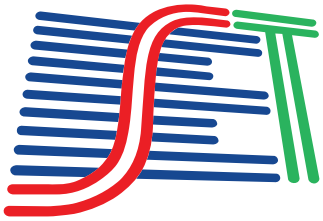
SET INTERNATIONAL JOURNAL OF BROADCAST ENGINEERING

SET IJBE V. 7, 2021

ISSN print: 2446-9246
ISSN online: 2446-9432



SET INTERNATIONAL JOURNAL OF BROADCAST ENGINEERING

SET IJBE V. 7, 2021

ISSN print: 2446-9246
ISSN online: 2446-9432

SET INTERNATIONAL JOURNAL OF BROADCAST ENGINEERING

ISSN PRINT: 2446-9246 | ISSN ONLINE: 2446-9432

INDEXED IN:

Google's system that offers tools for search of academic literature.

International Cataloging Data in the Publication - CIP - Librarian Zoraide Gasparini CRB/9 1529

S517 SET International Journal of Broadcast Engineering —— vol. 7, (Dec. 2021). – São Paulo: Brazilian Society of Television Engineering —— SET 2021.

Annual frequency
ISSN Print 2446- 9246
ISSN online 2446- 9432
Available at: https://www.set.org.br/ijbe

Broadcasting – Periodic. 2. Transmission Engineering. 3. Digital TV. I. SET. II. Title.

CDD: 384.54

ANNUAL | CIRCULATION : 500 EXEMPLARY | GRAPHIC DESIGN AND DIAGRAM: SOLANGE LORENZO

EDITORIAL BOARD

CORPORATE AUTHOR AND EDITOR
SET
Address: Av Auro Soares de Moura Andrade, 252, suite 31 - Barra Funda District - São Paulo - SP
Brazil - Postal Code: 01156-001

SET BOARD OF DIRECTORS

Deliberative Council

2021 - 2022

President: Carlos Fini

Vice-President: Claudio Eduardo Younis

OFFICE HOLDER	**SUBSTITUTE**
Carlos Fini	Carlos Cauvilla
Luiz Bellarmino Polak Padilha	David Estevam de Britto
Claudio Eduardo Younis	José Carlos Aronchi de Souza
Claudio Alberto Borgo	Luis Otavio Marcheze
Mauro Alves Garcia	Almir Antonio Rosa
Daniela Helena Machado e Souza	José Salustiano Fagundes de Souza
Vinicius Augusto da Silva Vasconcellos	Sergio Silva
Raymundo Costa Pinto Barros	Marcelo Santos Wance de Souza
Roberto Dias Lima Franco	Valderez de Almeida Donzelli
Emerson Weirich	Paulo Henrique Corona Viveiros de Castro
Sergio Eduardo di Santoro Bruzetti	Nelson Faria
José Eduardo Marti Cappia	Marco Tulio Nascimento
José Raimundo Lima da Cunha	Esdras Miranda de Araujo
Marcio Rogério Herman	Israel de Moraes Guratti
Cristiano Akamine	Marcelo Moreno
Rafael Duzzi de Oliveira	Fabio Ferraz
Jurandir Moreira Pitsch	Wagner Kojo

Fiscal Council

Nivelle Daou
José Chaves F. de Oliveira
Marcos Paulo Teixeira

Rafael Silveira Leal
Sandro Sereno

Council of Former Presidents

Adilson Pontes Malta
Carlos Eduardo Oliveira Capelão
Fernando Mattoso Bittencourt Filho
José Munhoz

Liliana Nakonechnyj
Olímpio José Franco
Roberto Dias Lima Franco

Regional representatives

North: Henrique Camargo e Eduardo Lopes
North East: Ronald Almeida e Gabriel Eskenazi
Midwest: Wender de Souza
Southeast: Geraldo Mello
South: Caio Klein

ABOUT THE JOURNAL

SET INTERNATIONAL JOURNAL OF BROADCAST ENGINEERING

The SET IJBE, (SET International Journal of Broadcast Engineering) is an open access, peer-reviewed article--at-a-time international scientific journal whose objective is to cover knowledge about communications engineering in the field of broadcasting. The SET IJBE seeks the latest and most compelling research articles and state-of-the-art technologies.

Publishing schedule and schema

On-line version – Once an article is accepted and its final version approved by the Editorial Board, it will be published immediately on-line on a one article-at-a-time basis

Printed version – Once a year, all articles accepted and published on-line over the previous twelve months will be compiled for publication in a printed version.

Types of papers

Regular (Full) Papers: Traditional and original research [from 6 to 20 pages]

Tutorial Papers: Brand new OTT (Over-The-Top) detailed implementation and fully set up state-of-the art systems [4 – 6 pages]

Letters: Short notes and consideration about current and relevant techniques, technologies and implementations involving engineering solutions [1-3 pages]

Open Access Policy

If an article is accepted for publication, it will be made available to be read and re-used under a Creative Commons Attribution (CC-BY) license.

Editorial Office:

If you require any additional information, please contact the SET IBJE (SET International Journal of Broadcast Engineering) administration staff:

Address: Av Auro Soares de Moura Andrade, 252, suite 31 - Barra Funda District - São Paulo - SP
Brazil - Postal Code: 01156-001

Aims and Scope include, but are not limited to:

Advanced audio technology and processing
Advanced display technologies
Advanced RF Modulation Technologies
Advanced technologies and systems for emerging broadcasting applications
Applying IT Networks in Broadcast Facilities
Broadcast spectrum issues – re-packing, sharing
Cable & Satellite interconnection with terrestrial broadcasters
Cellular broadcast technologies
Communication, Networking & Broadcasting
Content Delivery Networks – CDN
Digital radio and television systems: Terrestrial, Cable, Satellite, Internet, Wireless.
Electromagnetic compatibility issues between collocated services (e.g. broadcast and LTE)
General Topics for Engineers (Math, Science & Engineering)
Hybrid receiver technology
Interactive Technology for broadcast
IP Networks management and configuration
Metadata systems and management
Mobile DTV systems (all aspects, both transmission and reception)
Mobile/dashboard technology
Next-gen broadcast platforms and standards development
Non-real time (NRT) broadcast services
Ratings technology, second screen technology and services
Secondary service system design; mitigation of interference in primary services
Securing Broadcast IT Networks
Signal Processing & Analysis
Software Defined Radio – SDR Technologies
Streaming delivery of broadcast content
Transmission, propagation, reception, re-distribution of broadcast signals AM, FM, and TV transmitter and antenna systems
Transport stream issues – ancillary services
Unlicensed device operation in TV white spaces

We wish to inform you that the activities, events and publications of the Brazilian Society of Television Engineering – SET, including this one, enjoy international support, under formal agreements, from the following international organizations. We also take this opportunity to thank them and reiterate how proud we are that they support our work.

SUMMARY

SET IJBE v. 7, 2021, 56 pages, 4 articles

EDITORIAL

Dear reader,

This issue of IJBE is dedicated to the call for proposals (CfP) of the TV 3.0 project. The TV 3.0 project is the next generation of digital terrestrial television broadcasting proposed by the SBTVD Forum. As Phase 2 of the TV 3.0 project was completed in December 2021, there is no doubt that we are waiting for Phase 3 and the recommended technologies for the next issue. We have covered several physical layer-related technologies over the past few years. Now we are focusing on several other related technologies, such as video encoding, audio encoding, and application encoding layers.

The present issue presents several thematic papers covering different aspects of the proposed TV 3.0 technologies.

The search for quality in the development of technologies that make digital communication systems more efficient is the focus of researchers for whom IJBE offers the opportunity to disseminate their studies, experiments, and research in the scientific and technological areas of production and distribution of information content.

We hope you enjoy these articles and feel motivated to submit an article.

