An Overview of Audio Technologies, Immersion and Personalization Features envisaged for the TV3.0

Regis Rossi Alves Faria
Almir Antônio Rosa,
Eduardo Mendes,
Ana Amélia Benedito Silva,
Douglas Henrique Siqueira Abreu,
Henrique Rozena

Cite this article:

Faria, Regis Rossi Alves; Rosa, Almir Antônio; Mendes, Eduardo; Silva, Ana Amélia Benedito; Abreu, Douglas Henrique Siqueira; Rozena, Henrique; 2023. An Overview of Audio Technologies, Immersion and Personalization Features envisaged for the TV3.0. SET INTERNATIONAL JOURNAL OF BROADCAST ENGINEERING. ISSN Print: 2446-9246 ISSN Online: 2446-9432. doi: 10.18580/setijbe.2023.2. Web Link: https://dx.doi.org/10.18580/setijbe.2023.2

SET INTERNATIONAL JOURNAL OF BROADCAST ENGINEERING - SET IJBE V.9, 2023, Article 2, 15 p
2023 SET - Brazilian Society of Television Engineering / ISSN (Print): 2446-9246 / ISSN (Online): 2446-9432

An Overview of Audio Technologies, Immersion and Personalization Features envisaged for the TV3.0

Regis Rossi Alves Faria, Almir Antônio Rosa, Eduardo Mendes,
Ana Amélia Benedito Silva, Douglas Henrique Siqueira Abreu, Henrique Rozena

***Abstract*—In 2021 the Forum of the Brazilian Digital Terrestrial Television System (SBTVD) accomplished the phase 2 of the TV3.0 project, consisting of a series of tests of the technologies proposed for this next generation of TVD system in Brazil. The tests were conducted by research groups of Brazilian universities. Particularly referring to the audio coding layer of the system, we carried out at the University of São Paulo 13 groups of tests, as prescribed in a public Call for Proposals (CfP), and could assess the technologies capabilities and versatility in providing a series of new features for the next generation of audio for the digital broadcasting system. This paper summarizes the main results of this testing and evaluation phase, and brings an overview of the stimulating new features that content producers and audience would have available to create and consume immersive and personalized services at home.**

***Index Terms*— Next-generation Audio, Immersive Audio, Audio Coding, Audio Personalization, TV3.0**

I. Introduction

THIS paper aims at presenting an overview of the results of the evaluations and tests carried out on the candidate technologies proposed for the audio coding layer of the next-generation of DTV (TV 3.0 system), detaching their distinguished features on audio immersivity and content personalization. Two of them were tested in laboratory and could have their features evaluated and their compliance verified against the requirements defined in the public Call for Proposals (CfP) issued by the Forum of the Brazilian Digital Terrestrial Television System (SBTVD) [1]. The methodology included the execution of 13 tests groups, applied to each of the technology systems, and employing a common set of audio test content material, so as to permit assessment of performance and conformance issues based on a common ground.

The paper is to be organized as follows: first we present the 13 tests groups, their main features and technical requirements. Then, we present an overview of the technologies considered, and the audio test content employed in the tests. Following, the test cases performed in the laboratory are described, illustrating the facilities, employed setups, operating steps and conditions. Next, considering the fulfilled features and requirements, we give a glance on the potential of the new services expected in terms of audio immersion and content personalization control capabilities that the sound in the TV3.0 will offer, and finish with some pertinent discussions for the present and future.

II. Tests groups

The tests groups concerning the audio coding layer performed in the Testing and Evaluation phase of the TV3.0 SBTVD project were defined in the Phase 2 Call for Proposals (CfP) document first issued on December, 2020 by the Forum SBTVD and revised on March 15, 2021 [1]. The document describes 13 groups of Test Cases (TC) and the respective requirements (AC) for each Test Case, detailing the minimum technical specifications to be fulfilled. The tests were numbered from 1 to 13.

In this section we describe the Test Cases (TC), their requirements (AC), and the main features addressed to be evaluated, as specified in the CfP.

1. Test 1 (Immersive audio)

Test 1 addressed three TCs. In TC1.1, the requirements to the tested system were to demonstrate its ability to present audio in the specified channel mode, and to present the audio objects rendered together to various output setups (e.g., 2.0, 5.1, and 5.1+4H channels), in accordance with the target bitrate for this TC.

In TC1.2, the requirement was to demonstrate the system's ability to present scene-based HOA (Higher-Order Ambisonics) content in accordance with the target bitrate for this TC.

2. Test 2 (Interactivity and personalization)

Test 2 consisted of seven TCs. For TC2.1, the requirement "AC2.1: Language Selection" addressed the system's ability to allow end-users to select between multiple audio languages based on user interaction or automatic language selection (e.g., the receiver's preferred audio settings).

For TC2.2, the requirement AC2.1 (Selection of different preselections) targeted the system's ability to allow (or not) end-users to select between different preselections created in

This work was supported in part by CNPq, the Brazilian National Council for Scientific and Technological Development.

Regis Faria (regis@usp.br), Almir Almas (alalmas@usp.br), Ana Amélia B. Silva (aamelia@usp.br) and Eduardo Mendes (edusm@usp.br) are with the University of São Paulo.

Douglas Abreu (d229998@dac.unicamp.br) is with the School of Electrical and Computing Engineering, Universidade Estadual de Campinas, Campinas, Brazil.

https://set.org.br/ijbe doi: 10.18580/setijbe.2023.2 Web Link: https://dx.doi.org/10.18580/setijbe.2023.2

the production. TC2.3 evaluated the system's ability to switch between multiple commentators (e.g., during a sports event the user at home could switch between the usual commentator and the premium commentator or local team commentator).

For the Test Case TC2.4 the requirement (AC2.1: Display of textual labels) concerned the system's ability to display to the end-users correct textual labels for all audio objects that allow interactivity options and preselections as created in production. Test Case TC2.5 evaluated the system's ability to enable the end-users to interact with any audio object and adjust the object level at the end-user's device according to broadcaster settings in production (requirement AC2.2: Audio object loudness interactivity changing the level relative to the background). The user should be able to increase/decrease the object level (relative to the background) inside a range of min/max gain specified by the broadcaster, which might differ for each object.

The TC2.6 evaluated the requirement "AC2.3: Audio object interactivity, changing the object position". The test consisted of demonstrating the system's ability to enable the end-users to interact with any object and to adjust the sound object position at the end-user's device, according to broadcaster settings.

The last TC in this group – TC2.7 – addressed the requirement "AC2.4: Enable Interactivity when using external sound reproduction systems". It aimed to demonstrate the system's ability to enable interactivity when using external sound reproduction devices (e.g. soundbar/AVR, home theaters). The tests should demonstrate the system's ability to enable the interactivity options on the main receiving device (e.g., TV/STB) while the immersive sound is reproduced by the external sound reproduction device.

3. *Test 3 (Audio description)*

Test 3 consisted of four Test Cases addressing the selection and use of audio description content, delivered as an additional audio object with associated metadata.

For the TC3.1, the tested requirements were "AC3.1 and AC3.2: Audio description in the same stream as the main audio". It aimed to demonstrate the system's ability to enable audio description delivered in the same stream as the main audio (e.g., a single stream containing the main audio mix and alternative mix with audio description).

Test Case TC3.2 (requirement "AC3.3 Part 1: Audio description delivered as an additional audio object") consisted in demonstrating the system's ability to enable the audio description service, when available, and Test Case TC3.3 (requirement "AC3.3 Part 2: Audio description delivered as additional audio objects and language selection") aimed to demonstrate the end user's ability to enable/disable audio description available in multiple languages.

For the Test Case TC3.4, the requirement was: "AC3.3 Part 3: Audio description delivered as additional audio objects and spatial separation of main dialog and audio description". This test should demonstrate the system's ability to enable/disable audio description and spatially separate the main dialog and the audio description for better speech intelligibility".

4. *Test 4 (Audio emergency warning information)*

The test 4 consisted of only one TC (TC4.1) which should demonstrate how the audio system can deliver emergency warning information audio content. The test concerned showing what metadata is carried in the audio bitstream and how other applications could access or process this metadata to achieve the same result (e.g. some sort of API specification).

5. *Test 5 (Flexible audio playback configuration)*

For Test 5 (Flexible audio playback configuration), only one TC addressing the requirements "AC5.1 and AC5.2". The test should demonstrate the system's ability to decode and render the same content using multiple audio playback configurations and systems, including TV loudspeakers, soundbars, home theaters (immersive and 5.1 AVRs), and binaural.

6. *Test 6 (Consistent loudness)*

For Test 6, three TCs were conducted. TC6.1 addressed the requirement "AC6.1: Loudness Normalization Test - Programs" and the test should demonstrate the system's ability to achieve the target loudness level across multiple programs, i.e., to evaluate the effectiveness of solutions for guaranteeing a consistent loudness experience, without undesired (and sometimes exaggerated) volume changes between different programs.

TC6.2 addressed the requirement AC6.2 (Loudness Normalization Test for Preselections) and should demonstrate the ability to preserve the target loudness level across multiple preselections inside the same program. TC6.3 (requirement AC6.2: Loudness Compensation Test) should demonstrate the ability to preserve the target loudness level after user interaction (e.g., if the user increases the level of the dialog the overall loudness shall not increase).

7. *Test 7 (Seamless configuration changes and Audio/Video alignment)*

For the Test 7, seven TCs were evaluated. Test Case TC7.1 addressed the requirement "AC7.1: Seamless configuration changes" and considered demonstrating the system's ability to seamlessly play back content during configuration changes. Configuration changes between available configurations could include, for instance, combinations between 2.0, 5.1, and 5.1+4H output formats.

TC7.2 (requirement "AC7.2: Seamless content playback during user interaction") should demonstrate the ability to seamlessly playback content during user interaction, such as changes between different audio languages or preselections, increasing or decreasing the level of various audio objects, without audio drop-outs or glitches.

TC7.3 (requirement "AC7.3 Part 1: Seamless content playback during changes in production") should demonstrate the system's ability to seamlessly playback content during changes in production during a live broadcast. This test employed a special live setup, with equipment and software to permit a live edition of the broadcasting settings. Typical changes in a live broadcast should be tested, including:

a. Change the audio scene (objects, preselections, etc.);
b. Enable/disable dialogs in multiple languages;

https://set.org.br/ijbe doi: 10.18580/setijbe.2023.2 Web Link: https://dx.doi.org/10.18580/setijbe.2023.2

c. Enable/disable Audio Description in multiple languages;

d. Enable/disable interactivity options for one or more preselections;

e. Change the interactivity options (min/max gain and position values) for one or more objects;

f. Change the textual labels for one or more objects or preselections.

TC7.4 (requirement AC7.3 Part 2: Seamless content playback during changes in production using a contribution feed) should demonstrate the system's ability to seamlessly playback during changes in production in a live broadcast scenario.

