ATSC Candidate Standard: A/342 Part 1, Audio Common Elements

Doc. S34-187r9
15 June 2016

The Advanced Television Systems Committee, Inc., is an international, non-profit organization developing voluntary standards for digital television. The ATSC member organizations represent the broadcast, broadcast equipment, motion picture, consumer electronics, computer, cable, satellite, and semiconductor industries.

Specifically, ATSC is working to coordinate television standards among different communications media focusing on digital television, interactive systems, and broadband multimedia communications. ATSC is also developing digital television implementation strategies and presenting educational seminars on the ATSC standards.

ATSC was formed in 1982 by the member organizations of the Joint Committee on InterSociety Coordination (JCIC): the Electronic Industries Association (EIA), the Institute of Electrical and Electronic Engineers (IEEE), the National Association of Broadcasters (NAB), the National Cable Telecommunications Association (NCTA), and the Society of Motion Picture and Television Engineers (SMPTE). Currently, there are approximately 150 members representing the broadcast, broadcast equipment, motion picture, consumer electronics, computer, cable, satellite, and semiconductor industries.

ATSC Digital TV Standards include digital high definition television (HDTV), standard definition television (SDTV), data broadcasting, multichannel surround-sound audio, and satellite direct-to-home broadcasting.

Note: The user's attention is called to the possibility that compliance with this standard may require use of an invention covered by patent rights. By publication of this standard, no position is taken with respect to the validity of this claim or of any patent rights in connection therewith. One or more patent holders have, however, filed a statement regarding the terms on which such patent holder(s) may be willing to grant a license under these rights to individuals or entities desiring to obtain such a license. Details may be obtained from the ATSC Secretary and the patent holder.

This specification is being put forth as a Candidate Standard by the TG3/S34 Specialist Group. This document is a revision of the Working Draft (S34-187r3) dated 30 March 2016. All ATSC members and non-members are encouraged to review and implement this specification and return comments to cs-editor@atsc.org. ATSC Members can also send comments directly to the TG3/S34 Specialist Group. This specification is expected to progress to Proposed Standard after its Candidate Standard period.

Revision History

Version	Date
Candidate Standard approved	3 May 2016
Updated CS approved	15 June 2016
Standard approved	[date]

Table of Contents

Index of Figures and Tables

ATSC Candidate Standard:
A/342 Part 1, Audio Common Elements

[Note: Items shown in *blue italic* are to be completed during the Candidate Standard period.]

1. SCOPE

This document specifies the common framework for ATSC 3.0 Audio. It is intended to be used in conjunction with the specific audio technologies described in subsequent parts of this Standard [2] [3].

1.1 Introduction and Background

The ATSC 3.0 audio system provides immersive and personalizable sound for television. It is not compatible with the audio system used in ATSC 1.0 service [7].

1.2 Organization

This document is organized as follows:

- Section 1 – Outlines the scope of this document and provides a general introduction.
- Section 2 – Lists references and applicable documents.
- Section 3 – Provides a definition of general terms, acronyms, and abbreviations for this document.
- Section 4 – Audio Glossary (defines specialized audio terminology used in this document and its references, with mapping of those items that are identically defined but named differently in those references).
- Section 5 – System overview
- Section 6 – Specification of Common Elements for ATSC 3.0 Audio

2. REFERENCES

All referenced documents are subject to revision. Users of this Standard are cautioned that newer editions might or might not be compatible.

2.1 Normative References

The following documents, in whole or in part, as referenced in this document, contain specific provisions that are to be followed strictly in order to implement a provision of this Standard.

[1]   IEEE: "Use of the International Systems of Units (SI): The Modern Metric System," Doc. SI 10, Institute of Electrical and Electronics Engineers, New York, N.Y.

[2]   ATSC: "ATSC Candidate Standard, A/342 Part 2: AC-4 Audio System," Doc. A/342-2, Advanced Television Systems Committee, Washington, D.C., 15 June 2016. (Work in process.)