Best wishes,
SET IJBE Editors

Versatile Video Coding for 3.0 Next Generation Digital TV in Brazil

Thibaud Biatek
Mohsen Abdoli
Mickael Raulet
Adam Wieckowski
Christian Lehman
Benjamin Bross
Philippe De Lagrange
Edouard François
Ralf Schaefer
Jean Lefeuvre

CITE THIS ARTICLE
Biatek,Thibaud; Abdoli,Mohsen; Raulet,Mickael; Wieckowski, Adam; Lehman, Christian; Bross, Benjamin; De Lagrange, Philippe; François, Edouard; Schaefer, Ralf and Lefeuvre, Jean; 2021. Versatile Video Coding for 3.0 Next Generation Digital TV in Brazil . SET INTERNATIONAL JOURNAL OF BROADCAST ENGINEERING. ISSN Print: 2446-9246 ISSN Online: 2446-9432. doi: 10.18580/setijbe.2021.1. Web Link: http://dx.doi.org/10.18580/setijbe.2021.1

SET INTERNATIONAL JOURNAL OF BROADCAST ENGINNERING - SET IJBE V.7, 2021, Article 1, 8p.
© 2021 SET / ISSN Print: 2446-9246 | ISSN Online: 2446-9432

Versatile Video Coding for 3.0 Next Generation Digital TV in Brazil

Thibaud Biatek[1], Mohsen Abdoli[1], Mickael Raulet[1], Adam Wieckowski[2], Christian Lehman[2], Benjamin Bross[2], Philippe De Lagrange[3], Edouard François[3], Ralf Schaefer[3] and Jean Lefeuvre[4]

[1]ATEME, [2]Fraunhofer HHI, [3]InterDigital, [4]Télécom Paris

***Abstract*— In the past few decades, the video broadcast ecosystem has gone through major changes; Originally transmitted using analog signals, it has been more and more transitioned toward digital, leveraging compression technologies and transport protocols, principally developed by MPEG. Along this way, the introduction of new video formats was achieved with standardization of new compression technologies for their better bandwidth preservation. Notably, SD with MPEG-2, HD with H.264, 4K/UHD with HEVC. In Brazil, the successive generations of digital broadcasting systems were developed by the SBTVD Forum, from TV-1.0 to TV-3.0 nowadays. The ambition of TV-3.0 is significantly higher than that of previous generations as it targets the delivery of IP-based signals for applications, such as 8K, HDR, virtual and augmented reality. To deliver such services, compressed video signals shall fit into a limited bandwidth, requiring even more advanced compression technologies. The Versatile Video Coding standard (H.266/VVC), has been finalized by the JVET committee in 2021 and is a relevant candidate to address the TV-3.0 requirements. VVC is versatile by nature thanks to its dedicated tools for efficient compression of various formats, from 8K to 360°, and provides around 50% of bitrate saving compared to its predecessor HEVC. This paper presents the VVC-based compression system that has been proposed to the SBTVD call for proposals for TV-3.0. A technical description of VVC and an evaluation of its coding performance is provided. In addition, an end-to-end live transmission chain is demonstrated, supporting 4K real-time encoding and decoding with a low glass-to-glass latency.**

***Index Terms*— DTT, OTT, VVC, 8K, 4K, HDR, HFR, Broadcast**

I. Introduction

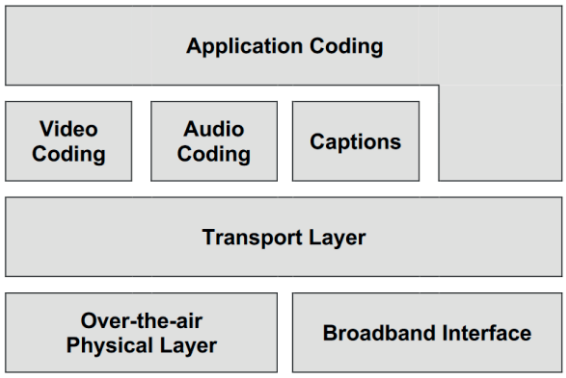
THE SBTVD Forum (Sistema Brasileiro de TV Digital [1]) is a Brazilian organization responsible for development of digital television in Brazil. The organization gathers more than eighty members, private and public, covering the complete ecosystem, including broadcasters, manufacturers, institutions, and universities. Over the years, SBTVD enabled the transition from analog to digital, releasing a number of specifications issued by the Brazilian Association of Technical Standards (ABNT). In its most recent version [2], the coding technology used remains H.264/AVC [3] even though advanced picture format such as High Dynamic Range (HDR) is covered.

In July 2020, SBTVD issued a call for proposals aiming at extending the current digital television system to new use-cases and futuristic applications. Among these applications, video format such as 8K, HDR, AR/VR shall be supported and delivered over broadcast, or hybrid broadband/broadcast networks, based on IP-centric protocols. To enable the delivery of such signals with high Quality of Experience (QoE) for the viewer, the usage of advanced compression systems is required.

Versatile Video Coding (VVC) [4][5], also known as H.266, is the most recent video coding standard issued by the Joint Video Experts Team (JVET) of ITU-T and ISO/IEC and approved in July 2020. VVC is the successor of High Efficiency Video Coding (HEVC) standard [6], providing around 50% of bitrate saving for the same visual quality. VVC has been designed to address a wide range of applications (e.g. broadcast, streaming, layered coding) and formats (e.g. 4K/8K, HDR, VR-360, Screen-Content), and is then a relevant candidate to address TV-3.0 requirements.

VVC was proposed by a consortium formed by ATEME, DiBEG, Fraunhofer HHI and InterDigital as coding technology for TV3.0 in November 30th 2020 and since then is going through the evaluation procedures. Beside VVC, other codecs have been proposed, such as HEVC, LCEVC [7] and AVS3 [8]. Together with the formal submission, an end-to-end delivery scheme was provided enabling to demonstrate the VVC readiness and use-cases in a live environment. The demonstrator includes a live VVC headend, an origin-/ad-server, and a video player capable of live VVC decoding and ad-insertion. In this paper, the VVC proposal brought to SBTVD for TV3.0 is thoroughly described, including technology, performance and demonstrator.

The rest of the paper is organized as follows. First, Section II introduces the TV-3.0 project, highlighting the required key features, evaluation procedure and timeline. Second, the VVC compression technology is presented in Section 0. The Section IV presents the end-to-end transmission chain delivered to SBTVD to assess the relevance of VVC in a live environment. Finally, the paper is concluded in Section V.



https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.1. Web Link: http://dx.doi.org/10.18580/setijbe.2021.1

Fig. 1: TV-3.0 Architecture [9]

II. THE TV-3.0 PROJECT

A. Overview

The TV3.0 call for proposals has been issued in July 2020 by the SBTVD forum [9], in order to proceed beyond the TV2.5 project, enabling enhanced features and new applications. The TV-3.0 architecture is composed of a number of components, described in Fig. 1, covering the application coding, audio/video coding, captions, transport and physical layers. It is expected that the system will rely on IP protocols for both over-the-air or over-the-top delivery. In addition, the system shall support accessibility and public-safety features which are included as requirements. In this paper, the response that proposes VVC to this call as the video-coding component is presented.

B. Video Coding Requirements

The TV3.0 video coding component is expected to provide state-of-the-art compression efficiency for live services with a minimal glass-to-glass latency. The native support of HDR and UHDTV is a requirement for the broadcast/broadband delivery, while the support for 8K resolution is only restricted to broadband. It is desired that a second video stream can be encoded and delivered for sign language purpose. In addition, support for public safety applications such as emergency warning or sign-language video is desired, as well as an anticipation of immersive technologies support (i.e VR, AR, XR, 3DoF and 6DoF). From an application perspective, the selected technology shall support seamless and frame-accurate stream splicing or ad-insertion for broadcast, broadband or hybrid delivery modes. In order to address hybrid use-cases, it is also expected that the selected codec supports scalable and layered coding. In order to be addressed by the proponents, SBTVD split all these features in several use-cases and requirements, reported in the Table 1.

Table 1 : TV-3.0 Video Coding Use-cases [9]

ID	Description
VC1	Provide improved video resolution, adequate to consumer electronics display evolution.
VC2	Provide improved video dynamic range and color space, adequate to consumer electronics display evolution.
VC3	Provide sharp images (reducing motion blur), even on content with fast motion (e.g. sports, action movies).
VC4	Provide state-of-the-art coding efficiency, to allow better quality video in limited capacity channels (over-the-air or the Internet).
VC5	Provide live video with minimum end-to-end latency.
VC6	Enable second video stream with a sign language interpreter to be optionally activated by the user (to be rendered at the side of the main video, that should be proportionally downscaled to fit the horizontal space left, with no overlap; an optional background still image can be defined by the broadcaster).
VC7	Enable emergency warning information delivery using sign language video.
VC8	Enable new immersive video services.
VC9	Enable seamless decoding and A/V alignment.
VC10	Enable interoperability with different distribution platforms (e.g. DTT, cable, IPTV, DTH satellite, fixed broadband, 4G/5G mobile broadband, home network).
VC11	Enable scalability (e.g. to improve over-the-air video quality with an Internet-delivered enhancement layer) and extensibility (support new settings and/or features in the future, in a backward-compatible way).

In the TV-3.0 CfP response jointly submitted by ATEME, Fraunhofer HHI and InterDigital, VVC has been proposed as technology addressing the complete range of mandatory use-cases defined by SBTVD. DiBEG also responded to the CfP with VVC and joint during the evaluation process the proposal of ATEME, Fraunhofer HHI and InterDigital. The video quality aspects VC1 to VC4 are fulfilled thanks to the approach used by JVET to design VVC which is tailored to efficiently address all formats (SD to 8K), dynamic range (SDR to HDR) and type of content (e.g. gaming, sport, movie, screen content, video conferencing) while achieving a maximal compression efficiency. The capability of delivering VVC compression in a live workflow (VC5) is addressed, demonstrated by VVC support in ATEME live encoding platform already demonstrated on the field [10][11].
The low-latency delivery case VC6 can also be addressed by VVC when combined with CMAF Low-Latency and HTTP Chunk Transfer Encoding [12]. VVC is fulfilling the application-oriented use-cases VC6 and VC7 by encoding the second video streams in an efficient manner. VVC natively supports VR-360 coding through specific high-level mechanisms and is thus addressing VC8. In addition, VVC can also be used as core 2D video codec in a V-PCC [13] compression engine and thus can be used to compress volumetric data for AR/VR, 3DoF or 6DoF applications. VVC through its defined picture types and encapsulation in ISOBMFF supports seamless splicing or ad-insertion (VC9). The interoperability of VVC with different distribution platforms (VC10) was demonstrated during several on-field trials [10][11]. Finally, VVC natively supports scalable coding (VC11).