TC7.5 (requirement "AC7.4 Part 1: Seamless Ad-Insertion") should demonstrate the system's ability to enable seamless advertisement insertion at any time instance, e.g., switch between the main feed authored live (e.g. a content playout 1) and an additional feed containing a pre-authored advertisement break (e.g. a content playout 2).

TC7.6 (requirement "AC7.4 Part 2: User selection persistency after the Ad-break") should demonstrate the ability to preserve the user interaction settings after the ad-break, e.g., if the user selects, before the ad- break, the English language (EN) and increases the dialog level with 7 dB, after the ad-break the content will start with the exact same settings.

TC7.7 (requirement "AC7.4 Part 3: Hybrid Delivery") should demonstrate the system's ability to synchronize and replace the main soundtrack delivered via broadcast for an alternative audio signal delivered via broadband (Internet link).

8. Test 8 (Audio coding efficiency)

For Test 8 (Audio coding efficiency), the CfP document specified that "the proponent's documentation provided on the Quality Assessment Reports should provide the data to analyze the audio coding efficiency". Therefore, this TC focused on the analysis of available technical assessments and previous studies, such as subjective tests, conducted by third parties [2-6], mainly to evaluate the system's ability to deliver the minimum MUSHRA (ITU-R BS.1534-3 Multiple Stimuli with Hidden Reference and Anchor, a subjective test methodology) quality scores for several audio formats and target bitrates

9. Test 9 (Audio End to end latency)

Test 9 considered two requirements to evaluate the system's ability to provide live audio with minimum end-to-end latency.

The requirement AC9.1 was specified to be tested during the execution of Test Case TC1.1, and considered if the proponent's system was able to encode and decode according to the requirements AC1.1.1, AC1.1.2, and AC1.1.3.

The requirement AC9.2 was not to be analyzed with a feature test, and should be verified through the analysis of the proponent's documentation provided in the Document Analysis phase. As specified in the CfP, the delay (latency) of each module of the real-time test setup should be documented, including the audio and video encoding delay, additional video buffering (if any) before the video encoder, audio decoding and rendering delay, transcoding to a different format delay, and final decoding delay in the external sound reproduction system.

10. Test 10 (Audio/Video synchronization)

Test 10 consisted of only one TC. TC10.1 addressed the requirement "AC10.1" targeted to demonstrate the system's ability to perform adequate A/V synchronization.

11. Test 11 (New immersive audio services)

For Test 11, the single Test Case TC11.1 addressed the requirement "AC11.1" and the test task was to verify the system's ability to perform playback of audio demonstrating one or more of those applications: VR / AR / XR / 3DoF (Degree of Freedom) / 6DoF. A video codec should be chosen by the proponent to be used in this test.

12. Test 12 (Interoperability with different distribution platforms)

Test 12 consisted of a single TC (TC12.1) whose requirements were AC12.1 and AC13.1. The test aimed to demonstrate the system's ability to send multiple audio contents over two or more communications channels.

13. Test 13 (Audio scalability and extensibility)

For Test 13 the CfP specified two requirements: "AC13.1" (for scalability) and "AC13.2" (for extensibility). The use test addressed the system's ability to enable scalability (e.g. to enhance the over-the-air audio experience with additional Internet-delivered audio content, such as new sports commentator options) and extensibility (e.g. support new settings and/or features in the future, in a backward-compatible way).

III. TECHNOLOGIES IN CONSIDERATION

Three international audio coding standards responded to the Call for Proposals (CfP) of technologies and were accepted as candidates to supply their systems for the audio component/layer of the TV3.0 system.

This section provides a concise introduction to the three audio coding technologies taken into consideration in the SBTVD TV3.0 testing and evaluation phase: the AVSA; Dolby Atmos (AC-4); and MPEG-H.

A. AVSA system

The acronym AVSA stands for AVS-Audio, the IEEE 1857.8-2020 Standard for Second Generation Audio Coding, also known as the audio stream that matches AVS2 audio standard [7]. This standard defines a set of tools for the compression, decompression, and packaging of multimedia data, aimed at efficient transmission and storage over the Internet. This standard, an evolution of the IEEE Std. 1857.2-2013, provides a flexible configuration of compression parameters to deliver an improved Quality of Experience (QoE). This includes bitrates ranging from 16 kb/s to 192 kb/s per channel, and supports up to 128 channels for audio signals with a sampling frequency from 8 kHz to 192 kHz and quantization resolutions of 8, 16, and 24 bits.

It presents a defined set of audio encoding tools for the transmission and decoding of recorded music, voice,

https://set.org.br/ijbe doi: 10.18580/setijbe.2023.2 Web Link: https://dx.doi.org/10.18580/setijbe.2023.2

environmental sounds, and instrumentals. The framework of AVS2 (Audio Video Coding Standard 2) is divided into two profiles: the base channel encoding profile (base_profile) and the 3D audio object encoding profile (3D_profile). The 3D object encoding includes object audio data and metadata, allowing for spatial configuration, movement localization, and additional descriptive information such as acoustic properties, directional cues, and volume levels.

AVS2 distinguishes itself by employing adaptive bit rate control and advanced psychoacoustic models to achieve high compression efficiency without compromising sound quality. It offers encoding options for both channel signal and audio object, allowing a flexible configuration between 128 sound objects and 128 channel signals. Furthermore, the GA (General Audio) encoding technology provides multiple encoding options that share a common core module, thus achieving high efficiency and low algorithmic latency.

When compared to its predecessor, AVS2 has improved the degree of 3D audio encoding, achieving a greater compression efficiency and sound quality, and saving up to 50% in bit rate. This improvement makes the standard especially useful for applications and services that include audio accompanying video, TV audio systems, digital audio storage, audio broadcasting, and communication. Hence, the standard is suitable for high-resolution digital broadcasting, digital storage media, broadband wireless multimedia communications, broadband Internet media streaming, digital cinema, and video surveillance.

In the context of the IEEE 1857.8-2020 standard, the data flow of the AVSA system is delineated in several steps, ranging from audio input to encoded audio output. A schematic for AVSA is shown in Fig. 1. Specifically, in addition to the AVS2 audio encoding profiles, the General Audio (GA) Encoding Technology also forms part of its ecosystem. AVSA is currently in use in broadcast services in China.

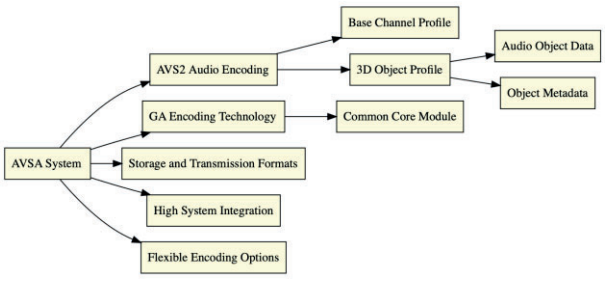


Fig. 1: Description diagram of the AVSA system as defined by the IEEE 1857.8-2020 standard.

B. AC-4 (Dolby) system

The Dolby AC-4 Audio Codec is defined in the technical specification ETSI TS 103 190 (Digital Audio Compression AC-4 Standard) published by the European Telecommunications Standards Institute (ETSI) [8], and represents an innovation in encoding efficiency and application versatility over the previous AC-3 format. It proposes several advancements in audio encoding and also offers flexibility and efficiency for a broad range of applications and output devices. As specified in ETSI TS 103 190-2 (Part 2: Immersive and personalized audio), this codec is designed to support a diverse range of content types, including legacy channel-based, object-oriented immersive, and customized audio [9]. Additionally, AC-4 is compatible with the ATSC (Advanced Television System Committee) 3.0 audio system, leveraging several specific features to enhance the quality and efficiency of audio transmission.

One of the highlighted features of AC-4 is the A/V frame alignment, aimed at mitigating complications associated with the synchronization of multimedia content at segmentation points. When enabled, this feature claims significant simplification on splicing workflows and transcoding to or from formats that utilize video-based frame alignment, such as HD-SDI (Serial Digital Interface, introduced by the Society of Motion Picture and Television Engineers - SMPTE).

The AC-4 system also introduces improvements in dialogue intelligibility through a user-controlled dialogue enhancement feature. Moreover, AC-4 incorporates support for the Extensible Metadata Delivery Format (EMDF), as defined in ETSI TS 103 190-1 (Part 1: Channel based coding). This format allows the transmission of third-party metadata and application data in AC-4 bitstreams, offering a framework for the inclusion of additional user data.

Regarding control of dynamic range and volume, AC-4 adheres to global standards, incorporating an extensive set of volume metadata and Dynamic Range Control (DRC) that are compliant with international norms, including ATSC A/85 [10]. This allows for flexible DRC implementation, adaptable to a wide range of device profiles and user application scenarios.

Another advancement of AC-4 is its volume verification engine, which ensures the accuracy of the transmitted volume metadata. This feature, combined with a real-time volume leveler, can be activated to guarantee audio output consistency.

The AC-4 system offers an array of innovative features for efficient encoding and application versatility, as depicted in Fig. 2. It includes features like A/V frame alignment and DRC, among others. AC-4 has been selected as the Next-Generation Audio (NGA) audio format for the United States, Canada and Mexico, formalized in the ATSC A/300 document.

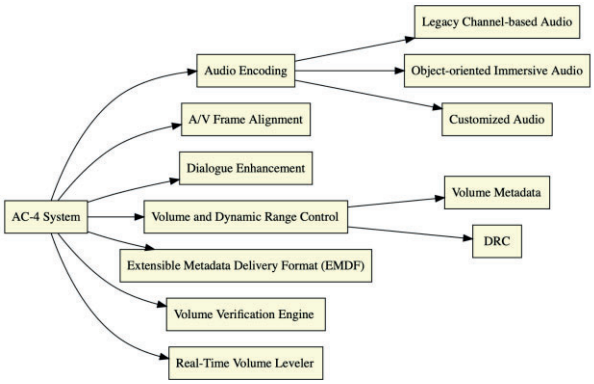


Fig. 2: Description diagram of the Dolby AC-4 system, as specified in ETSI TS 103 190-2.

https://set.org.br/ijbe doi: 10.18580/setijbe.2023.2 Web Link: https://dx.doi.org/10.18580/setijbe.2023.2

C. *ISO/IEC MPEG-H 3D Audio system*

The MPEG-H audio system represents an advanced paradigm in NGA systems. This innovation is formalized as an open international standard under ISO/IEC 23008-3, also known as MPEG-H 3D Audio [11][12]. One of the standout features of this audio system is its declared ability to heighten realism, allowing sound to come not only from the sides but also from above and below the listener. This multi-dimensional experience is further enriched by interactivity features that enable viewers to customize their auditory experience by choosing from various predefined audio presentations—referred to as Presets—or making manual adjustments to audio elements.

