SCTE · ISBE

STANDARDS

Digital Video Subcommittee

AMERICAN NATIONAL STANDARD

ANSI/SCTE 242-4 2018

Next Generation Audio Coding Constraints for Cable Systems: Part 4 – DTS-UHD Audio Coding Constraints

NOTICE

The Society of Cable Telecommunications Engineers (SCTE) / International Society of Broadband Experts (ISBE) Standards and Operational Practices (hereafter called "documents") are intended to serve the public interest by providing specifications, test methods and procedures that promote uniformity of product, interchangeability, best practices and ultimately the long-term reliability of broadband communications facilities. These documents shall not in any way preclude any member or non-member of SCTE•ISBE from manufacturing or selling products not conforming to such documents, nor shall the existence of such standards preclude their voluntary use by those other than SCTE•ISBE members.

SCTE•ISBE assumes no obligations or liability whatsoever to any party who may adopt the documents. Such adopting party assumes all risks associated with adoption of these documents, and accepts full responsibility for any damage and/or claims arising from the adoption of such documents.

Attention is called to the possibility that implementation of this document may require the use of subject matter covered by patent rights. By publication of this document, no position is taken with respect to the existence or validity of any patent rights in connection therewith. SCTE•ISBE shall not be responsible for identifying patents for which a license may be required or for conducting inquiries into the legal validity or scope of those patents that are brought to its attention.

Patent holders who believe that they hold patents which are essential to the implementation of this document have been requested to provide information about those patents and any related licensing terms and conditions. Any such declarations made before or after publication of this document are available on the SCTE•ISBE web site at http://www.scte.org.

All Rights Reserved

© Society of Cable Telecommunications Engineers, Inc. 2018
140 Philips Road
Exton, PA 19341

Table of Contents

Title **Page Number**

List of Figures

List of Tables

1. Introduction

1.1. Scope

This document is part four of a multi-part standard that specifies the coding constraints of Next Generation Audio system for cable television. In conjunction with SCTE 242-1 [1], this document defines the coding constraints on DTS-UHD for cable television. The carriage of the streams described in this specification is defined in SCTE 243-4 [8] in conjunction with SCTE 243-1[7].

2. Normative References

The following documents contain provisions, which, through reference in this text, constitute provisions of this document. At the time of Subcommittee approval, the editions indicated were valid. All documents are subject to revision; and while parties to any agreement based on this document are encouraged to investigate the possibility of applying the most recent editions of the documents listed below, they are reminded that newer editions of those documents might not be compatible with the referenced version.

2.1. SCTE References

[1] ANSI/SCTE 242-1 2017, Next Generation Audio Coding Constraints for Cable Systems: Part 1 – Introduction and Common Constraints

2.2. Standards from Other Organizations

[2] ATSC Standard A/342-1:2017, A/342 Part 1, Audio Common Elements
[3] ETSI TS 103 491 V1.1.1 (2017-04), DTS-UHD Audio Format; Delivery of Channels, Objects and Ambisonic Sound Fields
[4] Recommendation ITU-R BS.1770-4 (2015-10), Algorithms to measure audio programme loudness and true-peak audio level
[5] DVB BlueBook A038 (2017-12), Digital Video Broadcasting (DVB); Specification for Service Information (SI) in DVB systems (Final draft of ETSI EN 300 468 v 1.16.1)
[6] ISO/IEC 639-2:1998, "Codes for the representation of names of languages - Part 2: Alpha-3 code"

2.3. Published Materials

- No normative references are applicable.

3. Informative References

The following documents might provide valuable information to the reader but are not required when complying with this document.

3.1. SCTE References

[7] ANSI/SCTE 243-1 2017, Next Generation Audio Carriage Constraints for Cable Systems: Part 1 – Common Transport Signaling
[8] SCTE 243-4 2018, Next Generation Audio Carriage Constraints for Cable Systems: Part 4 – DTS-UHD Audio Carriage Constraints

3.2. Standards from Other Organizations

[9] ETSI TS 103 584 V1.1.1 (2018-01), DTS-UHD Point Source Renderer
[10] ATSC Doc A/85:2013, Techniques for Establishing and Maintaining Audio Loudness for Digital Television

3.3. Published Materials

- No informative references are applicable.