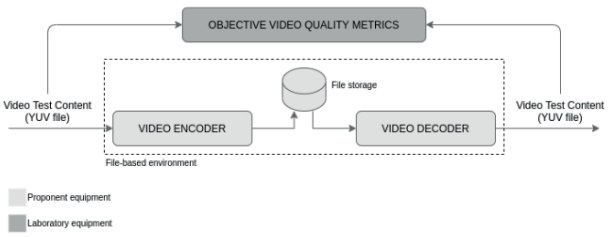


Fig. 2 : Evaluation procedure for non-real-time tests [14]

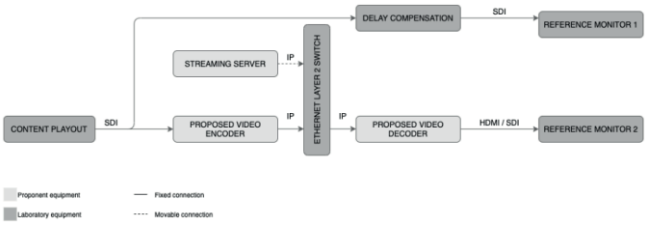


Fig. 3 : Evaluation procedure for real-time tests [14]

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.1. Web Link: http://dx.doi.org/10.18580/setijbe.2021.1

C. *Process and timeline*

It is specified that TV-3.0 is organized in two phases, with dedicated timelines. The first phase consisted of a formal submission, including contact information and how the proposed technology addresses the use-cases' requirements. The first phase response due date was expected on November 30th 2020. The second phase, expected to be delivered on January 29th 2021, comprised full specification of the solution as well as resources required to evaluate and test the proponents.

In order to evaluate and test the proposed technologies, the proponents should deliver equipment, following architectures described in Fig. 2 and Fig. 3, for non-real-time and real-time tests respectively. This includes video encoder, video decoder, streaming server and file storage. In terms of timeline, the evaluation takes place from July 2021 to December 2021.

III. Versatile Video Coding

A. *Overview*

Significantly higher performance and more versatility compared to HEVC have been two main elements in the VVC standardization agenda. To meet these requirements, several tools and functionalities were integrated into VVC in the course of its standardization.

From the versatility point of view, VVC aims to address compression of applications such as 8K, screen content, immersive video, and multi-resolution over-the-top streaming. At the same time, higher compression efficiency was achieved by introducing advanced coding tools to exploit signal redundancies that are either specific to abovementioned applications, or simply are more efficient than in the previous standards (e.g. HEVC).

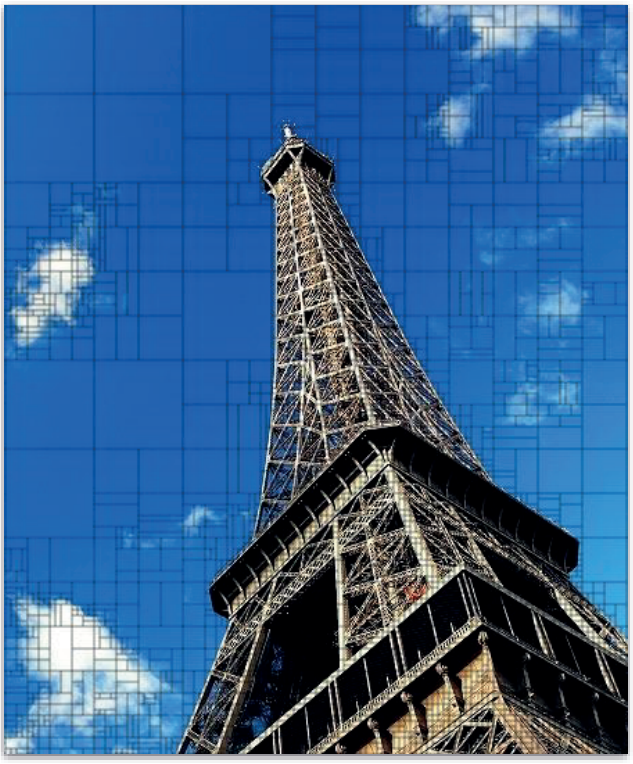


Fig. 4 : Illustration of VVC partitioning

B. *Core Coding Tools*

New tools adopted in VVC can roughly be categorized into four groups: partitioning, intra/inter prediction, residual coding, and in-loop filtering.

1) *Partitioning*

By far, the main advantage of VVC compared to other video codecs in terms of core coding tools, is its incredibly flexible partitioning scheme [15]. In addition to larger Coding Unit Tree (CTU) size (up to 128 compared to 64, in HEVC), this scheme allows splitting different regions of image in a way that their local is more appropriately represented for spatial and temporal prediction. Compared to HEVC's Quad-Tree (QT), this algorithm of VVC is integrated with binary and ternary (horizontal and vertical) splits in addition to quad split. Consequently, this partitioning method of VVC is known as Multi-Type Tree (MTT). **Erro! Fonte de referência não encontrada.** shows an example of how flexibly the MTT partitioning of VVC can split regions of an image based on its content complexity.

2) *Prediction*

Tens of new coding modes have improved both intra [16] and inter [17][18] prediction performance. In intra prediction, the main contributions are the use of 67 Intra Prediction Modes (IPM) instead of 35 (as in HEVC), Position-Dependent Prediction Combination (PDPC), data-driven block prediction called Matrix-based Intra Prediction (MIP), the possibility of explicitly choosing the reference line, Intra Sub-Partitioning (ISP) for short-distance pixel prediction, and last, but not least, advanced chroma intra prediction tools such as Cross-Component Linear Model (CCLM).

3) *Residual coding*

VVC introduces several new ways of encoding prediction residual. This results in a significant overall compression efficiency gain on different types of content [19]. Firstly, VVC allows rectangular transforms for encoding prediction residual in the MTT partitioning. This aspect helps in significant rate saving that was used in HEVC to split rectangular residual blocks into square shapes and transmit separately. Furthermore, VVC is integrated with more transform kernels using discrete sine and cosine transforms (DST, DCT), precisely, DST-I, DCT-II, DCT-V, DST-VII, DCT-VIII, compared to only DCT II, DST VII (limitedly) in HEVC. In addition to traditional separate transformation, VVC also allows a new mode called Low-Frequency Non-Separable Transforms (LFNST), that performs a low complexity non-separable transformation on the residual. This aspect alone accounts for about 2%-3% additional compression efficiency gain compared to HEVC.

As screen content coding has usually very specific signal characteristics, VVC is equipped with a dedicated residual coding method, called Transform Skip (TS). As the name suggests, TS performs the quantization step in the pixel domain without any transformation in between [19]. Finally, Dependent Quantization (DQ) enables a switching between two quantizers for decoding each transform coefficient. This choice depends on the previous quantized coefficient's value and a pre-defined state-machine.

4) *In-loop filtering*

As the processing capacity of both encoder and decoder sides have been significantly improved compared to the last decade, VVC has integrated several in-loop processing and filtering technologies that were not affordable in previous standards [20]. Adaptive Loop Filter (ALF) is a new

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.1. Web Link: http://dx.doi.org/10.18580/setijbe.2021.1

technique, which includes filter shapes, precision, and adaptive clipping processes, in order to tune the filtering parameters both in the sub-blocks and Coding Tree Blocks (CTB) levels. ALF is adaptive in the sense that the filtering coefficients are signaled in the bitstream and can be designed based on image content and distortion of the reconstructed picture. Moreover, a variation of the ALF, called Cross-Component ALF has also been adopted in VVC, which uses the luma sample values to refine the chroma sample values within the ALF calculation. Finally, Luma Mapping with Chroma Scaling (LMCS) is another pre-processing and in-loop technique to exploit the dynamic range of the signal. Precisely, LMCS consists of two steps of the Luma Mapping (LM), which remaps the luma code values, and the Chroma Scaling (CS), which allows flexible adjustment between luma and chroma adjustment.

C. *High-Level Features*

VVC has introduced new high-level syntax features [21]. Through its HLS, VVC offers flexible control over tools and features that deal with so-called versatile applications (e.g., layered coding, 360, virtual reality, screen content etc). Examples of such HLS functionalities are as follows. The concept of sup-pictures, tied with the newly introduced Picture Header (PH) and Decodable Capability Information (DCI), flexibly allows random access to subsets of the bitstream, representing localized regions of interest. Thanks to these dedicated syntaxes, subpictures can be arbitrarily recomposed without header rewriting.