The interactive potential is particularly evident in the system's integrated renderer and the advanced management of dynamic range and volume, which optimize content playback according to the capabilities of the playback device. This facilitates seamless audio content delivery across a variety of devices, ranging from headphones to high-quality speaker systems, making it a versatile tool in content creation.

The roots of MPEG-H trace back to the Moving Pictures Experts Group (MPEG), a joint initiative of the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), also responsible for the introduction of the popular MP3, DVD and MP4 multimedia formats. This audio compression standard evolved through a competitive and collaborative process involving experts in audio coding technology. MPEG-H audio encapsulates an intricate combination of highly efficient encoding technologies, the ability to represent audio in three different formats – channel-based, object-based, and scene-based – and advanced volume and dynamic range control (DRC).

Another notable element in the MPEG-H audio system is its classification into profiles and levels. Profiles are essentially a subset of available tools tailored for specific applications. For instance, the High profile incorporates all features and is essentially a theoretical construct. Low complexity and Baseline profiles are more targeted, with the former including additional encoding tools for specific applications such as Virtual Reality (VR) or Augmented Reality (AR). Levels introduce additional parameters that allow for finer tuning of these tools.

MPEG-H has gained international acceptance and has been included in many standards, such as ATSC 3.0, TTA in South Korea, and SBTVD in Brazil, as well as in 3GPP for 360° video streaming services over 5G (5th generation of mobile network).

The MPEG-H 3D Audio system offers an elaborate combination of functions, as illustrated in Fig. 3, whose features are determined by the profiles and levels used. In addition to being available on digital TV receivers, MPEG-H decoding and rendering is currently available in a variety of equipment, including AVR's and soundbar systems.

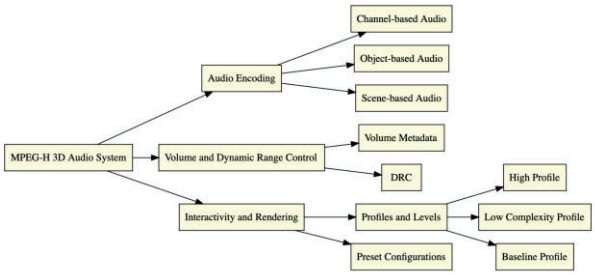


Fig. 3: Description diagram of the MPEG-H 3D Audio system, in accordance with the international ISO/IEC 23008-3 standard.

IV. AUDIO TEST CONTENT

The mandatory tests used audio test content items provided by four different sources, and included audio (wav) files, video (mov) files, and metadata (xml) files describing the audio program organization within the (sound items) payload. After verifying the provided material, 24 (twenty four) audio test content items were validated to be used in the test cases.

The audio test content items consisted of a rich set of types covering several channel setup organizations, program combinations, audio material in different languages, audio-description and emergency warning information. Twelve (12) program file types were defined in the CfP, with different program structures, number of channels, sound content, and bitrates.

In this section we present the 12 test content types used, describing the kind of content of specific items (e.g. what was in the scene, types of sound included) and describing the file type organization in terms of their ADM (Audio Definition Model) structure.

A. *An overview of the Audio Definition Model (ADM)*

The ITU Audio Definition Model (ADM) is a more recent standard proposed in 2019 for audio (scene and file) description in the Recommendation ITU-R BS.2076 [13]. The scheme has been connected to so-called "Next Generation Audio" (NGA) and found support in the broadcast community. In its construct, the audio elements are grouped within the file according to a hierarchy of audioProgrammes, audioContents and audioObjects, as shown in Fig. 4.

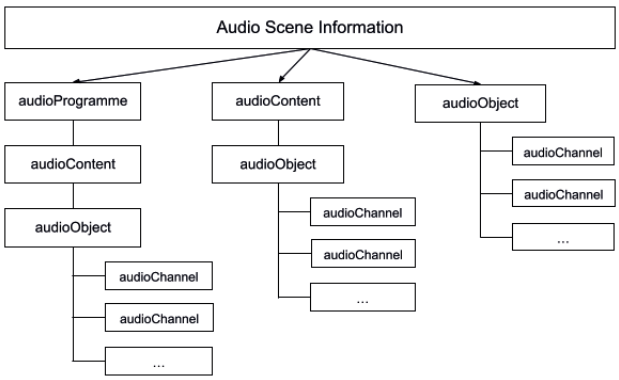


Fig. 4: ADM file structure, according to ITU-R BS.2076.

Audio programmes and audio contents have attributes such as name and language. Audio programmes bring all the audio contents together: they combine all the contents to make the

https://set.org.br/ijbe doi: 10.18580/setijbe.2023.2 Web Link: https://dx.doi.org/10.18580/setijbe.2023.2

complete ‘mix’.

An audioProgramme may contain, for instance, an audioContent for ‘narrator’ and another one for ‘background music’. Or, in another example, an audioProgramme may contain an audioContent for English speakers, called ‘dialogue-en’, together with a ‘backgroundMusic’ content, and another audioProgramme may be prepared for Portuguese speakers, which contains a ‘dialogue-port’ audioContent and the same ‘backgroundMusic’.

The objects are effectively sound source elements, they will have for example attributes such as azimuth, elevation and distance to describe the location of the sound in the scene.

The ADM model is divided into two sections: the content part, and the format part. The content part describes what is contained in the audio, describing things such as the language of any dialogue, the loudness, etc. The format part describes the technical nature of the audio (e.g. the formats of the tracks and/or streams) so it can be decoded/rendered correctly.

Several types of audio are possible, for example: a conventional track (e.g. front-left track); a HOA component; a group of channels, and so on. The type of audio stream will define which channels are inside. There may be audio channels associated to DirectSpeakers (which will then associate a specific speakerLabel), or to a HOA pack, a Matrix, or a Binaural set [13].

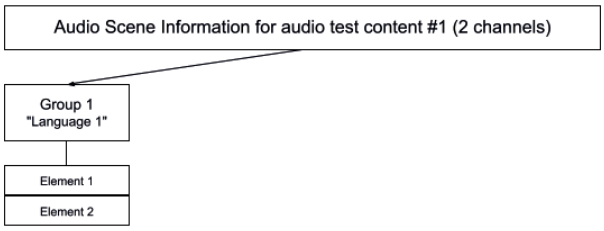
B. Tested audio content types

The following paragraphs present the 12 file types used in the tests, with examples of specific validated items of each type. The presentation of each file type begins by numbering the programs/contents included, with a brief description of their content, and information on the number of channels and bitrate of the file.

The rationale behind the segregation and/or grouping of elements throughout distinct programs, in different multichannel sets, is in facilitating their selection and access for experience personalization. It should be noted that Test Cases therefore derive from what each content organization can offer, and the content selection and personalization that might be possible, with each type, depends on that organization

1. *1: Stereo Mix (Language 1) - (2 Channels / 48kbps)*

This file type has the following ADM structure:

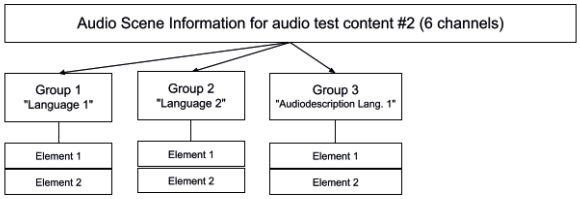


Three specific test items were used:

- 1.1 Aphorism on nature. Content: Nature documentary featuring a Portuguese narrator discussing the importance of nature. It includes ambient sounds (e.g. running water, footsteps, birds, bats) and background music.
- 1.2 Mountain bike. Content: GoPro-style biking trail recording, capturing the biker's breathing, the wheel against the ground, and wind sounds.
- 1.3 Trains passing by. Content: Urban subway footage with varying low and high sound levels.

2. *1: Stereo mix (Language 1) + 2: Stereo mix (Language 2) + 3: Stereo Audio Description (Language 1) - (6 channels / 144kbps)*

This file type has the following ADM structure:

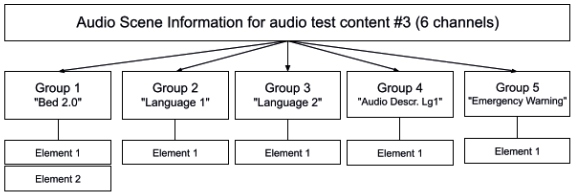


Two specific test items were used:

- 2.1 Aphorism on nature. Content: Two stereo mix contents with different narrators (one in Portuguese, one in English) backed by environment sounds (steps in a cave, trees being hit by the wind and a torch) and classic music playing as background soundtrack. One stereo audio description content (in Portuguese).
- 2.2 Phoenix - German & French. Content: European post-war movie with two stereo mixes (French and German dialogue), and background natural sounds like birds chirping and gravel footsteps.

3. *1: Channel Bed 2.0 + 2: Language 1 Mono + 3: Language 2 Mono + 4: Audio Description Mono (Language 1) + 5: Emergency Warning Information Mono - (6 channels / 192kbps)*

This file type has the following ADM structure:



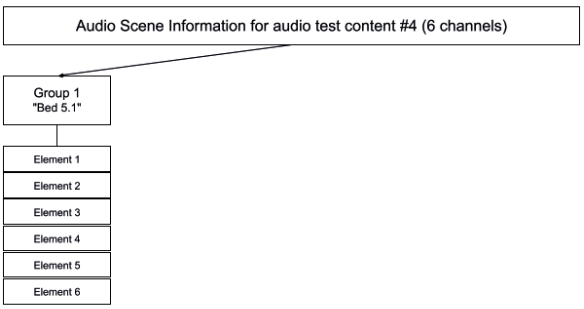
Two specific test items were used:

- 3.1 Aphorism on nature. Content: Similar to 2.1 with nature documentary content and three narrators.
- 3.2 4ever. Content: Short video clip for a French television company, featuring a city bell, classical music, and multiple languages.

4. *Channel Bed 5.1 - (6 channels / 144kbps)*

This file type has the following ADM structure:

https://set.org.br/ijbe doi: 10.18580/setijbe.2023.2 Web Link: https://dx.doi.org/10.18580/setijbe.2023.2

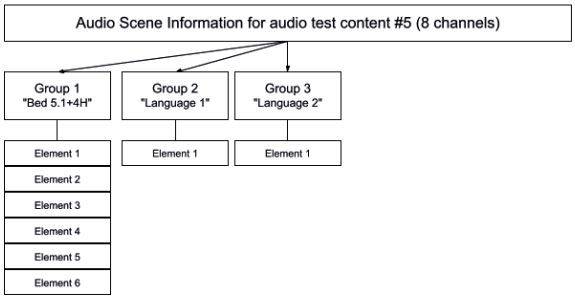


Three specific test items were used:

- 4.1 Aphorism on nature. Content: Nature documentary with Portuguese narration. Features similar ambient sounds as prior examples.
- 4.2 Mountain bike. Content: GoPro-style biking trail recording, similar to content 1.2..
- 4.3 Record's report. Content: Five-minute news report, in Portuguese with background music.