[3]   ATSC: "ATSC Candidate Standard, A/342 Part 3: MPEG-H Audio System," Doc. A/342-3, Advanced Television Systems Committee, Washington, D.C., 3 May 2016. (Work in process.)

[4]   ATSC: "ATSC Candidate Standard: Signaling, Delivery, Synchronization, and Error Protection," Doc. A/331:2016, Advanced Television Systems Committee, Washington, D.C., 5 January 2016.

[5]   IETF: "Tags for Identifying Languages," Doc. RFC 5646, Internet Engineering Task Force, Fremont, CA., September 2009.

[6]   ISO/IEC: "Information technology -- Dynamic adaptive streaming over HTTP (DASH) -- Part 1: Media presentation description and segment formats," Doc. 23009-1:2014, International Standards Organization, Geneva, Switzerland, 15 May 2014.

2.2   Informative References

The following documents contain information that may be helpful in applying this Standard.

[7]   ATSC: "Digital Audio Compression (AC-3) (E-AC-3) Standard," Doc. A/52:2015, Advanced Television Systems Committee, Washington, D.C., 24 November 2015.

[8]   *DASH-IF: "DASH-IF Profile for ATSC Broadcast" <Forthcoming document; reference to be completed during CS phase, or if document remains unpublished, reference to be removed.>*

3.   DEFINITION OF TERMS

With respect to definition of terms, abbreviations, and units, the practice of the Institute of Electrical and Electronics Engineers (IEEE) as outlined in the Institute's published standards [1] shall be used. Where an abbreviation is not covered by IEEE practice or industry practice differs from IEEE practice, the abbreviation in question will be described in Section 3.3 of this document.

3.1   Compliance Notation

This section defines compliance terms for use by this document:

shall – This word indicates specific provisions that are to be followed strictly (no deviation is permitted).

shall not – This phrase indicates specific provisions that are absolutely prohibited.

should – This word indicates that a certain course of action is preferred but not necessarily required.

should not – This phrase means a certain possibility or course of action is undesirable but not prohibited.

3.2   Treatment of Syntactic Elements

This document contains symbolic references to syntactic elements used in the audio, video, and transport coding subsystems. These references are typographically distinguished by the use of a different font (e.g., restricted), may contain the underscore character (e.g., sequence_end_code) and may consist of character strings that are not English words (e.g., dynrng).

3.2.1   Reserved Elements

One or more reserved bits, symbols, fields, or ranges of values (i.e., elements) may be present in this document. These are used primarily to enable adding new values to a syntactical structure without altering its syntax or causing a problem with backwards compatibility, but they also can be used for other reasons.

The ATSC default value for reserved bits is '1.' There is no default value for other reserved elements. Use of reserved elements except as defined in ATSC Standards or by an industry standards setting body is not permitted. See individual element semantics for mandatory settings and any additional use constraints. As currently reserved elements may be assigned values and

meanings in future versions of this Standard, receiving devices built to this version are expected to ignore all values appearing in currently-reserved elements to avoid possible future failure to function as intended.

3.3   Acronyms and Abbreviation

The following acronyms and abbreviations are used within this document.

ATSC – Advanced Television Systems Committee

HOA – Higher Order Ambisonics

ISOBMFF – International Standards Organization Base Media File Format

LFE – Low Frequency Effects (Channel)

NGA – Next Generation Audio

SAP – Secondary Audio Programming

VDS – Video Description Service

3.4   Terms

The following terms are used within this document.

reserved – Set aside for future use by a Standard.

4. AUDIO GLOSSARY

This section defines the specific terminology used for the ATSC 3.0 audio system. The terms defined in Section 4.1 are common terms, and may, in some cases, map to alternative terms used by individual systems specified in subsequent parts of this standard [2] [3]. A mapping to those terms is provided in Section 4.2. Figure 4.1 illustrates the relationship between several defined terms.

4.1   Common Terms

Common terms are given in Table 4.1. The relationship of key terms is illustrated in Figure 4.1.