4. Compliance Notation

shall	This word or the adjective "***required***" means that the item is an absolute requirement of this document.
shall not	This phrase means that the item is an absolute prohibition of this document.
forbidden	This word means the value specified shall never be used.
should	This word or the adjective "*recommended*" means that there may exist valid reasons in particular circumstances to ignore this item, but the full implications should be understood and the case carefully weighted before choosing a different course.
should not	This phrase means that there may exist valid reasons in particular circumstances when the listed behavior is acceptable or even useful, but the full implications should be understood and the case carefully weighed before implementing any behavior described with this label.
may	This word or the adjective "*optional*" means that this item is truly optional. One vendor may choose to include the item because a particular marketplace requires it or because it enhances the product, for example; another vendor may omit the same item.
deprecated	Use is permissible for legacy purposes only. Deprecated features may be removed from future versions of this document. Implementations should avoid use of deprecated features.

5. Abbreviations and Definitions

5.1. Abbreviations

LFE	Low Frequency Effects.
ISBE	International Society of Broadband Experts
SCTE	Society of Cable Telecommunications Engineers

5.2. Definitions

This specification uses the definitions defined in ANSI/SCTE 242-1[1] and, by incorporation, ATSC A/342-1[2]. The following terms have definitions specific to DTS-UHD and shall apply to all clauses in this document.

Audio Chunk	block of data within an audio frame containing compressed audio samples
Audio Frame	unit of coded audio that, when decoded, will generate defined number

	of uncompressed Linear PCM audio samples for each wave form
BroadcastChunk	block of data within an audio stream containing data that maps audio components to preselections
Frame Duration	time represented by one decoded Audio Frame
Metadata Chunk	block of data within an audio frame containing metadata describing an audio presentation
Object	Audio Element as defined in ATSC A/342-1[2] and referenced in ANSI/SCTE 242-1[1]
Object Group	selected collection of audio objects to be played together
Presentation	selected collection of Channels, or Objects and Object Groups used together to generate the rendered output

6. DTS-UHD System Description

The DTS-UHD coding system is the third generation of DTS audio delivery formats. It is designed to both improve efficiency and deliver a richer set of features than the second generation DTS system. The first two generations of DTS codecs were designed primarily for Channel Based Audio (CBA). DTS-UHD is primarily designed to support audio objects, where a given object can represent a channel based presentation, sound field channels, or audio objects used in Object Based Audio (OBA).

6.1. Terminology

Table 1 lists terms defined in ATSC A/342-1[2] and maps them to corresponding terms defined in ETSI TS 103 491[3]

Table 1 - Common Terms Cross Reference

Common Term	DTS-UHD (TS 103 491) terms
Audio Element	Object
Audio Element Metadata	Metadata Chunk
Audio Presentation	Presentation
Audio Program	Audio Program
Audio Program Component	Object, Object Group or Presentation*
Elementary Stream	Elementary Stream
* In a DTS-UHD stream, an Audio Presentation points to a list of Components and contains additional metadata for rendering. The Components may be Objects, Object Groups or other Presentations. When multiple elementary streams are used to create an Audio Program, each DTS-UHD Presentation in those streams is an Audio Program Component.	

6.2. Overview

The DTS-UHD bitstream supports 32 pre-defined channel locations and definition of up to 224 objects plus 32 object groups. A DTS-UHD object is a set of coded waveforms plus an associated metadata structure, which are referred to in ETSI TS 103 491 [3] as an audio chunk element, and a metadata chunk

element. It is worth noting that more than one object can reference the same audio chunk. These chunks are carried within an Audio Frame. In addition to metadata required to decode the audio chunk elements, metadata chunk elements may carry metadata that will be passed downstream to a renderer, e.g. as described in TS 103 584 [9].

The audio frame is organized by a frame table of contents (FTOC) at the beginning of the frame. The FTOC locates the metadata chunks and audio chunks within the frame, and creates associations of objects, object groups, and presentations. An object group is a collection of objects always played together and assigned to a single object ID. A presentation is a list of object IDs plus some additional metadata, usually including loudness and dynamics parameters. Up to 32 presentations can be defined in a stream. A more detailed introduction to the DTS-UHD stream architecture is provided in clause 4 of ETSI TS 103 491 [3].

The DTS-UHD decoder processes the selected audio chunks into sets of linear PCM waveforms. The waveforms are then passed to a renderer along with their object metadata, plus loudness and dynamic range metadata for the entire presentation. A reference renderer for DTS-UHD can be found in ETSI TS 103 584 [9].