5. *1: Channel Bed 5.1 + 2: Language 1 Mono + 3: Language 2 Mono - (8 channels / 240kbps)*

This file type has the following ADM structure:

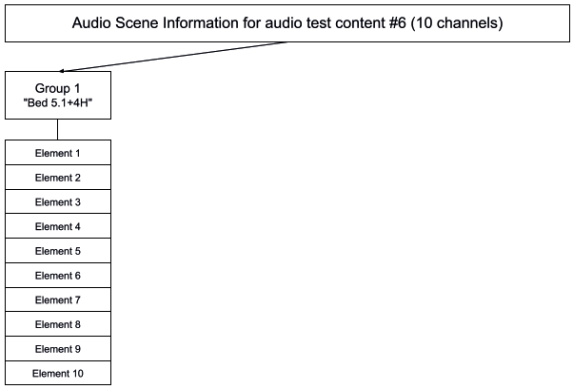


Two specific test items were used:

- 5.1 Aphorism on nature. Content: Nature documentary with dual narration in Portuguese and English translation, featuring similar ambient sounds.
- 5.2 One day in berlin. Content: Daily life in the city during summer with Portuguese and German narrations.

6. *Channel Bed 5.1+4H - (10 channels / 256kbps)*

This file type has the following ADM structure:

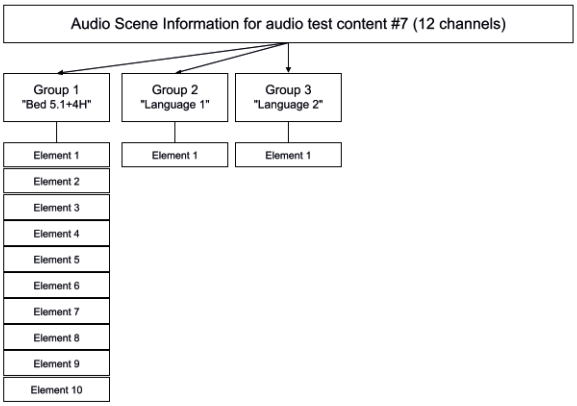


Three specific test items were used:

- 6.1 Aphorism on nature. Content: Features ambient nature sounds without narration.
- 6.2 Mountain bike. Content: GoPro-style biking trail recording.
- 6.3 Eurovision Sweden. Content: Audio from the Eurovision Israel 2019 event featuring piano, opening music, a singer, and a crowd.

7. *1: Channel Bed 5.1+4H + 2: Language 1 Mono + 3: Language 2 Mono - (12 channels / 352kbps)*

This file type has the following ADM structure:

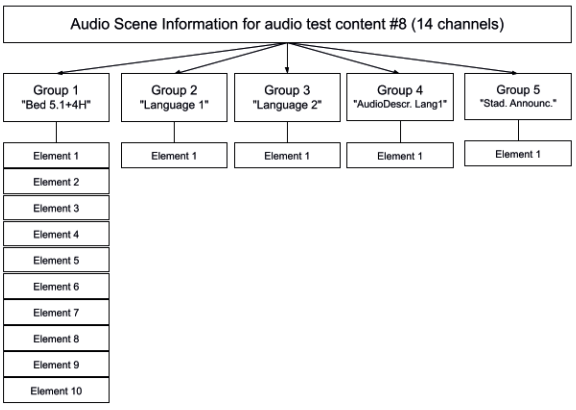


Two specific test items were used:

- 7.1 Aphorism on nature. Content: Nature documentary with ambient sounds.
- 7.2 Le Mans Astray. Content: First half features a high-speed car and traffic sounds; the second half transitions to a nature documentary.

8. *1: Channel Bed 5.1+4H + 2: Language 1 Mono + 3: Language 2 Mono + 4: Mix Stereo (Language 1) + 5: Mix Mono (Language 2) - (15 channels / 448kbps)*

This file type has the following ADM structure:



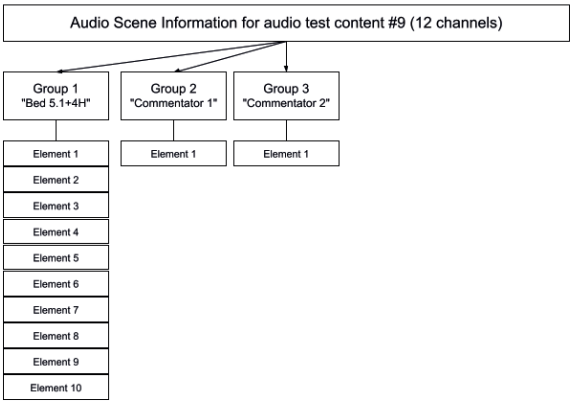
Two specific test items were used:

- 8.1 Aphorism on nature. Nature documentary with ambient sounds.
- 8.2 European championship Berlin/Glasgow 2018. Audio from a hurdle race championship featuring opening music, narration, and crowd sounds.

https://set.org.br/ijbe doi: 10.18580/setijbe.2023.2 Web Link: https://dx.doi.org/10.18580/setijbe.2023.2

9. *1: Channel Bed 5.1+4H + 2: Commentator 1 Mono + 3: Commentator 2 Mono - (12 channels / 352kbps)*

This file type has the following ADM structure:

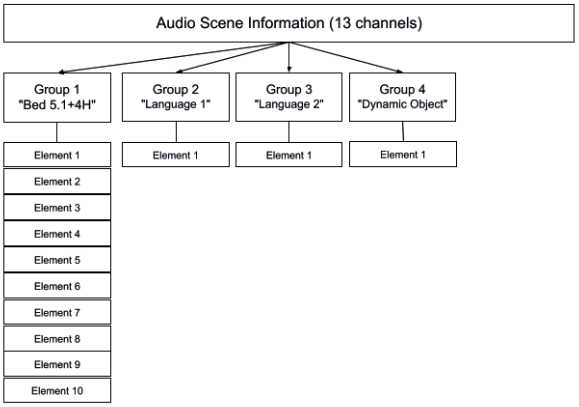


Two specific test items were used:

- 9.1 Rio de Janeiro carnival. Content: Samba school performance featuring two singers and instruments.
- 9.2 Carnaval 2020 Rio de Janeiro. Content: Audio from a Carnival Parade in 2020, featuring samba music.

10. *1: Channel Bed 5.1+4H + 2: Language 1 Mono + 3: Language 2 Mono + 4: Dynamic object Mono - (13 channels / 400kbps)*

This file type has the following ADM structure:

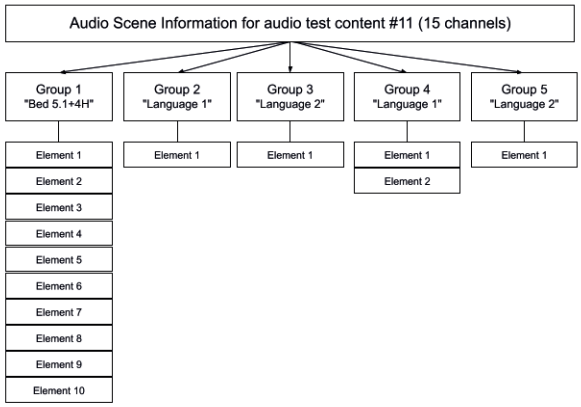


One specific test item was used:

- 10.1 Le Mans Astray (5.1+4H). First half features a high-speed car; the second half is a nature documentary.

11. *1: Channel Bed 5.1+4H + 2: Language 1 Mono + 3: Language 2 Mono + 4: Mix Stereo (Language 1) + 5: Mix Mono (Language 2) - (15 channels / 448kbps)*

This file type has the following ADM structure:



Two specific test items were used:

- 11.1 Aphorism on nature. Content: Nature documentary with ambient sounds.
- 11.2 European song contest Lisbon 2018. Content: features crowd sounds, music, and special effects.

Fig. 5 illustrates the audio test file 11.1, showing its 15 tracks.

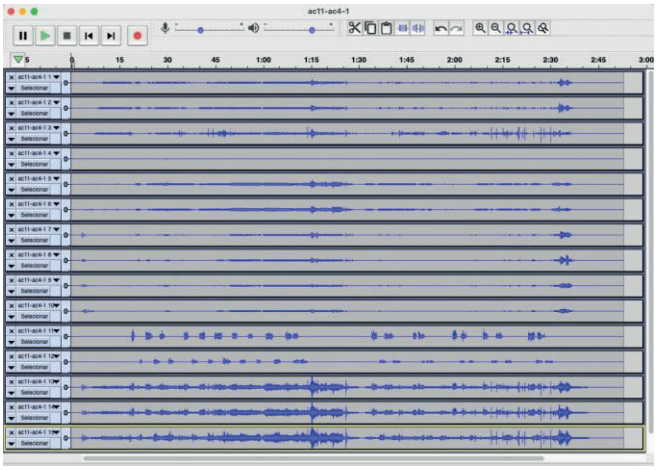
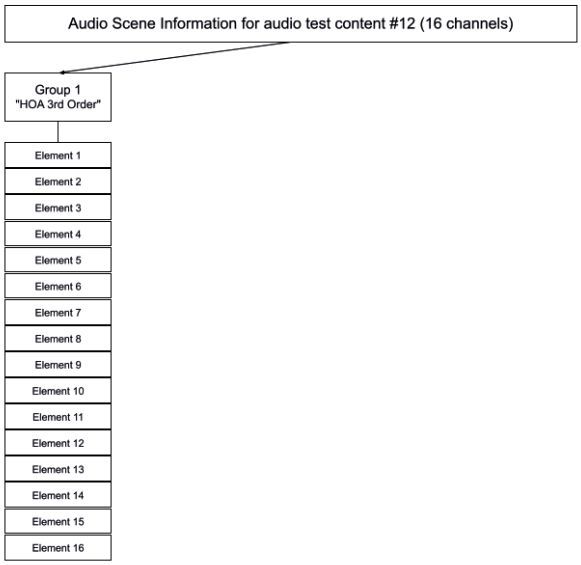


Fig. 5: Example of a file of type #11 with 15 channels.

12. *1: HOA (3rd order) - (16 channels / 320kbps)*

This file type has the following ADM structure:



https://set.org.br/ijbe doi: 10.18580/setijbe.2023.2 Web Link: https://dx.doi.org/10.18580/setijbe.2023.2

One specific test item was used:

- 12.1 Clouds of franconia. Content: Soundtrack introduction with guitar and drum (HOA formatted with ACN channel ordering, SN3D normalization).

V. Assessment and Testing Methodologies

The testing and evaluation phase 2 lasted six months (from July to December 2021) and was aimed at assessing the candidates' technologies in order to verify their compliance to the specified requirements and evaluating their performance in laboratory tests, conducted with a set of audio test content and conditions as close as possible to expected operation conditions at future broadcasting service.