Table 4.1 Common Terms as they Apply to this Standard

Term	Description
2.0	Nomenclature for stereo audio, with two audio channels (L, R), as found in legacy television audio systems.
5.1	Nomenclature for surround audio, with five full-range audio channels (L, C, R, LS, RS) and one low-frequency effects (LFE) channel, as found in the existing ATSC digital television audio system.
7.1+4	Nomenclature for a particular 11.1 loudspeaker arrangement suitable for Immersive Audio, consisting of three frontal loudspeakers (L, C, R) and four surround loudspeakers (left side [LS], left rear [LR], right side [RS], right rear [RR]) on the listener's plane, and four speakers placed above the listener's head height (arranged in LF, RF, LR and RR positions).
Audio Element	The smallest addressable unit of an *Audio Program*. Consists of one or more *Audio Signals* and associated *Audio Element Metadata,* and can be configured as any of three different *Audio Element Formats*. (See Figure 4.1**.**)
Audio Element Format	Description of the configuration and type of an *Audio Element*. Notes: There are three different types of Audio Element Formats. Depending on the type, different kinds of properties are used to describe the configuration: 　Channel-based audio: e.g., the number of channels and the channel layout 　Object-based audio: e.g., dynamic positional information

	Scene-based audio: e.g., HOA order, number of transport channels
Audio Element Metadata	Metadata associated with an *Audio Element*. Notes: Some examples of Audio Element Metadata include positional metadata (spatial information describing the position of objects in the reproduction space, which may dynamically change over time, or channel assignments), or personalization metadata (set by content creator to enable certain personalization options such as turning an element "on" or "off," adjusting its position or gain, and setting limits within which such adjustments may be made by the user). (See Section 4.2 for alternate nomenclature used for this term in other documents.)
Audio Object	An *Audio Element* that consists of an *Audio Signal* and *Audio Element Metadata,* which includes rendering information (e.g., gain and position) that may dynamically change. Audio Objects with rendering information that does <u>not</u> dynamically change may be called "static objects."
Audio Presentation	A set of *Audio Program Components* representing a version of the *Audio Program* that may be selected by a user for simultaneous decoding. Notes: An Audio Presentation is a sub-selection from all available *Audio Program Components* of one *Audio Program*. (See Figure 4.1.) A Presentation can be considered the NGA equivalent of audio services in predecessor systems, which each utilized complete mixes (e.g., "SAP" or "VDS") (See Section 4.2 for alternate nomenclature used for this term in other documents.)
Audio Program	The complete collection of all *Audio Program Components* and a set of accompanying *Audio Presentations* that are available for one Audio Program. (See **Figure 4.1.**) Notes: Not all *Audio Program Components* of one Audio Program are necessarily meant to be presented at the same time. An Audio Program may contain *Audio Program Components* that are always presented, and it may include optional *Audio Program Components*. (See Section 4.2 for alternate nomenclature used for this term in other documents.)
Audio Program Component	A logical group of *Audio Elements* that is used to define an *Audio Presentation* and may consist of one or more *Audio Elements.* (See Figure 4.1.) (See Section 4.2 for alternate nomenclature used for this term in other documents.)
Audio Program Component Type	Characterization of an *Audio Program Component* with regard to its content. Notes: Examples for Audio Program Component Types are: Complete Main Music & Effects (M&E): the background signal that contains a Mix of various Audio Signals except speech. Dialog: one or more Audio Signals that contain only speech Video Description Service
Audio Signal	A mono signal. (See Figure 4.1.)
Bed	An *Audio Element* that is intended to be used as the foundational element of an *Audio Presentation* (e.g., Music & Effects), to which other complementing Audio Elements (e.g., Dialog) are added.
Channel Set	A group of *Channel Signals* that are intended to be reproduced together.
Channel Signal	An *Audio Signal* that is intended to be played back at one specific nominal loudspeaker position.
Complete Mix	All *Audio Elements* of one *Audio Presentation* mixed together and presented as a single *Audio Program Component*.
Elementary Stream	A bit stream that consists of a single type of encoded data (audio, video, or other data). Notes: The *Audio Elements* of one *Audio Program* may be delivered in a single audio Elementary Stream or distributed over multiple audio Elementary Streams. (See Section 4.2 for alternate nomenclature used for this term in other documents.)
Higher-Order Ambisonics	A technique in which each produced signal channel is part of an overall description of the entire sound scene, independent of the number and locations of actually available loudspeakers.
Immersive Audio	An audio system that enables high spatial resolution in sound source localization in azimuth, elevation and distance, and provides an increased sense of sound envelopment.