Reference Picture Resampling (RPR) is another functionality that allows layered coding and is controlled in the SPS. Other than the traditional use-cases of scalability, RPR potentially saves significant coding efficiency by enabling open GOP in adaptive streaming applications, where the resolution changes frequently. Last, but not least, the support of Screen content coding (SCC) has been carefully considered in the VVC standardization, by dedicating several coding tools that are switchable in different levels of the HLS, in order to limit their usage where their particular content is present [22].

D. *Performance*

The initial agenda of VVC in terms of performance vs. complexity tradeoff was to strictly limit the decoder side complexity, while allowing a certain level of complexity increase at the encoder side. Consequently, software-based decoding of VVC streams (using all tools) is less than twice as complex as decoding of HEVC streams. Given the 10-year gap between the release of the standards, this complexity increase has been considered acceptable by decoder manufacturers since it can be afforded by their state-of-the-art technologies.

The VVC encodings for objective and subjective evaluation have been performed using the Fraunhofer Versatile Video Encoder, VVenC, a software VVC encoder implementation, in version 1.0.0 [23]. It is freely available under a 3-clause BDD license [24]. While providing very good compression efficiency, it is only suitable for offline usage. The objective performance evaluation results for three tests scenarios [14] are reported and compared to the HEVC test model (HM) encoder anchor:

- Test 1 contains two test cases requiring coding of video having different spatial resolutions from 720p to 4320p, one (TC 1.1) for Hybrid Log Gamma (HLG) High Dynamic Range (HDR) video and one for Perceptual Quantizer (PQ) HDR video (TC 1.2).
- Test 3 includes one case (TC 6.1) testing VVC encoding of 1080p Standard Dynamic Range (SDR) video for various temporal resolutions i.e., frame rates from 23.98 (24/1001) fps to 120 fps.
- Test 6 requires coding sign language video in two smaller resolutions (540x960 and 360x640) which is typically shown as picture in picture. One test case covers HLG HDR (TC 6.1) and the other test case covers PQ HDR (TC 6.2) video.

Fig. 5 : Average PSNR Y, U and V BD-rate savings of VVC (VVenC v1.0) over HEVC (HM-16.23).

The video in all test cases uses Wide Color Gamut (WCG) colorimetry according to recommendation ITU-R BT.2100 [25]. Fig. 5 summarizes the average PSNR Bjøntegaard Delta (BD) rates for all test cases. VVC provides steady and significant bit-rate savings over HEVC for all test cases. In terms of luma PSNR Y, savings over 30% are measured. Furthermore, the chroma PSNR U and V savings are significantly higher, ranging between 50% and 60%. This confirms that VVC has been designed for a high coding efficiency for this type of colorimetry.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.1. Web Link: http://dx.doi.org/10.18580/setijbe.2021.1

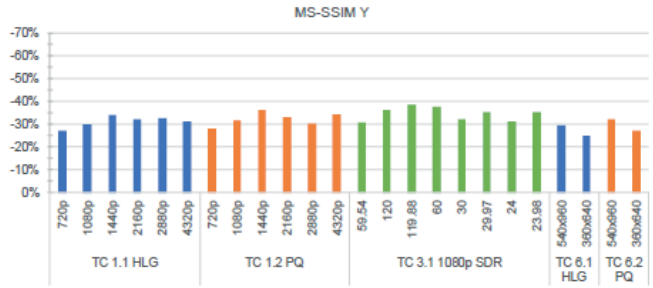


Fig. 6 : Average MS-SSIM Y BD-rate savings of VVC (VVenC v1.0) over HEVC (HM-16.23).

The average MS-SSIM Y based bit-rate savings over 30% shown in Fig. 6 are comparable to the PSNR Y results. For the U and V component we observed a strange behavior of MS-SSIM values, especially for HDR content, which is discussed in more details in this document. For the HDR PQ content item, the weighted PSNR (wPSNR) metric has been calculated as well and is shown in Fig. 7. Compared to PSNR, the wPSNR-based BD-rate savings are slightly higher. The VVC verification tests for HD and UHD video have shown that the "real" bit-rate savings for the same perceived video quality using subjective tests and Mean Opinion Scores (MOS) tend to be higher than the ones based on objective metrics such as PSNR or MS-SSIM [26][27]. Overall, VVC is capable of coding HDR WCG video at various spatial and SDR WCG video at various temporal resolutions as well as portrait mode sign language HDR WCG and PQ WCG videos.

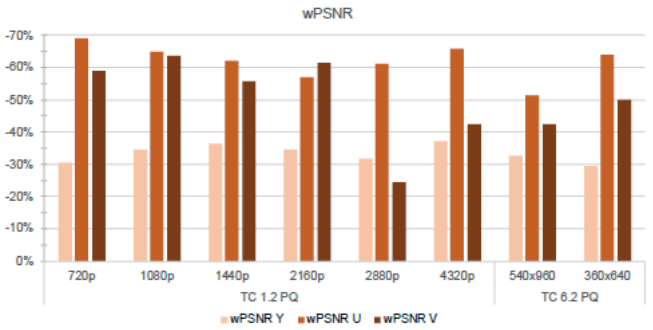


Fig. 7 : Average wPSNR BD-rate savings of VVC (VVenC v1.0) over HEVC (HM-16.23).

IV. Proposed End-to-End Transmission Chain

A. Overview

In order to test and demonstrate VVC readiness in a live environment, we delivered to the SBTVD Forum a complete end to end transmission chain, described in the Fig. 8.

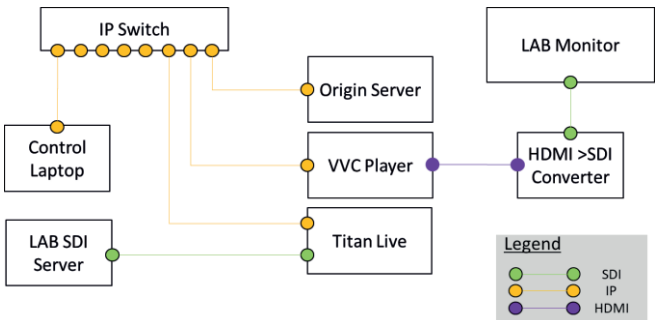


Fig. 8 : Proposed end-to-end transmission scheme

The chain includes several components. A Titan Live encoder from ATEME, providing live VVC-encoding, ISOBMFF packaging and CMAF Low-Latency encapsulation. The video chunks produced are pushed to an origin server, supporting HTTP chunk transfer encoding for low-latency. The origin server is also used as host for the ad-clips to demonstrate splicing and ad-insertion use-case. A mini-PC is also provided to be used as a live VVC player, supporting live decoding as well as ad-replacement, leveraging the following open-source software [28] :

- Telecom-Paris' GPAC player [29]
- FFmpeg/FFplay [30]
- Fraunhofer HHI's VVdeC [31]

B. Headend

The ATEME TitanLive solution provides software-based implementation of a wide variety of standards for audio/video coding, packaging and transport. This solution is currently used worldwide for broadcast and OTT head-end deployments. In order to support VVC, a number of components were upgraded.

As further described in [32], VVC and HEVC present some structural similarities making an upgrade from HEVC to VVC feasible in a cost effective manner. In order to do so, the VVC syntax has been implemented with support for the tools already implemented in HEVC, disabling the other ones in the APS. Then, the HEVC tools have been upgraded to comply with VVC specification and some tools offering a good complexity-vs-gains trade-offs were implemented. Relying on the same core coding engine enabled us to leverage the existing optimized function (assembly, intrinsic) to achieve VVC real-time encoding with interesting gains over HEVC, from 10% to 15% depending on the video content. The packager has been upgraded as well to support VVC encapsulation into MPEG2-TS and ISOBMFF following FDIS specifications.

C. Origin Server

The ATEME origin server provided to SBTVD for TV3.0 testing has two main roles to enable end-to-end low latency. The first one is to act as an HTTP Chunked Transfer Encoding proxy able to ingest and serve data as soon as it is available. This mechanism consists of storing the HTTP chunks received from the packager in such a way that it will be able to send them to the Content Delivery Network (CDN) when requested.

The second role is to manage the requests for playlists and media segments. The origin server is responsible of blocking the request to HLS playlists if needed. When a segment and a part number are included in the request, the response will be delayed until the playlist contains at least the specified part of the segment. Regarding the request to media segments, the origin server support open range requests already described in this document. More generally and independently of the request range, the origin server will immediately serve the bytes that are already available in his cache and the send the remaining data as soon as it is received while keeping the request handler opened as long as the whole payload has not been sent. Eventually, this origin server is also hosting pre-encoded ad-clips used to demonstrate ad-personalization.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.1. Web Link: http://dx.doi.org/10.18580/setijbe.2021.1

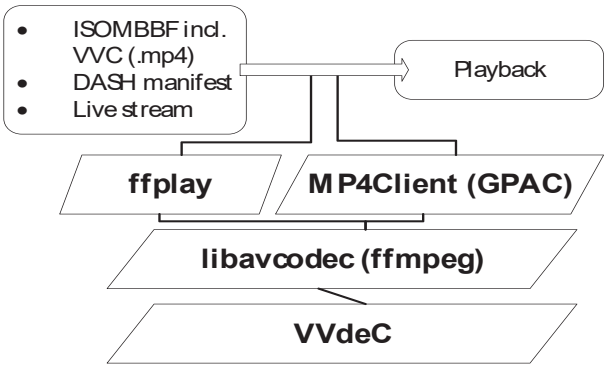


Fig. 9 :VVC player components [28]

D. Player

The delivered media player relies on what's is described in [28], and illustrated in Fig. 9. It is composed of the GPAC media player and the VVdeC optimized VVC decoding library.