As specified in the CfP methodology, the first evaluation stage was to analyze each candidate's technology documentation and technical specifications to verify theoretically if they fulfilled the mandatory requirements. All three candidate technologies properly submitted their documentation which were evaluated at this stage, also benefiting from additional documentation provided by standardization institutions and known previous studies [2-6].

Following this stage, the next one was to conduct laboratory tests with physical hardware and software provided by the candidates, in order to assess the actual functioning of the systems and the fulfillment of the requirements under operating conditions. In this stage, only 2 of the candidates submitted equipment and software and took part in the tests.

A. Laboratory setups

In this section we briefly describe the laboratory setups, presenting the studio and test room environments, delivery interfaces and equipment employed. Particularly of interest, is the audition test room prepared for simulating a common domestic audition environment, where, in addition to the TV set loudspeakers, we used an external 5.1+4H loudspeaker setup and a soundbar system to reproduce 2D/3D spatial sound-fields in the room (see Fig. 6). We also tested the immersive and functional capabilities using conventional stereo (2.0) reproduction and binaural setups using earphones, covering all required output layouts.

Fig. 6: Auditory test room and sound systems used in the tests.

Two scenarios for tests were considered, employing two different setup modes, as required by the CfP [1]: (1) real-time encoding/decoding setup, and (2) non-real-time encoding/decoding setup.

The real-time setup emulates a typical broadcast scenario, where the broadcast feed is authored in one location (e.g., event location or studio) and provided over a contribution link to the broadcast center where it is monitored and re-authored if needed, prior to delivery (over the air or over Internet). For this setup we employed two different locations: a simulated studio/broadcast facility, where it was possible to author metadata and prepare the broadcast feed; and the test room, where the end-users could experience the programs using appropriate receivers with decoders and renderer equipment and software. The live delivery between these facilities employed an IP-based link. Fig. 7 illustrates the real-time setup, as specified in the CfP [1].

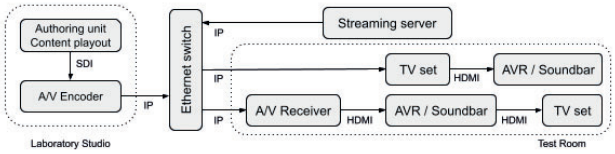


Fig. 7: Studio and Test room system setup for the real-time encoding/delivering/decoding.

The non-real-time setup, the transport layer between the audio encoder and the A/V receiver was specified to use an MP4 container format. All test content should be available as MP4 files stored in the receivers. For an end-user's setup with an A/V receiver, it should use an HDMI interface to feed the external sound system and demonstrate the audio system capabilities. The figure 8 illustrates the non-real time setup, as specified by the CfP.

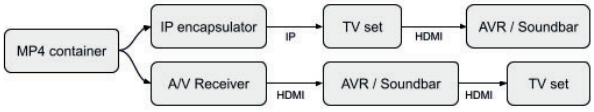


Fig. 8: Test room system setup for the non-real-time encoding/decoding. Sources in MP4 could be delivered to a TV set and external sound systems using IP and HDMI interfaces.

B. Test Cases performed

By conveniently breaking down the 13 test groups into a set of 46 test cases, and considering an average of (at least) 2 audio test content items (of 12 distinct types) used per TC, plus 2 technologies to evaluate, a minimum of 184 rounds of laboratory tests was required. Table I shows the expanded structure of the 46 Test Cases.

https://set.org.br/ijbe doi: 10.18580/setijbe.2023.2 Web Link: https://dx.doi.org/10.18580/setijbe.2023.2

TABLE I – EXPANDED TEST CASES

TC1.1.1 (immersive audio), AC1.1.1, audio 1 (2.0), real-time
TC1.1.2 (immersive audio), AC1.1.2, audio 4 (5.1), real-time
TC1.1.3 (immersive audio), AC1.1.3, audio 6 (5.1+4H), real-time
TC1.1.4 (immersive audio), AC1.2, audio 9, real-time
TC1.2 (immersive audio), AC1.3, HOA, audio 12, non-real-time *
TC2.1 (interactivity and personalization), AC2.1 language sel, audio 3/5/8, real-time
TC2.2 (interactivity and personalization), AC2.1 preselec, audio 8/11, real-time
TC2.3 (interactivity and personalization), AC2.1, coment switch, audio 9, real-time
TC2.4 (interactivity and personalization), AC2.1, labels display, audio 8/9/11, real-time **(AC08)**
TC2.4 (interactivity and personalization), AC2.1, labels display, audio 8/9/11, real-time **(AC09)**
TC2.4 (interactivity and personalization), AC2.1, labels display, audio 8/9/11, real-time **(AC11)**
TC2.5 (interactivity and personalization), AC2.2, loudness int, audio 8/9, real-time **(AC08)**
TC2.5 (interactivity and personalization), AC2.2, loudness int, audio 8/9, real-time **(AC09)**
TC2.6 (interactivity and personalization), AC2.3, object position, audio 8/10, real-time **(AC08)**
TC2.6 (interactivity and personalization), AC2.3, object position, audio 8/10, real-time **(AC10)**
TC2.7 (interactivity and personalization), AC2.4, int on external, audio 8/9/10, real-time **(AC08)**
TC2.7 (interactivity and personalization), AC2.4, int on external, audio 8/9/10, real-time **(AC09)**
TC2.7 (interactivity and personalization), AC2.4, int on external, audio 8/9/10, real-time **(AC10)**
TC3.1 (Audio description), AC3.1/AC3.2, AD in same stream/alternate mix, audio 2, real-time
TC3.2 (Audio description), AC3.3 P1, AD in additional audio, audio 3/8, real-time
TC3.3 (Audio description), AC3.3 P2, AD in addit audio and language, audio 3/8, real-time
TC3.4 (Audio description), AC3.3 P3, AD in addit audio and spatial, audio 3/8, real-time
TC4.1 (Audio emergency), AC4.1, audio 3, real-time
TC5.1 (Flexible audio playback), AC5.1/AC5.2, audio 8/11, real-time **(AC08)**
TC5.1 (Flexible audio playback), AC5.1/AC5.2, audio 8/11, real-time **(AC11)**
TC6.1 (Consistent loudness), AC6.1, loudness norm program, audio 8/11, non-real-time *
TC6.2 (Consistent loudness), AC6.2, loudness norm preselections, audio 8/11, non-real-time *
TC6.3 (Consistent loudness), AC6.2, loudness comp, audio 8/11, non-real-time *
TC7.1 (Seamless config changes), AC7.1, seamless config changes, audio 1/4/6/8, real-time
TC7.2 (Seamless config changes), AC7.2, seaml. playback user interact, audio 8/11, real-time
TC7.3 (Seamless config changes), AC7.3 P1, seaml. playback changes in product., audio 8/9/11, real-time
TC7.4 (Seamless configuration changes), AC7.3 P2, seaml. plbck chngs in prod. w/ feed, audio 8/11, real-time
TC7.5 (Seamless configuration changes), AC7.4 P1, seamless ad-Insertion, audio 1/11, real-time
TC7.6 (Seamless configuration changes), AC7.4 P2, user select persist. after ad-break, audio 1/11, real-time
TC7.7 (Seamless configuration changes), AC7.4 P3, hybrid delivery, audio 3, real-time
TC8 (Audio coding efficiency), AC8.1 kbps @ MOS 4 / MUSHRA > 80
TC9.1.1 (Latency), AC9.1 during TC1.1, AC.9.2 latency (ms)
TC9.2 (Latency), AC9.1 during TC1.1, AC.9.2 latency (ms)
TC10 (A/V Sync), AC10.1, adequate A/V sync, audio 3/5/7/9, real-time **(AC03)**
TC10 (A/V Sync), AC10.1, adequate A/V sync, audio 3/5/7/9, real-time **(AC05)**
TC10 (A/V Sync), AC10.1, adequate A/V sync, audio 3/5/7/9, real-time **(AC07)**
TC10 (A/V Sync), AC10.1, adequate A/V sync, audio 3/5/7/9, real-time **(AC09)**
TC11.1 (New immersive audio services), AC11.1, VR / AR / XR / 3DoF / 6DoF
TC12.1 (Interoperability), AC12.1 and AC13.1 w/ different distrib. plats, audio 8, real-time
TC13 (Scalability/Extensibility), AC13.1/AC13.2, during TC12

The actual number of test rounds executed, however, was nearly 10 times as much, considering that there were several audio test files of each type; rounds were done by different testers/evaluators; several real-time tests required changing attributes' values in real-time authoring; and that tests were repeated in order to verify the result consistency for a range of attributes' values and reproducibility of the operations.

For each test case, the respective requirements were evaluated and the results of the tests were to produce a classification in three possible outputs:

- "Fulfilled", if the proponent's system was fully able to satisfy the requirements;
- "Partially Fulfilled", if the proponent's system fails to satisfy part of the requirements;
- "Not Fulfilled", if the proponent's system fails to satisfy the requirements.

The next paragraphs describe how the TCs sessions were organized and conducted for each test group.

Test 1 (Immersive audio):

This test was divided in two parts, one aimed at channel-based setups (three TCs to test the system's ability to present audio in the specified channel modes 2.0, 5.1, and 5.1+4H channels and target bitrates during real-time set up) and one aimed at a scene-based setup (one TC to demonstrate the ability to present scene-based HOA content in the expected target bitrate of 20 kbps per HOA-channel, therefore a total HOA bitrate equal to 320 kbps) through non-real-time setup.

In both parts, the videos (with the audio content) were played continuously and they were heard through the AVR system and soundbar to evaluate the overall 3D experience.

Test 2 (Interactivity and personalization):

This test was divided into seven TCs. The first one examined the ability to select between multiple audio languages based on user interaction or automatic language selection through real-time set up. The system should allow authoring the metadata in the studio, enabling the desired personalization options. Then, during playback, these options should be displayed on the receiver side. This test also investigated the capability of the audio system in maintaining the chosen language when restarting the receiver.

The second TC evaluated the capability of the audio system to display all preselections authored in production; the capability of the end-user to manually switch between different preselections during live playback; and the capability of the audio system to correctly render the preselections on the receiver side.

The third TC examined the ability of end-users to switch between multiple available commentators, and evaluated the display on the receiver side of all available commentators authored in the production through real-time set up.

The fourth TC consisted in real-time authoring metadata at the studio for several audio elements (e.g., Dialogs, Commentators, Stadium Announcers) and several preselections (e.g., Main mix, Dialog+, Stadium). The authored signal was then encoded and in the test room it was evaluated the capability to display the textual labels for all preselections and audio objects allowing user interactivity as well as the capability of the audio system to correctly render the preselections

The fifth TC consisted of demonstrating the ability to interact with any audio object and adjust the level through real-time set up. It evaluated if the user was able to increase or decrease the object level (relative to the background) inside the range specified by the broadcaster during live playback.