| LFE	Low-frequency effects channel. A limited frequency response channel that carries only low frequency (e.g., 100 Hz and below) audio.
Mix	A number of *Audio Elements* of one *Audio Program* that are mixed together into one *Channel Signal* or into a *Bed*.
Rendering	The realization of aural content for acoustical presentation.
Track	Representation of an *Elementary Stream* that is stored in a file format like the ISO Base Media File Format. Notes: For some systems, it may be possible to directly store the unmodified data from the *Elementary Stream* into a Track, whereas for other systems it may be necessary to re-format the data for storage in a Track.
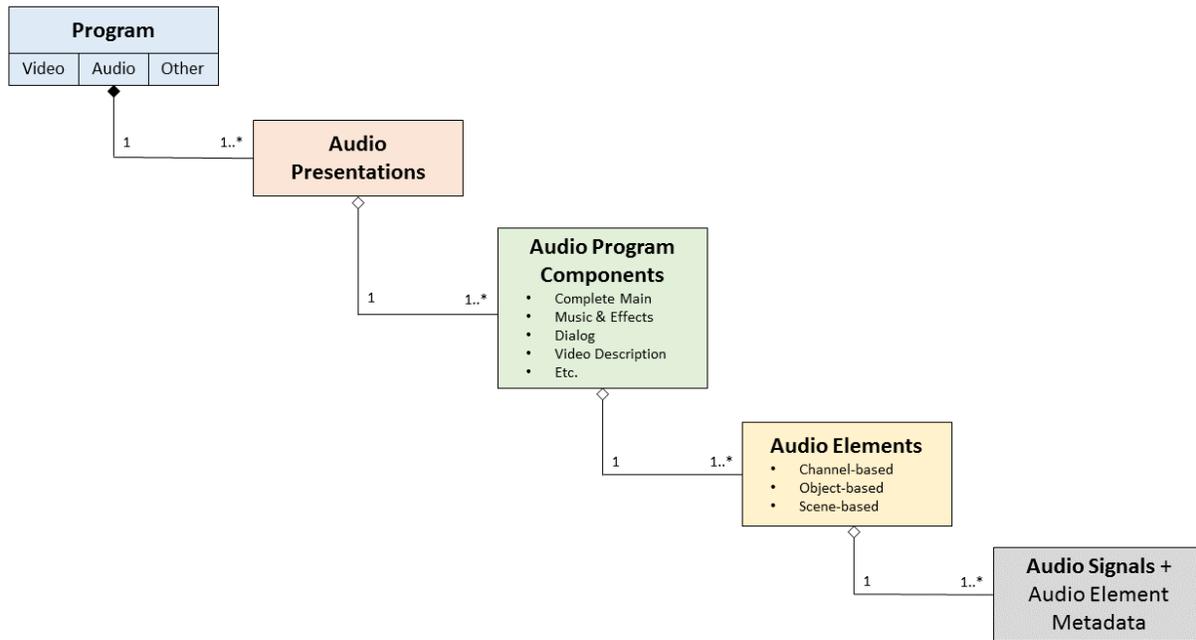
Figure 4.1 Relationship of key audio terms.

4.2 Mapping of Terms to Specific Technologies

Table 4.2 lists the alternative terms used for the items defined above by the individual systems defined in subsequent parts of this standard, and by the DASH-IF.