1) Media player

GPAC is a multimedia framework providing tools to package, stream and playback multimedia content. It is well-known for its MP4Box tool, an mp4 file packager and HTTP streaming (HLS) segmenter. GPAC also enables end users to build custom multimedia processing pipelines through its filter-based architecture [29]. The project is distributed under LGPL v2.1+ license.

The systems aspect of VVC have been implemented in the master branch of GPAC and are part of the nightly builds of the project. This covers the most common application use cases:

- MPEG-2 broadcasting: multiplexing and demultiplexing an MPEG2-TS program with VVC content
- MP4 file packaging and dumping of VVC bitstreams
- MPEG-DASH and HLS content packaging, with transport over HTTP or ROUTE

GPAC supports all encoders and decoders integrated in FFmpeg's libavcodec library, including VVdeC as included in [28]. The GPAC player embeds the customized libavcodec which itself embeds the VVdeC core decoding library. Until integration of VVenC is finalized in libavcodec, GPAC can consume the VVenC output through files and pipes.

2) VVC decoding library

The Fraunhofer Versatile Video Decoder, VVdeC, is an optimized VVC software decoder implementation, freely available on GitHub under a 3-clause BSD copyright license [33]. Analogue to VVenC, the license covers both commercial and non-commercial use. The latest release, v1.2.0, was published in September 2021. It is compliant with the VVC Main 10 profile and correctly decodes the current VVC conformance testing set [34]. A major part of this release is providing browser playback functionality based on open standards.

The decoder software has been derived from VTM as well, with subsequent optimizations and parallelization. The decoder allows playback of HD video at 60 frames per second (fps) on most modern computers using only 2 or 3 threads, and 60 fps UHD playback on more powerful modern workstations with sufficient processing cores to fully exploit the multi-threading potential [35].

The VVdeC package contains a simple and easy-to-use C library interface and a standalone decoder application capable of decoding elementary VVC bitstreams into raw YUV video data. The integration into the described frameworks allows actual playback of the decoded VVC content.

V. Conclusion

In this paper, the VVC proposal to the TV-3.0 CfP is described. The adequateness of VVC is discussed showing that it addresses all the TV-3.0 use-cases and requirements in an efficient manner. An overview of VVC is provided to highlight the key tools as well as a performance evaluation. To reinforce the relevance of VVC an end-to-end transmission chain is provided to show its readiness for commercial deployment in the context of TV-3.0 in Brazil.

References

[1] SBTVD Forum, https://forumsbtvd.org.br/
[2] *ABNT NBR 15602-1, "Digital terrestrial television – Video coding, audio coding and multiplexing, Part 1: Video Coding"*
[3] T. Wiegand, G. J. Sullivan, G. Bjontegaard and A. Luthra, "Overview of the H.264/AVC video coding standard," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560-576, July 2003, doi: 10.1109/TCSVT.2003.815165.
[4] *B. Bross et al., "Overview of the Versatile Video Coding (VVC) Standard and its Applications," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 10, pp. 3736-3764, Oct. 2021, doi: 10.1109/TCSVT.2021.3101953.*
[5] *W. Hamidouche et al., "Versatile Video Coding Standard : A Review from Coding Tools to Consumers Deployment," in IEEE Consumer Electronics Magazine, Accepted (to be published).*
[6] *G. J. Sullivan, J. Ohm, W. Han and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1649-1668, Dec. 2012, doi: 10.1109/TCSVT.2012.2221191.*
[7] *F. Maurer et al., "Overview of MPEG-5 Part 2 – Low Complexity Enhancement Video Coding (LCEVC)", in ITU Journal: ICT Discoveries, Vol. 3(1), 8 June 2020.*
[8] *J. Zhang, C. Jia, M. Lei, S. Wang, S. Ma and W. Gao, "Recent Development of AVS Video Coding Standard: AVS3," 2019 Picture Coding Symposium (PCS), 2019, pp. 1-5, doi: 10.1109/PCS48520.2019.8954503.*
[9] *SBTVD Forum, "Call for Proposals: TV 3.0 Project", July 17, 2020, https://forumsbtvd.org.br/wp-content/uploads/2020/07/SBTVDTV-3-0-CfP.pdf*
[10] ATEME Press-Release, "ATEME and The Explorers to Launch the First OTT Channel Promoting VVC", https://www.ateme.com/ateme-and-the-explorers-to-launch-the-first-ott-channel-promoting-vvc/
[11] ATEME Press-Release, "ATEME Joins Forces with SES to Trial First-Ever Live Over-The-Air UHD Broadcast Using VVC", https://www.ateme.com/ateme-joins-forces-with-ses-to-trial-first-ever-live-over-the-air-uhd-broadcast-using-vvc/
[12] ISO/IEC 23000-19:2020, "Information technology – Multimedia application format (MPEG-A) – Part 19: Common Media Application Format (CMAF) for segmented media".
[13] Information technology - Coded representation of immersive media - Part 5: Visual volumetric video-based coding (V3C) and video-based point cloud compression (V-PCC), ISO/IEC 23090-5:2021
[14] SBTVD Forum, "CfP Phase 2 / Testing and Evaluation : TV3.0 Project", https://forumsbtvd.org.br/wp-content/uploads/2021/03/SBTVD-TV_3_0-P2_TE_2021-03-15.pdf
[15] Huang, Yu-Wen, et al. "Block partitioning structure in the VVC standard." *IEEE Transactions on Circuits and Systems for Video Technology* (2021).
[16] Pfaff, Jonathan, et al. "Intra prediction and mode coding in VVC." *IEEE Transactions on Circuits and Systems for Video Technology* (2021).
[17] Chien, Wei-Jung, et al. "Motion Vector Coding and Block Merging in the Versatile Video Coding Standard." *IEEE Transactions on Circuits and Systems for Video Technology* 31.10 (2021): 3848-3861.
[18] Yang, Haitao, et al. "Subblock-based motion derivation and inter prediction refinement in versatile video coding standard." *IEEE Transactions on Circuits and Systems for Video Technology* (2021).

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.1. Web Link: http://dx.doi.org/10.18580/setijbe.2021.1

[19] Zhao, Xin, et al. "Transform coding in the VVC standard." *IEEE Transactions on Circuits and Systems for Video Technology* 31.10 (2021): 3878-3890.
[20] Karczewicz, Marta, et al. "VVC in-loop filters." *IEEE Transactions on Circuits and Systems for Video Technology* (2021).
[21] Wang, Ye-Kui, et al. "The high-level syntax of the versatile video coding (VVC) standard." *IEEE Transactions on Circuits and Systems for Video Technology* (2021).
[22] Nguyen, Tung, et al. "Overview of the screen content support in VVC: Applications, coding tools, and performance." *IEEE Transactions on Circuits and Systems for Video Technology* (2021).
[23] A. Wieckowski et al., "VVenC: An Open And Optimized VVC Encoder Implementation," 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2021, pp. 1-2, doi: 10.1109/ICMEW53276.2021.9455944.
[24] Fraunhofer HHI VVenC software repository. Retrieved from https://github.com/fraunhoferhhi/vvenc.
[25] ITU-R, "Image parameter values for high dynamic range television for use in production and international program exchange", Rec. ITU-R BT.2100-2, 2018.
[26] V. Baroncini and M. Wien, "VVC verification test report for UHD SDR video content," doc. JVET-T2020 of ITU-T/ISO/IEC Joint Video Experts Team (JVET), 20th JVET meeting: October 2020.
[27] M. Wien and V. Baroncini, "VVC Verification Test Report for High Definition (HD) and 360° Standard Dynamic Range (SDR) Video Content," doc. JVET-V2020 of ITU-T/ISO/IEC Joint Video Experts Team (JVET), 22st JVET meeting: April 2021.
[28] A. Wieckowski, C. Lehmann, B. Bross, D. Marpe, T. Biatek, M. Raulet and J. LeFeuvre, *"A Complete End-To-End Open Source Toolchain for the Versatile Video Coding (VVC) Standard",* In *Proceedings of the ACM Multimedia Conference.* ACM, October, 2021, Chengdu, CN.
[29] Jean Le Feuvre. 2020. GPAC filters. In *Proceedings of the 11th ACM Multimedia Systems Conference*. ACM, New York, NY, USA, 249–254. DOI: https://doi.org/10.1145/3339825.3394929
[30] FFmpeg fork with full VVdeC integration. Retrieved from https://github.com/tbiat/FFmpeg/releases/tag/vvc.
[31] A. Wieckowski, G. Hege, C. Bartnik, C. Lehmann, C. Stoffers, B. Bros, and D. Marpe. 2020. Towards a Live Software Decoder Implementation for the Upcoming Versatile Video Coding (VVC) Codec. In *2020 IEEE International Conference on Image Processing (ICIP)*, October, 2020, Abu Dhabi, UAE 3124–3128. DOI: https://doi.org/10.1109/ICIP40778.2020.9191199
[32] Biatek, T., et al. "Future MPEG standards VVC and EVC: 8K broadcast enabler." *Proc. Int. Broadcast. Conv.*. 2020.
[33] Fraunhofer HHI VVdeC software repository. Retrieved from https://github.com/fraunhoferhhi/vvdec.
[34] J. Boyce, E. Alshina, F. Bossen, K. Kawamura, I. Moccagatta and W. Wan. 2021. *Conformance testing for versatile video coding (Draft 6)*. Doc. JVET-U2008 of ITU-T/ISO/IEC Joint Video Experts Team (JVET), 21st JVET meeting: January 2021.
[35] A. Wieckowski, G. Hege, C. Bartnik, C. Lehmann, C. Stoffers, B. Bros, and D. Marpe. 2020. Towards a Live Software Decoder Implementation for the Upcoming Versatile Video Coding (VVC) Codec. In *2020 IEEE International Conference on Image Processing (ICIP)*, October, 2020, Abu Dhabi, UAE 3124–3128. DOI: https://doi.org/10.1109/ICIP40778.2020.9191199