The sixth TC evaluated if the end-user was able to move audio elements inside an area (space) specified by the broadcaster in real-time setup.

The seventh TC tested the interactivity options when using external sound reproduction devices through real-time set up (e.g., Soundbar/AVR). This TC tested if the end-user could interact with the audio scene (e.g., change an object level or position, change the preset) during live playback, and evaluated the immersive experience reproduced on the external sound device (e.g., objects moved in the 3D space should be perceived at specific positions according to the user interaction).

Test 3 (Audio description):

This test was divided into four TCs. The first two consisted of demonstrating the system's ability to display the available audio description in multiple languages (as authored in production) and to enable and to switch between audio description elements during live playback through real-time set up.

The third TC examined the ability to enable/disable audio description in multiple languages, as authored in production and the capability of the audio system to start the playback of the audio description in the authored stream.

The fourth TC evaluated the system's ability to enable/disable audio description and spatially separate the

https://set.org.br/ijbe doi: 10.18580/setijbe.2023.2 Web Link: https://dx.doi.org/10.18580/setijbe.2023.2

main dialog and the audio description through real-time setup. It evaluated the capability of the audio system to correctly reproduce the main dialog and the audio description at the desired locations in each preselection during live playback according to the metadata authored in production.

Test 4 (Audio emergency warning information):

This test consisted of demonstrating audio emergency warning information presentation through real-time set up. During this test, it was verified the continuous playback of the content before, during, and after its delivery, and the capability of the audio system to signal the Emergency Information in the authoring system and the flexibility to control it (i.e., if the audio object should be active in all preselections or a dedicated preselection, should mute the main dialog or playback over the main dialog).

Test 5 (Flexible audio playback configuration):

This test evaluated the system's ability to present the same content on TV loudspeakers, soundbar, AVR connected to a 5.1 and 5.1+4H loudspeaker setups as well on headphones in real-time set up.

Test 6 (Consistent loudness):

This test was divided into three parts. The first one evaluated the ability to achieve the target loudness level across multiple programs through non-real-time set up. The audio content was decoded using the proponent software audio decoder to three different target loudness levels: -31, -24, and -16 LKFS and the loudness consistency across multiple test items was evaluated. The program loudness was measured according to ITU-R BS.1770-4 with a tolerance of +/-3 dB, using the FFMPEG tool. From the FFMPEG tool output results, the loudness measurement was the value of the parameter labeled as "Integrated Loudness" (I), in LUFS units (equal to LKFS as defined in Rec. ITU-R BS.1770).

The second part examined the system's ability to preserve the target loudness level across multiple preselections through non-real-time set up. The test demanded re-authoring the content using the proponent authoring tool and adding more preselections. After that, the audio was encoded and decoded to the target loudness levels: -31, -24, and -16 LKFS using the proponent encoder and the loudness consistency was measured according to ITU-R BS.1770-4 with a tolerance of +/-3 dB, using FFMPEG tool.

The third part evaluated the preservation of the target loudness level after user interaction. The test demanded re-authoring the content using the proponent authoring tool and changing the minimum and maximum gain interactivity options for several dialog objects to at least +/- 10 dB. After that, the audio content was encoded and MP4 files mixing audio and video streams were created. The MP4 files were playbacked using the proponent video player and the increase of level of dialog objects was evaluated as well the overall perceived loudness before and after the user interaction.

Test 7 (Seamless configuration changes and Audio/Video alignment):

This test was divided into seven TCs. The first one tested the system's ability to seamless playback content during configuration changes through real-time set up. The expected result was a continuous and seamless playback during all configuration changes.

The second test examined the seamless playback content during user interaction through real-time set up. It evaluated the occurrence of audio drop-outs or glitches in real-time playback during changes between different audio languages or preselections, increasing or decreasing the level of various audio objects.

The third test consisted of demonstrating the system's ability to seamlessly playback content during changes in production during a live broadcast through real-time set up. All typical changes in a live broadcast were tested.

The fourth test examined the seamless playback during changes in production in a live broadcast scenario through real-time set up. For this test, the pre-recorded output of the proponent's authoring system was used as production content in the broadcaster studio. Using the authoring system in the studio, the metadata was re-authored, and potential errors in the original authoring were treated. The ability of enabling/disabling interactivity for one and more preselections and one and more audio objects were evaluated as well as the interactivity options (min/max gain and position values) and the capability to seamlessly playback the content and correctly display the user interaction options while making the changes in live production.

The fifth test analyzed the system's performance during advertisement insertion. Using a clean SDI switch, two different contests were switch in order to evaluate the capability of the audio system to display on the receiver side the various interactivity options corresponding to each configuration and seamlessly update the user interface at each ad-insertion and evaluate continuously and seamlessly playback the content during the ad-insertion.

The sixth test examined the system's ability to preserve the user interaction settings after the ad-break. After re-authoring metadata in the authoring system, a clean SDI switch was used to switch between contents. After one minute, the contents were switched and verified the ability to preserve the user selections after the ad-break.

The seventh test consisted of demonstrating the system's ability to synchronize during the replacement of the main soundtrack delivered via broadcast for an alternative audio signal delivered via broadband through real-time set up. In the beginning of the test, the content was encoded offline and prepared as multiple ISOBMFF streams ready for DASH streaming from the Streaming Server containing a Channel Bed 2.0 program and different objects. In the test room, it was evaluated the ability: to synchronize the multiple streams received live; to display the options available (e.g. the playing starts always with stream 1 but options from stream 2 and 3 shall be displayed); to switch to additional languages coming from IP chain 2 (Streaming Server) and to switch back to the main language if the IP chain 2 is disconnected (e.g., stopped from the Streaming Server).

Test 8 (Audio coding efficiency):

This test examined the proponent's documentation provided on the Quality Assessment Reports, without laboratory tests.

https://set.org.br/ijbe doi: 10.18580/setijbe.2023.2 Web Link: https://dx.doi.org/10.18580/setijbe.2023.2

Test 9 (Audio End to end latency):

This test was performed during the execution of Test Case TC1.1, and its requirement AC9.1 was classified correlated to the requirements AC1.1.1, AC1.1.2, and AC1.1.3.

Test 10 (Audio/Video synchronization)

This test examined the A/V synchronization through real-time set up. It evaluated the downmix and rendering when the content was played back over a 5.1 setup and a stereo setup and the overall 3D experience when the content was played back over an external sound system.

Test 11 (New immersive audio services)

This test examined the ability of the system to perform playback of audio in VR / AR / XR applications, with 3DoF. or 6DoF. It evaluated: features of the codec; the readiness for real-time coding/decoding; the readiness of delivery of the format; and how the application works and could manipulate the audio codec stream. The test verified the demonstration of 3D audio VR and 3DoF support (none of the proponents included a 6DoF example) and was partially fulfilled, as the provided schemes did not permit the evaluation of the capability of real-time encoding and delivering/decoding in a streaming/broadcasting fashion. For the time being, it demonstrated, however, the capability of offline encoding and decoding on a cell phone application.

Test 12 (Interoperability with different distribution platforms)

This test analyzed the system's ability to send multiple audio contents over two or more communications channels through real-time set up. It needed multiple ISOBMFF streams ready for DASH streaming from the Streaming Server to be created in order to verify the abilities to: synchronize multiple streams received live; display the options available; switch to additional languages coming from IP chain 2 (Streaming Server) and the ability to switch back to the main language if the IP chain 2 is disconnected (e.g., stopped from the Streaming Server).

Test 13 (Audio scalability and extensibility):

This test was performed during the execution of Test Case 12.1, verifying its requirement AC13.1. Concerning the requirement AC13.2, it was not analyzed with a feature test but through the analysis of the proponent's documentation provided in the Document Analysis phase.

VI. Results and discussions

The results obtained in the conducted test cases provided a rich showcase of the actual capabilities delivered by the current versions of the candidate technologies and their potential to deliver a set of new services for audio presentation in a variety of program types.

A final and comprehensive report of the results – indicating the fulfillment classification for each requirement and for all tested technologies – was prepared by our team, and then published publicly by the Forum SBTVD [14].

This section presents an overview of the audio personalization and immersion features that were evaluated and would be possible in the next generation of digital TV; communicates the decisions made by the technical Forum regarding the official adoption of technologies; and brings some discussions on present limitations, future prospects, aesthetics challenges, technical issues, and ongoing activities towards the implementation of the new TV3.0.

A. Audio personalization and immersion features

The personalization of audio program presentation – in terms of enabling the user to control sound selection, positioning, distance, loudness, equalization and effects – has being thought for television markets even before the first generation of DTV in 2007, when ISO/IEC issued the call for technology proposals for MPEG SAOC (Spatial Audio Object Coding). The concept of authoring presets for final users' selection and personalization, for instance, had already been addressed by Lee et al in 2006, at that time using the MPEG-4 BIFS (BInary Format for Scenes) 3D audio scene description tool in the MPEG-4 Systems [15].

The ISO/IEC call for proposals for developing the MPEG-H 3D Audio standard was issued 10 years ago (Jan. 2013), targeting application scenarios such as personal home theaters, handheld smartphones, 3D video, telepresence rooms, cloud-based gaming, and accurate sonic localization for audio-only program listening (e.g. virtual concert halls), seen as potential markets for the future. For broadcast application, a special attention was given to cover demands for transmitting several groups of audio content and making them accessible independently from each other, such as multi-language announcer and commentator voices, environment sounds, and different sound mix presets.

Sport events and karaoke were typical program examples, where the user should be able to select and to enable or not several sounds out of a set of options (e.g. turning on/off the stadium announcements and audience sound; selecting the preferred language and commentator; selecting specific instruments, vocal option, or presets with different instrumentation) and to adjust localization and loudness levels for the sounds. Keeping timbre, sound localization and envelopment were also requirements for MPEG-H, as well as the capability of downmixing the number of channels and rendering the spatial sound to a lower hierarchy of loudspeaker setup, with 10.1 (ten full channels plus a low frequency one) or 8.1 channels.

The TV3.0 personalization capabilities will include the selection and activation of available sound programs and the control of loudness (prominence) levels. End-users shall be able to switch between components (e.g. alternative mix substreams or audio objects), to adjust their loudness, and to enable services such as audio description and emergency warning information.

https://set.org.br/ijbe doi: 10.18580/setijbe.2023.2 Web Link: https://dx.doi.org/10.18580/setijbe.2023.2

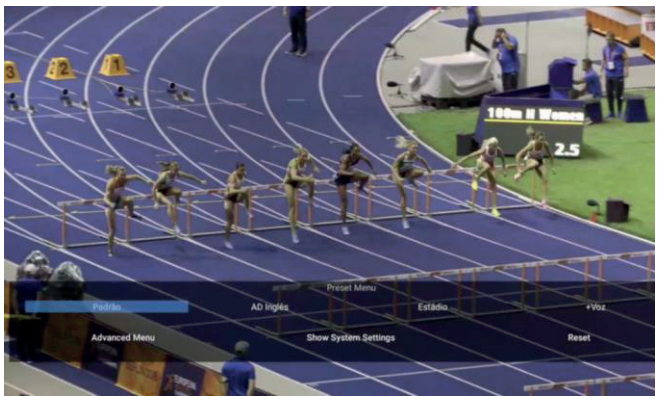


Fig. 9: Menu for selection of programs/contents and personalization of sound system settings.