A presentation is referenced by a single parameter when invoking the decoding process. This parameter is referred to in ETSI TS 103 491 [2] as *ucAudPresIndex*. Note that a given presentation may internally reference other presentations defined within the same stream. All objects and object groups will be combined into presentations for the purposes of linear broadcast applications, so the preselection interfaces in the consumer premises equipment will not reference audio objects or object groups directly. Referring to Table 4-1 in ETSI TS 103 491 [3], only the first two of the three API calls to the bitstream decoder are utilized by SCTE broadcast implementations:

1. In the case of default playback, *ucAucPresInterfaceType* is set to API_PRES_SELECT_DEFAULT_AP, so the default audio program indicated in the stream is played back.
2. In the case of defined playback, *ucAucPresInterfaceType* is set to API_PRES_SELECT_SPECIFIC_AP, where the parameter passed in the API is *ucAudPresIndex*.

A special case of default playback is when *bFullChannelBasedMixFlag* is set to 1. In this case, the stream is organized as a single "object" containing all channels with locations described by a single channel mask. This stream is processed using API_PRES_SELECT_DEFAULT_AP.

6.3. Sync frames and non-sync frames

A given DTS-UHD audio frame is either sync-frame or non-sync frame. Properties of sync frames are specified in ETSI TS 103 491 [3]. This clause provides an overview of some of the important implications of these frame types.

A sync frame contains all parameters necessary to unpack metadata and audio chunks, describe audio chunks, render and process audio samples and generate a frame of linear PCM samples. A decoder can attempt to establish initial synchronization only in a sync frame.

A non-sync frame may only contain parameters that have changed in value since the previous frame or sync frame to minimize payload size.

The time period between sync frames is a sync interval. The sync interval can be any duration, but it is recommended to be at least 500 ms, and is nominally about 2 seconds.

7. Multi-stream playback

When an audio program contains multiple DTS-UHD streams, there shall always be one "main" stream, and a maximum of seven "auxiliary" streams. The main stream contains a required default Audio Preselection and may contain additional Preselections and Components. The auxiliary streams contain additional Components to create new Preselections. Every Preselection shall contain all audio and metadata assets needed to render the final output to the speakers.

The main stream and auxiliary streams shall use the same values for the following parameters:

- Sampling rate (*m_unClockRateInHz* as defined in ETSI TS 103 491 [3])
- Frame duration (*m_unFrameDuration* as defined in ETSI TS 103 491 [3])

The conceptual model of multi-stream playback is that of multiple decoder sessions running in parallel. An implementation choice may be a single decoder processing the frames from the various streams sequentially, then rendering all waveforms from the given time interval together to generate the final output to the speakers.

When an Audio Program contains multiple elementary streams, the indexing of the streams contributing to the Audio Preselection determine which metadata will be used to render the final output. The final rendering metadata for scaling the output is always provided by the highest indexed stream in the sequence that contains such metadata. An example is shown in Figure 1. Here we see three streams contributing to a preselection.