Thibaud Biatek received the Ph.D. degree in signal and image processing from the Institut National des Sciences Appliquées, Rennes, France, in 2016. From 2013 to 2017, he was a Doctoral and Post-Doctoral Fellow with TDF, Cesson-Sévigné, France. From 2017 to 2019, he was a Video Coding Expert with TDF, where he was involved in MPEG and DVB standardization activities. In 2019, he was a Senior Engineer with Qualcomm working on VVC standardization. Since 2020, he has been Director of Technology and Standards with ATEME, working on partnership projects and multimedia standards, contributing to MPEG, DVB and 3GPP groups. His research interests include compression, processing, and delivery of audiovisual signals over broadcast and broadband networks.

Mohsen Abdoli received his Master of Engineering degree in 2013 from Sharif University of Technology, Tehran, Iran, and his Doctor of Philosophy in 2018, jointly from Université Paris-Saclay and CentraleSupélec, Paris, France. In 2015, he joined Orange Labs, Rennes, France, where he actively contributed to development of several video compression techniques, targeting the VVC standard by JVET, particularly in the domain of intra prediction and residual coding of natural/screen content. In 2018, he joined Ateme, Rennes, France, where he contributed to the world-first implementation of real-time VVC transcoder for VoD and live streaming applications. He is currently with IRT b<>com, Rennes, France, holding the position of standardization engineer. His research interests include areas of signal processing, encoder optimization, machine learning and quality assessment.

Mickaël Raulet is CTO at ATEME, where he drives research and innovation with various collaborative R&D projects. He represents ATEME in several standardization bodies: ATSC, DVB, 3GPP, ISO/IEC, ITU, MPEG, DASH-IF, CMAF-IF, SVA and UHDForum. He is the author of numerous patents and more than 100 conference and journal scientific papers. In 2006 he received his Ph.D. from INSA in electronic and signal processing, in collaboration with Mitsubishi Electric ITE (Rennes, France).

Adam Wieckowski received the M.Sc. degree in computer engineering from the Technical University of Berlin, Berlin, Germany, in 2014.,In 2016, he joined the Fraunhofer Institute for Telecommunications–Heinrich Hertz Institute, Berlin, as a Research Assistant. He worked on the development of the software, which later became the test model for VVC Development. He contributed several technical contributions during the standardization of VVC. Since 2019, he has been a Project Manager coordinating the technical development of decoder and encoder solutions for the VVC standard.

Christian Lehmann received the M.S. degree in computer science from University Leipzig, Germany, in 2006. He joined the Fraunhofer Institute for Telecom-munications – Heinrich Hertz Institute, Berlin, Germany in 2009, where he is a research associate in the Video Coding Systems group.His research interests include the development of optimized software video codecs, multimedia frameworks and systems integration. This includes real-time video decoder and encoder solutions for High Efficiency Video Coding (HEVC) and the Versatile Video Coding (VVC) open source implementations VVenC and VVdeC.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.1. Web Link: http://dx.doi.org/10.18580/setijbe.2021.1

Benjamin Bross received the Dipl.-Ing. degree in electrical engineering from RWTH Aachen University, Aachen, Germany, in 2008. In 2009, he joined the Fraunhofer Institute for Telecommunications – Heinrich Hertz Institute, Berlin, Germany, where he is currently heading the Video Coding Systems group at the Video Coding & Applications Department and in 2011, he became a part-time lecturer at the HTW University of Applied Sciences Berlin. Since 2010, Benjamin is very actively involved in the ITU-T VCEG | ISO/IEC MPEG video coding standardization processes as a technical contributor, coordinator of core experiments and chief editor of the High Efficiency Video Coding (HEVC) standard [ITU-T H.265 | ISO/IEC 23008-2] and the new Versatile Video Coding (VVC) standard [ITU-T H.266 | ISO/IEC 23090-3]. In addition to his involvement in standardization, his group is developing standard-compliant software implementations. This includes the development of an HEVC live software encoder that is currently deployed in broadcast for HD and UHD TV channels and most recently, the open and optimized VVC software implementations VVenC and VVdeC. Benjamin Bross is an author or co-author of several fundamental HEVC and VVC-related publications, and an author of two book chapters on HEVC and Inter-Picture Prediction Techniques in HEVC. He received the IEEE Best Paper Award at the 2013 IEEE International Conference on Consumer Electronics – Berlin in 2013, the SMPTE Journal Certificate of Merit in 2014 and an Emmy Award at the 69th Engineering Emmy Awards in 2017 as part of the Joint Collaborative Team on Video Coding for its development of HEVC.

Edouard François is currently Principal Scientist at InterDigital where he is leading a research team focused on video and media compression. In previous years, he has been part of Technicolor and Canon research. He received an Engineering degree from the IMT Atlantique, France and a Ph.D. degree in computer science from University of Rennes 1, France. He has been participating to several standardization activities, including the specifications of the scalable extension of AVC/H.264 (known as SVC), the HEVC/H.265 standard and its extensions, and the recent VVC/H.266 standard. He had previously co-chaired the MPEG activity on HDR and WCG video coding. His main research interests include signal and video processing, conventional and ML-based video coding, and high dynamic range video.

Philippe de Lagrange received the Eng. degree in electrical engineering from the INSA of Rennes, France, in 1999. He is currently with Interdigital R&D France, in the Core Video Coding team, where he has contributed to VVC through his active participation to the JVET group. He also leads a team that promotes VVC through optimized software implementations (e.g. full-featured multi-threaded VTM decoder, including multi-layer), performance testing, publications, partnerships, and participation to applicative standards. He provided the system infrastructure for offline testing of VVC and remote support during the evaluation phase of TV3.0 in Brazil. Previously, he worked on broadcast video encoders (TVN/Harmonic and Envivio/MediaKind), decoders (Technicolor), and also worked for a radiocommunication research lab (Mitsubishi ITE), and in satellite positioning R&D (Galileo).

Ralf Schaefer received his engineering degree (Dipl.–Ing.) from the Technical University Kaiserslautern/Germany and joined InterDigital in June 2019. As full time standard professional, Ralf chairs the MPEG WG7 AhG on Video-based Graphics Coding, is vice-chair of the DVB Commercial Module and is elected member of the DVB Steering Board. Furthermore, Ralf is an active contributor to standards working groups in DVB, ETSI, SBTVD Brazil, NorDiG and FAVN France. Previously he chaired the MPEG 3DG AhG on Point Cloud Compression, which advanced the work on Video-based Point Cloud Compression (V-PCC) between MPEG meetings and led to the publication of ISO/IEC 23090-5:2021. Before joining Interdigital, Ralf occupied a similar position in Technicolor, where he contributed to ATSC3.0 and NGBF South Korea. Earlier he held R&D positions at various levels in Thomson and chaired working groups in DVB related to IPTV, Home Networking and Companion Screens.

Jean Le Feuvre received his Ingénieur (M.Sc.) degree in Telecommunications in 1999, from TELECOM Bretagne. He has been involved in MPEG standardization since 2000 for his NYC-based startup Avipix,llc and joined TELECOM Paris in 2005 as Research Engineer within the Image, Data and Signal Department. His main research topics cover multimedia authoring, delivery and rendering systems in broadcast, broadband and home networking environments. He is the project leader and maintainer of GPAC, a multimedia framework based on standard technologies (MPEG, W3C, IETF). He is the author of many scientific contributions (peer-reviewed journal articles, conference papers, book chapters, patents) in the field and is editor of several ISO standards."