Sound personalization possibilities illustrated in Fig. 9 include:

- change of audio program (from the available options delivered)
- change language of announcer / commentator
- activation of the audio description feature (or not)
- change playback configuration (e.g. output system such as soundbar, external 5.1+4H system or embedded stereo loudspeakers of the TV set)
- choose which sound elements (objects) available in the programs the end-user wants (or not) to listen
- modify the presence level of specific sounds (loudness adjustment), e.g. turning up speech or lowering the background environment sound
- reset for the default configurations set by the broadcaster

Considering the immersion features, the next generation will enable a flexible selection and seamless switch between available audio playback configurations, such as 2.0, 5.1, and 5.1+4H channel-based layouts (home-theaters) or binaural renderization (for earphones). It shall also permit at the end-user device to control the sound panorama and to adjust sound object position in the listening area, according to available broadcaster's settings and within the range conceived by the program content producer.

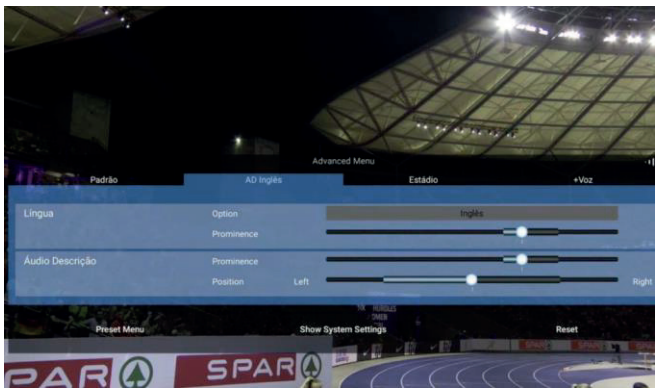


Fig. 10: Menu for audio description and language selection, loudness (prominence) and sound spatial positioning.

Immersion personalization possibilities illustrated in Fig. 10 include:

- define the positioning of specific sounds in the selected program around the user (panorama setup: left-right within the range set by the broadcaster)
- define the prominence of the specific selected sounds
- reset the configurations back to the broadcaster's default settings.

The Test Cases verified some possible immersive personalization capabilities, such as the end-user ability to move audio elements at the end-user's device inside an area specified by the broadcaster (e.g., min/max position interactivity values), and have shown that this operable area may be set differently for each object. However, the present end-user's interfaces for changing the spatial scene were restricted to a simple panorama (left-right) adjustment, controlled with the remote control keys. Although not available in the actual tested interactive interfaces, it would be possible to change the sound object positioning not only on the planar situation (2D) but also on its elevation (for 3D playback).

Considering the choices for loudspeaker systems and possible enhancements in immersive experience, it remains open space for further investigations. The determination of the immersion level delivered by a certain system could be a valuable tool both for manufacturers – in calibrating and assessing the level of perceived spatial immersion achieved, and for final users – in optimizing their listening setups. Faria has proposed in 2005 a discrete 6-level scale for immersion degree assessment, starting from 0 (no spatial information) up to 6 (3D coherent spatial impression) [16], but no practical usage of this or similar scales has been tracked so far.

Considering the choices for multichannel playback layouts, we comment on some interesting points. Television sets shall continue offering the 2.0 (stereo) output, but for taking advantage of 2D and 3D experience the end-user must use an external multi-loudspeaker sound system, which will receive the audio stream from the DTV receptor typically using a HDMI cable, and then will distribute the output signals for each of its loudspeakers. An AVR unit might be one choice, in which case it should be considered if it handles MPEG-H decoding and rendering functions in an integrated manner (which has been the tested situation using a commercial native MPEG-H AV receiver).

Alternatively, some DTV devices might include native multichannel loudspeakers support. A flat HDTV television set prototype with over a hundred micro-loudspeakers embedded onto its frames has been tested by NHK in 2012, aiming to enhance frontal spatialization using a wave-field synthesis scheme [17]. However, such types of loudspeaker arrays have not shown popularity like newer soundbar systems – a single multichannel audio receiver with an (embedded) array of loudspeakers. A study from Ando (2011) has pointed out that a minimum of 8 channels would be required for providing sound localization and envelopment to satisfactorily recreate a spatial impression at home using loudspeakers [18]. Employing layouts with loudspeaker at 3 height levels, that outcome converges to support configurations beyond the regular planar ITU 5.1 surround

https://set.org.br/ijbe doi: 10.18580/setijbe.2023.2 Web Link: https://dx.doi.org/10.18580/setijbe.2023.2

scheme, such as unusual 8.1, 9.1, and 10.1 layouts (which mixes loudspeakers in the middle, upper, lower and top positions in different azimuths and elevations) and the more recent 5.1+4H (10 loudspeakers) proposed setup.

Notwithstanding, the popularity of multiple loudspeakers in home environments is not borne out by history, being binaural/stereo output or using a single speaker array the most likely successful choices for 3D audio consumption. Even so, current soundbars' ability to provide height effects are still limited. A sense of envelopment from above is noticeable, but the accuracy of depth and elevation positioning is less clear. It is worth recalling that the studies by Ando [18] evaluated loudspeaker layouts with 3 height levels, a requirement not explored in the present tests. Finally, improvements are expected in the definition of the elevation dimension for speaker arrays.

B. Official deliberations of technologies to adopt

Although both final tested candidate technologies have demonstrated fair sound quality – and less complexity/greater ease of use in particular test cases and functionalities – after weighting the complete fulfillment of all CfP requirements, the tests results were evaluated by the Forum SBTVD and a final deliberation pointed towards the formal adoption of the international open-standard MPEG-H 3D Audio for the future TV3.0 Over-the-Air (OTA) and Internet distribution (OTT) – ISO/IEC Standard 23008-3, now in its 3rd. edition, revised in 2022 [11]. It also decided to maintain the E-AC3 and AAC (MPEG-4 Advanced Audio Coding) audio formats currently supported in TV2.5 for the distribution of alternative content over the Internet, including optional Dolby AC-4 support.

Most use scenarios highlight the capability for the program producers to assemble different setups of audio presentations (which could be seen as "presets") and to deliver them in different sets of channels and programs. The systems take the advantage of being compliant to NGA program description schemes, such as with the xml-based ITU Audio Definition Model (ADM) metadata format [13], which conveys programs descriptions and how they are organized and could be selected and presented at the receiver (player) side.

C. Discussions on challenges, prospects on aesthetics and technical issues

The TV3.0 system opens and welcomes a broad avenue for novel aesthetics and ways to explore the system's resources to code and deliver interactive and personalizable content. There are challenges, however, on the current content production chain, not considering yet the benefits and impacts on the production and post-production methods, and the unexplored creative potential for artists in inventing new forms of programs, both to take advantage of the new features offered by the technology and to engage the users in active participation in the program experience. Some system's capabilities, such as the "emergency warning delivery" could be, for example, explored by TV3.0 application coding experts to incorporate new features into the future specification.

The ISO/IEC 23090-4 MPEG-I Immersive Audio forthcoming standard – currently under development – is expected (possibly in a shorter time-to-market than usual) to further expand the possibilities for immersive program creation, delivery, consumption and personalization, especially for Virtual and Augmented Reality audio presentations. MPEG-I will permit 6DoF experience (6 Degrees of Freedom) i.e. translation {x,y,z} and rotation {yaw,pitch,roll} directed by the physical movement of the user in the space. Several enhancements in terms of immersive experience are expected, such as better directivity perception and positioning of sound sources, ambience and reverberation, and new rendering technologies.

The new technology also brings novelties in the transport layer, employing novel and alternative methods to the conventional MPEG TS (Transport Stream) container format used for transmission and storage of media. MPEG-H content can be packetized in MPEG-H Audio Stream (MHAS) format, or encapsulated into ISOBMFF files (which we have tested in the present study using the non-real-time encoding/decoding setup). Besides MPEG TS, delivery could be implemented over the ROUTE/DASH protocol (Real-Time Object Delivery over Unidirectional Transport/ Dynamic Adaptive Streaming over HTTP) which we have tested in the present study using the real-time encoding/ decoding setup.

Finally, as incorporated in the mandatory tested requirements, our studies also unveiled the possibilities for binaural rendering for Virtual and Augmented Reality applications, and scene-based audio formats, such as HOA, which are establishing new landmarks for film and communications in the 21st century.

Scene-based streams carry full sound field representations, agnostic of the final required loudspeaker feeds to recreate them at playback, but are sensitive to channel bandwidth and compressive artifacts. MPEG-H (and future MPEG-I) use several techniques for spatial compression of HOA signals. Bleidt et al (2017) states that broadcast quality (scoring a minimum of 80 MUSHRA points) of HOA content up to the sixth order (which sums up to 49 HOA coefficients) is achieved at bitrates as low as 300 kb/s and transparent quality transmission (minimum of 90 MUSHRA points) are achieved at 500 kb/s independent of the HOA order [19]. In a standard SDI environment, however, limited to 16 channels, MPEG-H is capable of delivering a full set of 15 audio channels plus 1 control channel (for metadata), which means a (pure) HOA signal up to the third order (16 channels).

VII. Final Remarks

The effectiveness of personalizing a sound scene is an enterprise dependent not only on the technology of coding, transmitting and rendering, but also on the content motivation and interactivity aspects. One has to observe that while current technologies offer the possibility of defining specific loudness and 3D positioning for a sound in a scene, the final user or spectator ability to control these attributes are limited by the user interface, which can, at last, encourage or demotivate spectators in the task of interacting with the scene or setting it up. Intuitive means to select sounds and pinpoint their desired positions in the space by voice command or gestural tracking should be much more taken for granted than operating over a remote control.

https://set.org.br/ijbe doi: 10.18580/setijbe.2023.2 Web Link: https://dx.doi.org/10.18580/setijbe.2023.2

Besides the user interface, one has also to consider the program's ability to entice the spectator to participate in its flow by selecting options and settings. Content providers and producers are ultimately responsible for the creativity and innovation in the next generation of attractive and popular interactions.

These features open a new and vast avenue to be explored by program content producers, artists, and broadcasters, as they can assemble and deliver diverse audio contents in the same stream, for alternative presentation options, and also permit the user to select or not some sound items, languages and accessibility services, to modify their spatial distribution around, and hence to personalize the content and the immersive aspects of the listened experience.