Table 4.2 Mapping of Alternative Terms to Audio Glossary Common Terms

Common Term	DASH-IF Term [8]	AC-4 Term [2]	MPEG-H Audio Term [3]
Audio Element Metadata		Metadata, Object Audio Metadata	Metadata Audio Elements (MAE), Object Metadata (OAM)
Audio Presentation	Preselection	Presentation	Preset
Audio Program	Bundle	Audio Program	Audio Scene
Audio Program Component	Referred to as Audio Element	Audio Program Component	Group
Elementary Stream	Representation in an Adaptation Set	Elementary Stream	Elementary Stream

5. SYSTEM OVERVIEW

5.1   Audio System Features

5.1.1   Immersive and Legacy Support

The ATSC 3.0 audio system supports Immersive Audio with enhanced performance when compared with existing 5.1 channel-based systems.

The system supports delivery of audio content from mono, stereo, 5.1 channel and 7.1 channel audio sources, as well as from sources supporting Immersive Audio. Immersive features are supported over the listening area. Such a system might not directly represent loudspeaker feeds but instead could represent the overall sound field.

5.1.2   Next Generation Audio System Flexibility

The ATSC 3.0 audio system enables Immersive Audio on a wide range of loudspeaker configurations, including loudspeaker configurations with suboptimum loudspeaker locations, and headphones.

The system enables audio reproduction on loudspeaker configurations not designed for Immersive Audio such as 7.1 channel, 5.1 channel, two channel and single channel loudspeaker configurations.

5.1.3   Personalization and Interactive Control

The ATSC 3.0 audio system enables user control of certain aspects of the sound scene that is rendered from the encoded representation (e.g., relative level of dialog, music, effects, or other elements important to the user).

The system enables user-selectable alternative audio Tracks to be delivered via terrestrial broadcast or via broadband and in Real Time or Non-real Time. Such audio Tracks may be used to replace the primary audio Track or be mixed with the primary audio Track and delivered for synchronous presentation with the corresponding video content.

The system enables receiver mixing of alternative audio Tracks (e.g., assistive audio services, other language dialog, special commentary, music and effects) with the main audio Track or other audio Tracks, with relative levels and position in the sound field and receiver adjustments suitable to the user.

The system enables broadcasters to provide users with the option of varying the loudness of a TV program's dialog relative to other elements of the audio Mix to increase intelligibility.

5.1.4   Next Generation Audio System Loudness Management and Dynamic Range Control

The ATSC 3.0 audio system supports information and functionality to normalize and control the loudness of reproduced audio content.

The system enables adapting the loudness and dynamic range of audio content as appropriate for the receiving device and environment of the content presentation.

5.1.5   Accessible Emergency Information

The ATSC 3.0 audio system supports the inclusion and signaling of audio (speech) that provides an aural representation of emergency information provided by broadcasters in on-screen text display (static, scrolling or "crawling" text).

Note that this is not Emergency Alerting, but rather contains additional emergency information provided by broadcasters.

5.1.5.1 Accessible Emergency Information Signaling

<TBD – to be completed during CS Phase.>

5.1.5.2 Insertion of Accessible Emergency Information by Specific Technologies

<TBD – to be completed during CS Phase.>

5.2 Audio System Architecture

The ATSC 3.0 system is designed with a "layered" architecture in order to leverage the many advantages of such system, particularly pertaining to upgradability and extensibility. A generalized layering model for ATSC 3.0 is shown in Figure 5.1. The ATSC 3.0 audio system resides in the upper layer (Applications & Presentation). Audio system signaling resides primarily in the middle layer (Management & Protocols).
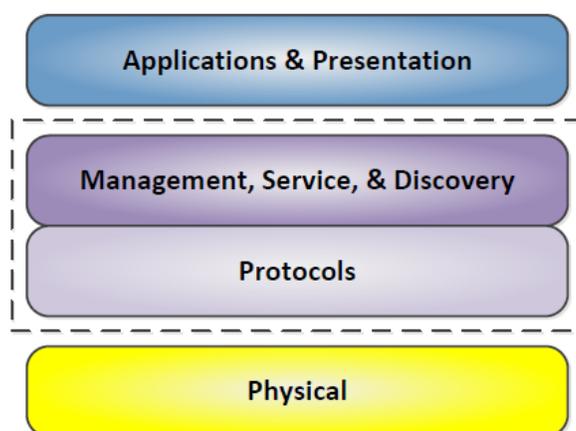

Figure 5.1 ATSC 3.0 generalized layer architecture.