Received in 2021-10-29 | Approved in 2021-12-20

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.1. Web Link: http://dx.doi.org/10.18580/setijbe.2021.1

HDR10+ Concepts, Principles, Capabilities and Advantages for the Next-Generation of Brazilian Broadcasting System (TV 3.0)

Rodrigo Admir Vaz
Luiz Gustavo Pacola Alves
Steve Larson

CITE THIS ARTICLE
Vaz, Rodrigo Admir; Alves, Luiz Gustavo Pacola; Larson, Steve; 2021. HDR10+ Concepts, Principles, Capabilities and Advantages for the Next-Generation of Brazilian Broadcasting System (TV 3.0). SET INTERNATIONAL JOURNAL OF BROADCAST ENGINEERING. ISSN Print: 2446-9246 ISSN Online: 2446-9432. doi: 10.18580/setijbe.2021.2. Web Link: http://dx.doi.org/10.18580/setijbe.2021.2

SET INTERNATIONAL JOURNAL OF BROADCAST ENGINNERING - SET IJBE V.7, 2021, Article 2, 11p.
 / ISSN Print: 2446-9246 | ISSN Online: 2446-9432

HDR10+ Concepts, Principles, Capabilities and Advantages for the Next-Generation of Brazilian Broadcasting System (TV 3.0)

Vaz, Rodrigo Admir; Alves, Luiz Gustavo Pacola; Larson, Steve

SIDIA Instituto de Ciência e Tecnologia - DTV Lab.
Samsung – SRA

***Abstract* – This paper aims to briefly introduce the current Brazilian DTV scenario towards the next generation DTT system (TV 3.0) as well as HDR (High Dynamic Range) principles, features and applications. Different HDR formats have been proposed along the history with different features and purposes. One of them is the HDR10+ technology that can provide powerful advantages and benefits related to picture quality and based on dynamic metadata system. HDR10+ is being adopted by major international digital terrestrial TV systems and other standardizations bodies. In this scenario TV 3.0 can take advantage of HDR10+ to improve the overall system from industry perspective, and consequently delivering a consistent user experience.**

Key Words – HDR, HDR10+, TV 3.0, Video Quality, Next-generation of broadcasting standards and systems, Video coding and processing

I. INTRODUCTION

In the second half of 2020, SBTVD (*Sistema Brasileiro de Televisão Digital* - Brazilian Digital Television System) Forum – association responsible to the make DTV (digital terrestrial television) system specifications used in Brazil – released CfP (*Call for Proposals*) with the requirements of the next Brazilian terrestrial DTV system (named TV 3.0), open to all international associations to deliver respective proposals as solutions to the requirements. TV 3.0 CfP requirements were divided in groups, as physical and transport layers, audio and video coding, captions, and application coding.

During the first half of 2021, SBTVD Forum defined the test procedures to evaluate the candidate solutions, and during the second half 2021, SBTVD Forum Test Labs will be testing proposed solutions [1].

Several features are part of the requirements of TV 3.0, as C/N ≤ 0 dB, MIMO (Multiply Input Multiple Output) 2x2 support, channel bonding, vital Broadband and Broadcast integration, emergency warnings, accessibility, immersive audio, scalable audio, and video. Especially for video, UHD (Ultra-High Definition) 4K and 8K image, HFR (High Frame Rate), WCG (Wide Color Gamut) and HDR (High Dynamic Range) support are required.

Varied solutions were proposed by the consortiums to meet TV 3.0 requirements, as MMTP (MPEG Multimedia Transport Protocol), ROUTE (Real-time Object delivery over Unidirectional Transport) for transport layer; MPEG-H, AVSA and AC-4 for audio coding; H.266, H.265 and AVSA for video coding; SMPTE ST 2094-10 (Dolby Vision), SMPTE ST 2094-20, SMPTE ST 2094-30 (SL-HDR (1/2/3)) and SMPTE ST 2094-40 (HDR10+) for HDR Dynamic Mapping Codec [1].

HDR is a technology used to improve picture quality, better representing luminance, and colors in videos images. It is contrasted with SDR (standard dynamic range), which has become the term for older technology. HDR offers the possibility to represent more realistic images with substantially brighter highlights, darker shadows, and more colors than what was previously possible. HDR enables better use of displays that have high brightness, contrast, and color capabilities.

A HDR technology encompasses various technical characteristics, such as transfer function, dynamic or static metadata, bit depth, maximum luminance range, and color volume. Some HDR formats have been proposed since 2014 and the most common formats are Dolby Vision, HLG, SL-HDR, HDR10 and HDR10+ [2].

HDR10+ has been established as SMPTE standard ST 2094-40 [3], and is a video codec agnostic solution which aligned with TV 3.0 video requirements [1] can leverage next generation TV experiences and bring several advantages to the new Brazilian DTV system: enhanced video quality,better user experience on consumer TVs.

Dynamic tone mapping is another important characteristic offered by HDR10+. Such feature applies a different tone mapping curve from scene to scene, maximizing image quality while preserving the creative intent. Dynamic tone mapping eliminates clipping in highlights and crushing in the darks [3].

As a royalty free technology with a simple production workflow, the adoption and device expansion have experienced significant growth since HDR10+'s launch, allowing worldwide consumers to access the new technology. An ever-expanding catalog of international video content is already produced with HDR10+ and also available through major content and streaming video providers. Professional tool development is facilitated with a number of commercial and royalty free SW development kits and cloud CLI

Vaz, Rodrigo Admir, studies Ph.D. in digital TV at Mackenzie Presbyterian University, besides being a senior engineer at SIDIA DTV Lab (e-mail: rodrigo.vaz@samsung.com).
Alves, Luiz Gustavo Pacola is MSc by University of São Paulo and technical coordinator at SIDIA DTV Lab (e-mail: luiz.alves@samsung.com).
Larson, Steve is project manager at Samsung SRA US Lab (e-mail: steve.l@samsung.com).

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.2. Web Link: http://dx.doi.org/10.18580/setijbe.2021.2

(command-line interface) tools available for implementing HDR10+ into almost any application.

This paper describes several advantages offered by HDR10+ technology to improve TV 3.0 system and offer enhanced experience to the end user and is organized as follows. Section II provides an introduction to fundamentals of image quality and importance of HDR. Section III shows HDR working principle. Section IV illustrates the workflows to produce HDR10+ video contents. Section V shows HDR10+ in practice. Section VI details benefits, and advantages offered by HDR10+ paving the way to TV 3.0 and Section VII concludes this paper.

II. IMAGE QUALITY FUNDAMENTALS AND HDR IMPORTANCE

Display devices such as LCD (Liquid Crystal Displays) with global backlight screens, OLEDs (Organic Light-Emitting Diodes), QD (Quantum Dot) enhancement layers, direct-view LED technology and others are acquiring the ability to show an increased range of brights, darks, and color, improving image quality, leading to an increased realism [4].

A. Image Quality Elements

Considering this image quality improvement, it is possible to highlight five elements, which have a huge influence over image quality [2], [4], [5].

- *Spatial Resolution*: Bigger spatial resolution shows better image details and increases the image perceived depth sensation. Currently, the highest video resolutions used are 4K (3840x2160) and 8K (7680x4320) UHD (Ultra high Definition).

- *Frame Rate (also known as temporal resolution)*: Higher frame rate is especially beneficial for scenes with significant motion, such as sporting events, or action movies since it decreases the image blur. Most modern video content are produced using 60 or 120 fps.

- *Wide Color Gamut*: Modern HDR displays are capable of showing a wider color spectrum than ever before. Current displays accept an input signal with a BT 2020 [6] color gamut.

- *Color Depth (also known as Bit Depth)*: Is the quantity of bits used by each color component to compose a pixel. The larger the number of bits used, the smoother the color transition inside the image with be. UHD content supports at least 10 bits per color.

- *Luminance*: With higher luminance capabilities now available, more realistic luminance variation between scenes such as sunlight, indoor, and night scenes can be shown, as well higher contrasts can be explored. Presently, luminance limits used in the video content converted from SDR (maximum of 100 nits) to HDR limits can reach up to 10,000 nits.

Figure 1 illustrates possible combinations of these elements.

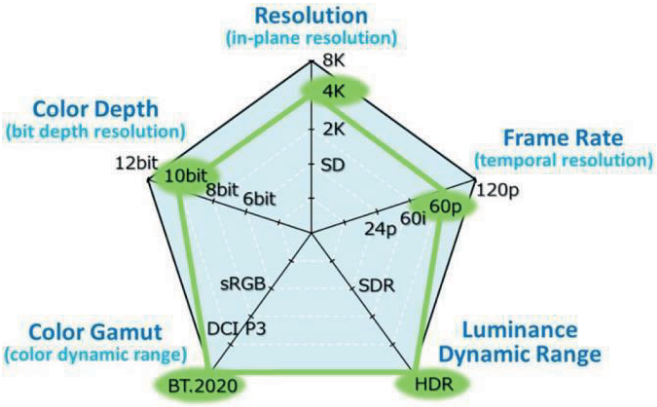


Figure 1- Five Elements of Image Quality Combined – Taken from Reference [2].