Concerning the MPEG-I forthcoming standard, its compatibility with the MPEG-H architecture, codecs, bitstream and transport mechanisms is expected to facilitate the eventual integration of its new features into the next TVD realm, but that is a future addressable target, not subject to further speculation at the moment.

By the time we conclude this article, the TV3.0 project is undergoing. Still deliberations are expected on other layers, as necessary developments and tests are in progress (for instance in the applications coding layer, which also employs MPEG-H). Working Groups (GTs) are currently addressing technology issues and needs (such as the necessary developments in authoring tools and innovative personalizable and interactive programs for additional tests) and challenges in establishing regulatory policies for Internet OTT (Over the Top) and broadcast OTA (Over the Air) delivery.

REFERENCES

[1] Brazilian Digital Terrestrial TV Forum, CfP Phase 2 / Testing and Evaluation: TV 3.0 Project (15 March 2021), available at the TV3.0 Project website at https://forumsbtvd.org.br/tv3_0/

[2] ATSC 3.0 Audio Testing Report, Doc. S34-2B-048r7 (12 August 2015).

[3] ATSC Standard: A/342:2021 Part 1, Audio Common Elements (Doc. A/342:2021 Part 1) (9 March 2021)

[4] ATSC Standard: A/342:2021 Part 2, AC-4 System (Doc. A/342:2021 Part 2)" (10 March 2021).

[5] Advanced Television Systems Committee, "ATSC 3.0 Audio Testing Report, Doc. S34-2B-048r7", Washington, USA, August 2015.

[6] MPEG-H 3D Audio Verification Test Report, ISO/IEC JTC1/SC29/WG11 MPEG2017/N16584 (January 2017, Geneva)

[7] IEEE Standard for Second Generation Audio Coding, IEEE Std. 1857.8-2020, pp.1-470, 25 Nov. 2020, doi: 10.1109/IEEESTD.2020.9271961.

[8] ETSI TS 103 190 Digital Audio Compression (AC-4) Standard, version 1.1.1, 2014. Available at https://www.etsi.org/deliver/etsi_ts/103100_103199/103190/01.01.01_60/ts_103190v010101p.pdf

[9] ETSI TS 103 190-2 Digital Audio Compression (AC-4) Standard; Part 2: Immersive and personalized audio, version 1.2.1, 2018. Available at https://www.etsi.org/deliver/etsi_ts/103100_103199/10319002/01.02.01_60/ts_10319002v010201p.pdf.

[10] Advanced Television Systems Committee, "Techniques for Establishing and Maintaining Audio Loudness for Digital Television," Doc. A/85:2013, Washington, D.C., Mar. 12, 2013. Corrigendum No. 1, "SPL," approved Feb. 11, 2021.

[11] Information Technology—High Efficiency Coding and Media Delivery in Heterogeneous Environments—Part 3: 3D Audio, Standard ISO/IEC 23008-3:2022, 3rd ed., 2022. Available at https://www.iso.org/standard/83525.html.

[12] Y. Grewe, A. Murtaza, S. Meltzer. "MPEG-H Audio System for SBTVD TV 3.0 Call for Proposals", *SET International Journal of Broadcast Engineering*, v. 7, 2021.

[13] Recommendation ITU-R BS.2076-2 (10/2019), Audio Definition Model, https://www.itu.int/rec/R-REC-BS.2076/en.

[14] R. R. A. Faria, A. A. Rosa, E. Mendes, A. A. B. Silva, D. H. S. Abreu, S. D. Costa, H. F. Rozena, G. K. F. Komatzu, "Testing and Evaluation Report: TV 3.0 Project – Audio Coding", Brazilian Digital Terrestrial Television System Forum, University of São Paulo, Dec. 3, 2021. Available at https://forumsbtvd.org.br/tv3_0/#panel-phase2.

[15] D. Jang, T. Lee, Y. Lee, & Yoo, J. H. (2006, October). A personalized preset-based audio system for interactive service. In Audio Engineering Society Convention 121. Audio Engineering Society Available: http://www.aes.org/e-lib/browse.cfm?elib=13738

[16] R. R. A. Faria, M. K. Zuffo and J. A. Zuffo (2005, July). Improving spatial perception through sound field simulation in VR. In IEEE Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2005. IEEE, doi: 10.1109/VECIMS.2005.1567573.

[17] Okubo, H., Sugimoto, T., Oishi, S., & Ando, A. (2012, October). A method for reproducing frontal sound field of 22.2 multichannel sound utilizing a loudspeaker array frame. In Audio Engineering Society Convention 133. Audio Engineering Society. Available: http://www.aes.org/e-lib/browse.cfm?elib=16456.

[18] A. Ando, "Conversion of Multichannel Sound Signal Maintaining Physical Properties of Sound in Reproduced Sound Field," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1467-1475, Aug. 2011, doi: 10.1109/TASL.2010.2092429.

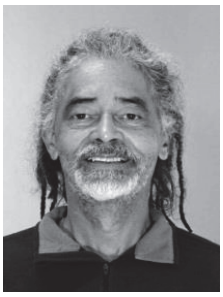
[19] R. L. Bleidt, D. Sen, A. Niedermeier, B. Czelhan, ... & M. Y. Kim, "Development of the MPEG‐H TV audio system for ATSC 3.0", *IEEE Transactions on broadcasting*, 63(1), pp. 202‐236, 2017, doi: 10.1109/TBC.2017.2661258.

Regis Rossi A. Faria received the B.S. degree in electrical engineering from the Federal University of Minas Gerais (UFMG), Brazil, in 1990 and the M.S. degree and Ph.D. degree in electrical engineering from the University of São Paulo (USP), Brazil, respectively in 1997 and 2005.

In 2004 he was a researcher at USP collaborating to the development of the first generation of the Brazilian digital television system (SBTVD) in the audio layer. He has coordinated the testing and evaluation of audio technologies for the SBTVD TV3.0 project phase 2 in 2021, being a current collaborator with the SBTVD technical Forum.

Dr. Faria is an Associate Professor with the University of São Paulo, where he coordinates the Laboratory of Audio and Music Technology (LATM) at the School of Arts, Sciences and Humanities (EACH), and has research interest in spatial audio, sound and music computing, audio engineering, immersive aesthetics and production technologies. He is member of the Audio Engineering Society and of the Brazilian Computer Society.

https://set.org.br/ijbe doi: 10.18580/setijbe.2023.2 Web Link: https://dx.doi.org/10.18580/setijbe.2023.2



Almir Almas is PhD in Communication and Semiotics by the Pontifical Catholic University of Sao Paulo and an Associate Professor of the Department of Film, Radio and Television at the School of Communications and Arts of the University of São Paulo, and Researcher of the Program of Postgraduate Studies in Media and Audiovisual Processes, where he is the General Coordinator of the Research Group LabArteMídia (Laboratory of Art, Media and Digital Technologies) and Obted (Brazilian Observatory of Digital Television and Technological Convergence).

He is currently Visiting Professor and Researcher (Support by FAPESP BPE) at the Faculty of Humanities and Social Sciences and School of English and Modern Languages at Oxford Brookes University. Author of 'Televisão digital terrestre: sistemas, padrões e modelos' (Digital terrestrial television: systems, standards and models), among other books and articles.

Dr. Almas is Filmmaker/Videoartist/VJ, and Artist of the Cobaia Art Collective and Formigueiro. He is a Member of the Board of the Brazilian Society of Television Engineering (SET) and Member of the Brazilian Digital Terrestrial Television System Forum (FORUM SBTVD).

Eduardo Santos Mendes was born in São Paulo, Brazil. He received the B.S. and M.S. degrees in Film from USP (São Paulo University) and the Ph.D. degree in Arts/Film Sound in 2000 from USP.

He is a sound supervisor since 1984 and professor at USP since 1990 where he taught in the Bachelor's Degree in Film, Radio and TV and Film and Video courses. At the moment, he teaches in the Bachelor's Degree in Audiovisual Course. He is a member and advisor in the Postgraduate Program in the Audiovisual Media and Processes (USP) since 2002. Dr. Mendes was also a guest teacher in schools in Mexico and Belgium. In his work as a sound supervisor, he collaborated with main Brazilian directors such as Carlos Adriano, Carlos Reichenbach, Lina Chamie, Tata Amaral and Walter Hugo Khouri.

Dr. Mendes is part of the board of directors of CIBA/CILECT, was awarded in Brazilian and International festivals such as Brasilia, FestRio and Havana (Cuba). He is a researcher of audiovisual technology and narrative, focusing on sound.

Ana Amelia Benedito Silva is a Brazilian Professor of the Postgraduate Program in Modeling Complex Systems and the Undergraduate Course in Information Systems at the School of Arts, Sciences and Humanities at the University of São Paulo. She achieved her bachelor's, master's, and PhD degrees in Electrical Engineering at the University of São Paulo.

Ana Amelia Benedito Silva has experience in the field of studies of rhythmic phenomena, with an emphasis on sleep, and the application of mathematical and statistical models to the analysis of empirical data mainly linked to the area of biology and health. Themes of activity: sleep, sleep-wake cycle, shift worker sleep, biological rhythms, biostatistics, mathematical modeling, series analysis temporal, malnutrition, exercise physiology.

Douglas H. S. Abreu was born in Lavras - MG, Brazil. He earned his bachelor's degree in Information Systems from the Federal University of Lavras, Brazil, in 2015, and his master's degree in Systems Engineering and Automation from the same university in 2017. Currently, he is a professor at the Polytechnic School of PUC Campinas and a doctoral candidate in Electrical Engineering and Computing at UNICAMP.

Throughout his career, Douglas has focused on Machine Learning, audio and acoustics, and recently cybersecurity, working on notable projects such as mapping critical infrastructures and their cyber vulnerabilities in the Brazilian electrical sector, and testing and evaluating audio technologies for Brazilian TV 3.0. He is currently involved in the Technological Training Program in Artificial Intelligence (AI) promoted by CPQD and PUC-Campinas with the support of the Ministry of Science, Technology, and Innovations (MCTI).

Prof. Abreu is a member of the Laboratory of Acoustics of Communications at FEEC/UNICAMP (LAC-Unicamp).

Henrique F. Rozena was born in São Paulo, capital of the state of São Paulo, Brazil, in 2000. He graduated from high school at Escola de Aplicação and continues his studies at the FATEC State Technology college, within the digital media design course.

In 2021 he was an assistant fellow in the research and testing project for audiovisual equipment in the TV 3.0 Project, and has since then helping to conduct the X-Reality events at the University of São Paulo where the project is conducted.

Received in 2023-06-08 I Approved in 2023-07-08

https://set.org.br/ijbe doi: 10.18580/setijbe.2023.2 Web Link: https://dx.doi.org/10.18580/setijbe.2023.2