Normative portions approved by September Technical Module
V1.0.07	January 2004	Informative Annex A added and submitted to Technical Module for approval
A084	July 2004	Processing to yield A084 (equivalent to draft TS 102 005 V1.1.1)