5.3 Central Concepts

Several concepts are common to all audio systems supported by ATSC 3.0. This section describes these common concepts.

5.3.1 Audio Program Components and Presentations

Audio Program Components are separate pieces of audio data that are combined to compose an Audio Presentation. A simple Audio Presentation may consist of a single Audio Program Component, such as a Complete Main Mix for a television program. Audio Presentations that are more complex may consist of several Audio Program Components, such as ambient music and effects, combined with dialog and video description.

  Audio Presentations are combinations of Audio Program Components representing versions of the audio program that may be selected by a user. For example, a complete audio with English dialog, a complete audio with Spanish dialog, a complete audio (English or Spanish) with video description, or a complete audio with alternate dialog may all be selectable Presentations for a Program.

  The Components of a Presentation can be delivered in a single audio Elementary Stream or in multiple audio Elementary Streams. Signaling and delivery of audio Elementary Streams is documented in ATSC A/331 [4].

5.3.2   Audio Element Formats

The ATSC 3.0 audio system supports three fundamental Audio Element Formats:

1) Channel Sets are sets of Audio Elements consisting of one or more Audio Signals presenting sound to speaker(s) located at canonical positions. These include configurations such as mono, stereo, or 5.1, and extend to include non-planar configurations, such as 7.1+4.

2) Audio Objects are Audio Elements consisting of audio information and associated metadata representing a sound's location in space (as described by the metadata). The metadata may be dynamic, representing the movement of the sound.

3) Scene-based audio (e.g., HOA) consists of one or more Audio Elements that make up a generalized representation of a sound field.

5.3.3   Audio Rendering

Audio Rendering is the process of composing an Audio Presentation and converting all the Audio Program Components to a data structure appropriate for the audio outputs of a specific receiver. Rendering may include conversion of a Channel Set to a different channel configuration, conversion of Audio Objects to Channel Sets, conversion of scene-based sets to Channel Sets, and/or applying specialized audio processing such as room correction or spatial virtualization.

5.3.3.1   Video Description Service (VDS)

Video Description Service is an audio service carrying narration describing a television program's key visual elements. These descriptions are inserted into natural pauses in the program's dialog. Video description makes TV programming more accessible to individuals who are blind or visually impaired. The Video Description Service may be provided by sending a collection of "Music and Effects" components, a Dialog component, and an appropriately labeled Video Description component, which are mixed at the receiver. Alternatively, a Video Description Service may be provided as a single component that is a Complete Mix, with the appropriate label identification.

5.3.3.2   Multi-Language

Traditionally, multi-language support is achieved by sending Complete Mixes with different dialog languages. In the ATSC 3.0 audio system, multi-language support can be achieved through a collection of "Music and Effects" streams combined with multiple dialog language streams that are mixed at the receiver.

5.3.3.3   Personalized Audio

Personalized audio consists of one or more Audio Elements with metadata, which describes how to decode, render, and output "full" Mixes. Each personalized Audio Presentation may consist of an ambience "bed", one or more dialog elements, and optionally one or more effects elements. Multiple Audio Presentations can be defined to support a number of options such as alternate language, dialog or ambience, enabling height elements, etc.

There are two main concepts of personalized audio:

1) Personalization selection – The bit stream may contain more than one Audio Presentation where each Audio Presentation contains pre-defined audio experiences (e.g. "home team" audio experience, multiple languages, etc.). A listener can choose the audio experience by selecting one of the Audio Presentations.