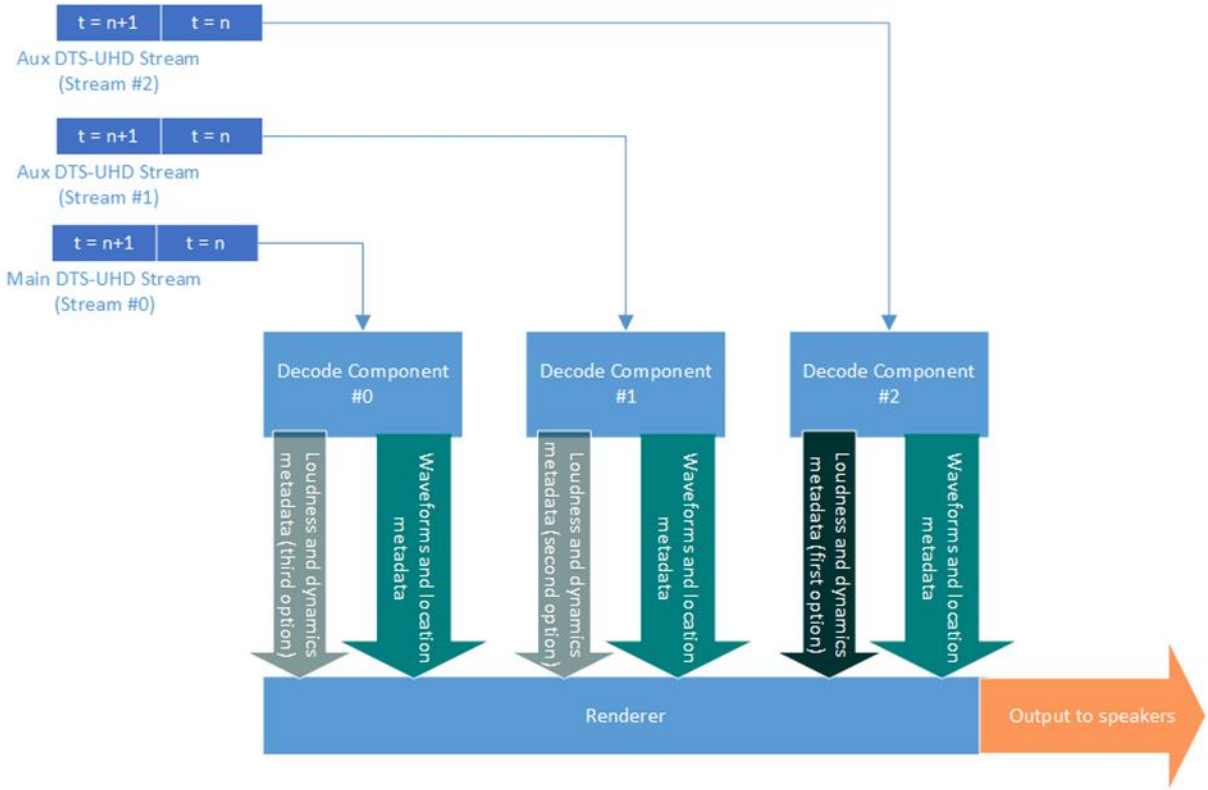

Figure 1 - DTS-UHD Multi-stream Example

The individual components making up the preselection carry positional information for each waveform. For each component the decoder also passes the available loudness and dynamic metadata to the renderer. The metadata from the component in the highest indexed stream from which it is available shall be used for scaling the final output.

For example:

Component #2 is from the highest indexed stream in a multi-stream preselection. The renderer first looks for metadata from Component #2 to perform the final scaling of the mix. If some metadata is not included in Component #2, then the renderer looks at the metadata delivered with Component #1, and finally Component #0, in order, to fill in the missing metadata.

To illustrate this example, consider that the component from elementary stream #0 carries music and effects, the component from elementary stream #1 carries dialogue and the component from elementary stream #2 is adding spoken subtitles. Multiple dialogue objects might be able to use the same music and effects, so the mixing metadata with the dialogue will be preferred when only these two components are played. The spoken subtitle stored in stream #2 was mastered with the M&E from stream #0 and the dialog from stream #1, so it was the only Component mastered with the awareness of the other Components.

8. DTS-UHD Preselections

8.1. Overview

A maximum of one Audio Program Component is contributed from each elementary stream to create a Preselection. Note that a Component, in this context, may contain contributions from other Components within the same elementary stream. Therefore, in the case of a single contributing stream, an Audio Program Component is also a Preselection.

The additional information required to coordinate Components from the various elementary streams is carried in the DTS-UHD BroadcastChunk, defined in clause 8.2. When multiple Preselections are carried in a single stream, the BroadcastChunk is still useful in exposing certain properties of each preselection.

8.2. DTS-UHD BroadcastChunk

The DTS-UHD BroadcastChunk is a map of the Audio Program. It carries high-level metadata about the available Preselections, and a mapping of the Components to specific Preselections. The BroadcastChunk shall be carried as a separate frame, distinct from Audio Frames and unlike Audio Frames there is no duration associated with a frame carrying a BroadcastChunk as it contains no audio samples.

The BroadcastChunk applies to the Sync Interval that follows. It may or may not be valid for adjacent Audio Frames in the stream..

The structure of the BroadcastChunk shall be as defined in Table 2. This metadata block shall be delivered in the main DTS-UHD elementary stream. If auxiliary streams are also present, these streams shall not carry a BroadcastChunk. The components delivered in auxiliary streams shall be referenced by the BroadcastChunk in the main stream.