2) Personalization control – Listeners can modify properties of the complete audio experience or parts of it (e.g., increasing the volume level of an Audio Element, changing the position of an Audio Element, etc.).

6. SPECIFICATION

6.1 Audio Constraints

The following constraints are applied to all audio content in ATSC 3.0 services.

6.1.1 Sampling Rate

The sampling frequency of Audio Signals shall be 48 kHz.

6.1.2 Audio Program Structure

An Audio Program shall consist of one or more Audio Presentations. One Audio Presentation shall be signaled as the default (main), and shall have all of its Audio Program Components present in the broadcast stream. The main Audio Presentation is intended to be the default in cases where no other selection guidance (user-originated or otherwise) exists.

Audio Presentations shall consist of at least one Audio Program Component of any Audio Element Format.

Audio Program Components may be delivered in more than one Elementary Stream. For example, one Elementary Stream may be delivered over broadcast and an additional Elementary Stream may be delivered over a broadband connection. Audio Presentations other than the default Presentation may include Audio Program Components from multiple Elementary Streams. Audio Presentations shall not utilize Audio Program Components from more than three Elementary Streams.

Further constraints are defined in subsequent Parts of this standard.

6.1.3 General Elementary Stream Structure

Audio Elementary Streams shall be packaged and signaled in ISOBMFF in a configuration specified by the A/331 standard [4].

6.2 Signaling of Audio Characteristics

Table 6.1 describes the audio characteristics that are signaled in the delivery layer [4].

Table 6.1 Audio Characteristics

Item	Name	Description	Options
1	Codec	Indicates the codec and resources required to decode the bit stream.	FourCC (i.e., ac-4, mhm1, mhm2) followed by codec specific level or version indicators.
2	Role	Indicates the role of the default (entry point) presentation or preset	Values as defined by ISO/IEC 23009-1 [6].
3	Language	Indicates the language of a presentation or preset	RFC 5646 language codes [5]
4	Accessibility	Indicates the accessibility features of a presentation or preset	*TBD*
5	Sampling Rate	Output sampling rate	48000
6	Audio channel configuration	Indicates the channel configuration and layout.	Codec specific
7	Presentation or preset identifier	Indicates IDs for each presentation or preset	Codec specific

The audio system shall operate according to A/342-2 when the transport layer signals that the item 1 codec parameter is equal to 'ac-4', and according to A/342-3 when the transport layer signals that the item 1 codec parameter is equal to 'mhm1' or 'mhm2'.

Annex A: Examples of Common Broadcast Operating Profiles

A.1   OPERATING PROFILES

Table A.1.1 lists some broadcast operating-profile examples and shows how the input elements for each profile fit into presentations or presets within a single elementary stream. Figure A.1.1 illustrates the encoding of some of the broadcast operating-profile examples. Note that these examples are not exhaustive and are included to demonstrate common/practical operating profiles.

The following notations are used in Table A.1.1 and Figure A.1.1:

- CM = Complete Main
- M&E = Music and Effects
- Dx = Dialog element (mono)
- VDS = Video Descriptive Service (mono)
- O = Other object (mono), i.e. PA feed
- O(15).1 = 15 object or spatial object groups + LFE
- HOA(X) = 6th Order Higher Order Ambisonics sound-field represented by X Audio Signal transport channels