SelectionSets define preselections and are composed of two parts. The first part is a preamble that identifies properties and roles of the Preselection. The second part is a list of Components needed to compose the Preselection. The preamble has several pre-defined flags to indicate specific features, and an optional byte to further differentiate program content

The BroadcastChunk is organized by language. For each language indicated there shall be one or more SelectionSets, each defining a preselection. Each component shall be identified with a *StreamID* and a *ComponentID*.

Table 2 - BroadcastChunk

Syntax	Number of bits	Identifier
DTSUHD_BCHUNK	32	bslbf
ByteCount	8	uimsbf
Version	3	uimsbf
numLanguages	5	uimsbf
for (i=0; i ≤ numLanguages; i++) { // Language index = i		
ISO639_code // Language Table	24*numLanguages	bslbf
}		
for (i=0; i ≤ numLanguages; i++) { // Language groups		
b_UserByte // Language group header	1	bslbf
reserved_bits	2	blsbl
numSelectionSets [i] // preselections per group	5	uimsbf
for (j = 0; j ≤ numSelectionSets[i]; j++) { // ProgramIndex = j		
AudioDescription // properties of preselection	1	bslbf
SpokenSubtitle	1	bslbf
DialogueEnhancement	1	bslbf
if (b_UserByte)		
UserByte	8	bslbf
numComponents	3	uimsbf
reserved_bits	2	bslbf
for (k = 0; k ≤ numComponents; k++) { // preselection		
StreamID	3	uimsbf
ComponentID	5	uimsbf
}		
}		
}		
CRC16	16	bslbf

8.3. DTS-UHD BroadcastChunk Parameters

8.3.1. DTSUHD_BCHUNK

This is a 4-byte syncword identifying the broadcast chunk and shall have the value shown in Table 3.

Table 3 - BroadcastChunk Syncword

Name	Syncword	Description
DTSUHD_BCHUNK	0x2A3E2523	DTS-UHD BroadcastChunk

8.3.2. ByteCount

ByteCount is the size in bytes of the DTS-UHD BroadcastChunk. This value shall reflect the number of bytes in the BroadcastChunk excluding the syncword, but inclusive of *ByteCount* and CRC16.

8.3.3. Version

Version shall be set to 0.

8.3.4. numLanguages

A 5-bit unsigned integer representing the number of language codes in the language table. *numLanguages* + 1 shall equal the number of language codes in the language table and the number of language groups defined in the BroadcastChunk.

8.3.5. ISO639_code

ISO639_code shall indicate the language of the preselections in the language group as a 3 byte language code according to ISO 639-2 [6].

8.3.6. b_UserByte

The flag *b_UserByte* shall be set to 1 to indicate that a user defined byte is included in each preselection preamble for the associated language group. If user defined bytes are not present, *b_UserByte* shall be 0.

8.3.7. numSelectionSets[i]

A 5-bit unsigned integer representing the number of preselections defined for a particular language group. *numSelectionSets*[i] + 1 shall equal the number of Preselections listed in language group i.

8.3.8. AudioDescription

AudioDescription shall be equal to 1 if the given preselection includes Video Description Service. Otherwise, *AudioDescription* shall equal 0.

8.3.9. SpokenSubtitle

SpokenSubtitle shall be equal to 1 if the preselection includes spoken subtitles as defined in EN 300 468. Otherwise, *SpokenSubtitle* shall equal 0.

8.3.10. DialogueEnhancement

DialogueEnhancement shall be equal to 1 if this audio playback option has been processed to make the dialog more easily understood. Otherwise, *DialogueEnhancement* shall be set to 0.

8.3.11. UserByte

If *b_UserByte* is set to 1, then *UserByte* shall be present in each program preamble for the associated language group. If *b_UserByte* = 0, then *UserByte* shall not be present.

8.3.12. NumComponents

A 3-bit unsigned integer representing the number of Components creating the given preselection. *NumComponents* + 1 shall equal the total number of Components listed by *StreamID* and *ComponentID* to define the given preselection.

8.3.13. reserved_bits

These 2 bits are reserved for future definition. They shall both be set to 0.

8.3.14. StreamID

StreamID shall identify which elementary stream is contributing a given component. *StreamID* = 0 is assigned to the main stream, which is the first stream in the multiplex. *StreamID* = 1, represents the first auxiliary stream, and so on, to a maximum of 8 elementary streams.

8.3.15. ComponentID

ComponentID shall identify a specific contribution within the stream indicated by *StreamID*. This is the corresponding value of *ucAudPresIndex*, as defined in TS 103 491 [3].

8.3.16. CRC16

A BroadcastChunk shall be terminated with a 16-bit CRC starting from (including) numLanguages, through the last instance of *ComponentID*. The CRC shall be calculated by initializing to 0xFFFF and computed using the polynomial $x^{16}+x^{12}+x^5+1$.

8.4. DTS-UHD BroadcastChunk Requirements

The following requirements and constraints apply to the DTS-UHD_BroadcastChunk:

- The DTS-UHD Broadcast Chunk shall be present when multiple DTS-UHD streams are used to construct Preselections. The BroadcastChunk should be present if multiple Preselections are available in the main audio stream.
- When the BroadcastChunk is present, it shall be transmitted in the main DTS-UHD stream and may be located between any two audio frames.
- The BroadcastChunk shall not be encrypted.
- When the BroadcastChunk is present, it shall be present at least once per sync interval.
- If the BroadcastChunk is present more than once per sync interval, all instances of the BroadcastChunk in that sync interval shall be identical.
- The BroadcastChunk shall always apply to the next sync interval.

9. DTS-UHD Coding Specifications

9.1. General Requirements

The following encoding requirements apply to DTS-UHD elementary streams used in SCTE broadcast systems:

- DTS-UHD audio elementary streams shall comply with the syntax and semantics contained in ETSI TS 103 491 [3], and the present document.

- The main DTS-UHD stream shall contain a default preselection. The lowest numbered *ucAudPresIndex* that represents a complete preselection will be played by the decoder in the absence of any explicit instruction, as defined in ETSI TS 103 491 [3].
- For all DTS-UHD elementary streams, a sync frame shall be present at every random-access point.
- If DTS-UHD auxiliary streams are present in the Audio Program, the sync frames of the auxiliary streams shall be aligned to the sync frames of the main stream.
- The Audio Chunk type and audio frame duration shall not change for the duration of the audio program where seamless playback is expected.

A given preselection shall be limited to a total of 16 full-bandwidth waveforms and 2 LFE waveforms for delivery to the renderer.

9.2. Loudness and Dynamics Settings

9.2.1. Loudness

The DTS-UHD bitstream is capable of carrying multiple loudness parameters, some of which include (nominally) the complete preselection, the speech objects only, and composition of all components excluding the speech objects. The content encoder may use these to signal loudness values and the method(s) used to calculate the values.

Multiple measurement methods are supported, and the bitstream syntax has provisions to carry as many as 16 sets of loudness parameters..

For cable television applications the DTS-UHD stream shall contain at least one set of loudness parameters measured according to local loudness regulations (e.g., ATSC A/85[10]).

Each set of loudness parameters consists of 3 values:

m_rLoudness is the measured loudness, stored with an associated asset type code

m_ucAssociatedAssetType which describes the type of asset being measured

m_ucLoudnessMsrmType which indicates the method used to calculate m_rLoudness.

Details regarding loudness measurements and signaling for DTS-UHD are documented in ETSI TS 103 491 [3], clause 7.7.6.5.

Typically, *m_ucLoudnessMsrmType* is used in its shorted form (2-bits).

If *m_ucLoudnessMsrmType* = 0x2 then *m_rLoudness* was calculated using methods described in ATSC A/85.

If *m_ucLoudnessMsrmType* = 0x3, then *m_rLoudness* was calculated using methods described in EBU R128. If the signaling of other measurement methods are required, refer to ETSI TS 103 491 [3], clauses 7.7.5 and 7.7.6 for more detail.

9.2.2. Dynamic Range Personalization

Multiple selectable and custom dynamic range compression curves can optionally be associated with DTS-UHD Programs and signaled in the bitstream to facilitate adaptation to various listening

environments by the renderer. Different curves can be assigned to high, medium and low compression use cases.

The presence of a custom DRC curve shall be indicated by the bitstream metadata parameter *m_bCustomDRCCurveMDPresent* as defined in TS 103 491 [3]. If custom DRC curves are present the DRC parameters shall be encoded as defined in TS 103 491 [3] (for more information refer to clause 7.7.6.8).