Table A.1.1 Encoding of Example Broadcast Operating Profiles

|	Profile Type	Input Elements	Presentations/Presets	Elements Referenced by Presentation/Preset
1	Complete Main	2.0 CM	CM	CM
2		5.1 CM	CM	CM
3		HOA(6) CM	CM	CM
4		5.1.2 CM	CM	CM
5		7.1.4 CM	CM	CM
6		HOA(12) CM	CM	CM
7		O(15).1 CM	CM	CM
8	M&E + Objects	2.0 M&E + D	English	M&E + D
			M&E Only	M&E
9		5.1 M&E + D1 (en) + D2 (es) + VDS (en)	English	M&E + D1
			English + VDS	M&E + D1 + VDS
			Spanish	M&E + D2
			M&E Only	M&E
10		HOA(6) + D1 (en) + D2 (es) + VDS (en)	English	M&E + D1
			English + VDS	M&E + D1 + VDS
			Spanish	M&E + D2
			M&E Only	M&E
11		5.1.2 M&E + D1 (en) +D2 (es) + VDS (en)	English	M&E + D1
			English + VDS	M&E + D1 + VDS
			Spanish	M&E + D2
			M&E Only	M&E
12		7.1.4 M&E + D1 (en) + D2 (es) + VDS (en) + O	English	M&E + O + D1
			English + VDS	M&E + D1 + VDS
			Spanish	M&E + O + D2
			M&E	M&E + O
13		O(15).1 M&E + D1 (en) + D2 (es) + VDS (en)	English	M&E + D1
			English + VDS	M&E + D1 + VDS
			Spanish	M&E + D2
			M&E Only	M&E
14		HOA(12) M&E + D1 (en) + D2 (es) + VDS (en) + O	English	M&E + O + D1
			English + VDS	M&E + D1 + VDS
			Spanish	M&E + O + D2
			M&E	M&E + O
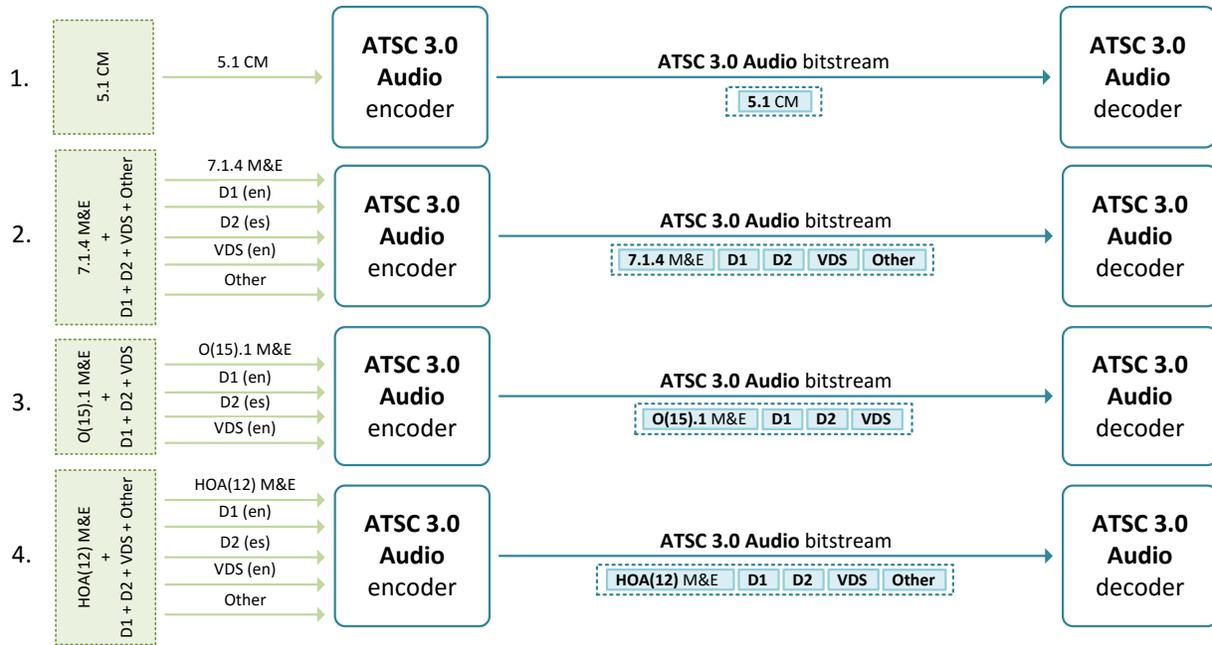
Figure A.1.1 Encoding of example broadcast operating profiles.

End of Document