B. HDR Importance

All these elements combined leads to image quality improvement, allowing the display device to show more realist images, with greater picture detail, more vibrant colors, and smoother motion.

HDR has a great impact on image quality, delivering darker "darks" and brighter "brights" along with more nuanced gradations for better delineation of on-screen shapes. This feature offers some benefits as below [4].

- *Perceptual Benefits*: The human eye is capable of sensing light values from starlight to bright sunlight, a 10^{14} range of illuminance. However, total dynamic range is only achieved after many minutes of night vision adaptation, at any given moment, only a fraction of that range is available: estimated at around 13 to 16 stops (about 10^5 dynamic range).

Using SDR, the original TV imaging system could only capture, record, transmit and display less than 10 stops of dynamic range, limiting how content creators could convey scene brightness. This problem persisted with the launch of HDTV in the 1990's.

HDR takes advantage of the latest cameras, processing, storage, distribution platforms and displays, to maintain as much of the full range of human vision as possible from the original scene all the way through to the display.

- *Technical Benefits*: HDR technology is often associated with "brighter" pictures measured in "nits". While brightness is a major aspect of HDR – it is only one part of the story. HDR is really about the entire dark-to-light range of tonal values, what cinematographers call "grayscale." HDR enables us to see the entire grayscale: not just the highlights, but all the shadows and all the subtle gradations in between, as depicted in Figure 2.

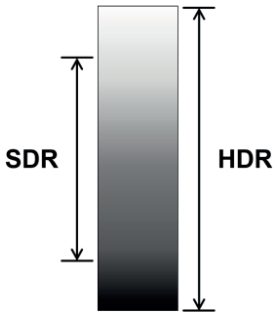


Figure 2 - Today's Cameras Capture Deep Shadows and Bright Highlights. SDR Cannot Reproduce Them. HDR Can. – Taken from Reference [4].

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.2. Web Link: http://dx.doi.org/10.18580/setijbe.2021.2

• *Creative Benefits*: Grayscale improvements go directly to the heart of visual storytelling, enabling a greater range of expression. Careful control of grayscale's light and shadow helps to establish mood, convey realistic skin tones, and identify what to look for in a scene. In fact, it is so important that for each individual scene, movie crews typically set up a unique configuration of lights, reflectors, and diffusers to achieve just the right effect.

Moreover, the director and cinematographer continue to perfect grayscale values during postproduction, via color correction and mastering. By expanding and refining the grayscale, HDR dramatically improves movies, TV programs and videogames, making it more engaging, dynamic, and closer to the creative intent.

III. HDR WORKING PRINCIPLE

HDR is a technology used to improve picture quality, better representing luminance, and colors in videos images, offering the possibility to represent more realistic images with substantially brighter highlights, darker shadows, and more colorful colors than what was previously possible. This feature enables better use of displays that have high brightness, contrast, and color capabilities.

To take advantage of the increase in the peak luminance, different transfer functions were proposed, namely the PQ (Perceptual Quantizer) curve [7], [8]. This curve is capable of representing luminance level up to 10,000 nits and down to 0.0001 nits.

A transfer function, also known as tone mapping curve, renders incoming HDR contents on a display having a smaller dynamic range, improving image contrast, therefore showing more realistic pictures.

Figure 3 shows the need for HDR technology presenting different video contents.

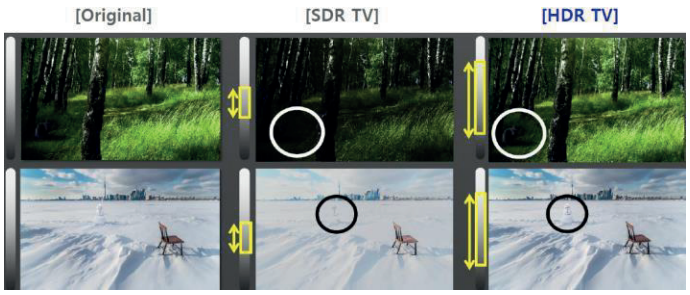


Figure 3 – Difference Among Original, SDR and HDR Contents.

A. *Static Tone Mapping*

Early HDR Technologies use static tone mapping, applying one fixed tone mapping curve (transfer function) to all frames of the content. One example of this feature is HDR10 [8], [9], which is a widespread technology, that uses PQ [6], [7] curve and static metadata.

PQ curve enables content creators to exactly specify the image color and brightness as viewed on a reference monitor in the grading suite. It is up to displays to play this directly or adapt to the consumer's viewing environment.

Static metadata includes information about the display on which the content was mastered, information such as Maximum Frame Light Level (MaxFALL) and Maximum Content Light Level (MaxCLL).

HDR10 contents production is shown in Figure 4. Such information is used by the receiving display, to adjust its own brightness for the content, as depicted in Figure 5. The horizontal axis represents complete range of possible light levels in PQ-encoded content. The vertical axis is the light level produced by an HDR10 display.

However, these values remain static throughout the runtime for HDR10 content, which produces less than ideal results wherein some scenes are not as bright as they can be while others could be brighter than they needed to be. This is not efficient approach since dark scene gets dimmer if tone mapping is designed to avoid highlight saturation and highlight gets saturated (banding) if designed to avoid dimming in dark area.

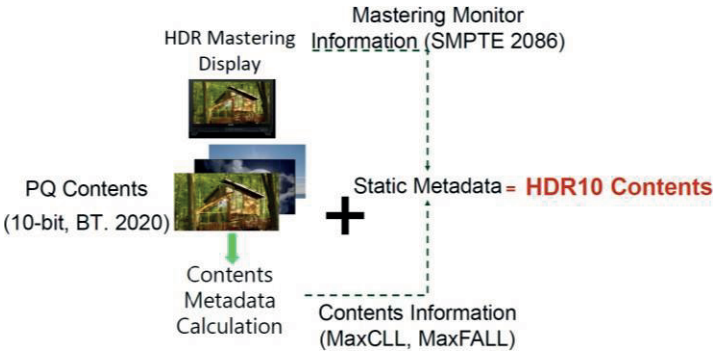


Figure 4 - HDR10 Contents Production.

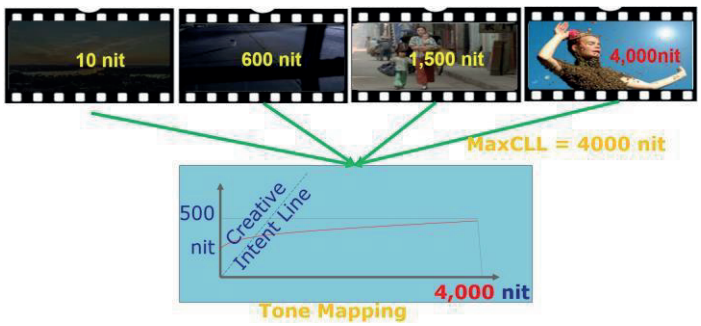


Figure 5 – Static Tone Mapping Applies One Fixed TM Curve to Entire Contents.

B. *Dynamic Tone Mapping*

More modern HDR Technologies use dynamic tone mapping, applying scene specific tone curves via dynamic metadata.

With the support of dynamic metadata, information is created for the content on a scene-by-scene or frame-by-frame basis for adjusting the content light level. This information can then be used by receiving display to adjust luminance to match the mastered content exactly,reproducing the creator's intent more closely. It is efficient tokeep the creative intent: no excessive dimming, maintained saturation and highlight detail.

An illustration of scene based dynamic tone mapping is depicted in Figure 6. It uses 500 nits as the display peak luminance. Dynamic tone mapping happens when the source content peak luminance is higher than the display device peak luminance.

HDR10+ is a HDR technology, which adds dynamic metadata on top of already existing static HDR10 technology, and is standardized in SMPTE ST 2094-40 [3] spec.

ST-2094-40 metadata includes percentile information calculated on the actual scene content, a guided Opto-optical Transfer Function (OOTF) curve. It allows precise adaptation

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.2. Web Link: http://dx.doi.org/10.18580/setijbe.2021.2

of tone mapping based on Bezier curve (N[th] order Bernstein polynomial) to various display capabilities.

Figure 7 shows tone mapping function based on N[th] Bezier curve. The tone mapping function is composed of two sections: The first section is linear while the second section is a polynomial. The two sections are conjoined at a knee point (Kx, Ky). The coefficients of the Bernstein polynomial along with the coordinates of the knee point are part of the dynamic metadata that defines the scene-specific tone mapping curve [3], [10].

The Figure 8 and Figure 9 show image quality difference between Static and Dynamic tone mapping.

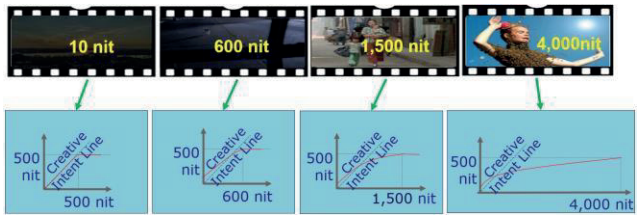


Figure 6 - Dynamic Tone Mapping: Scene Based Tone Curves.

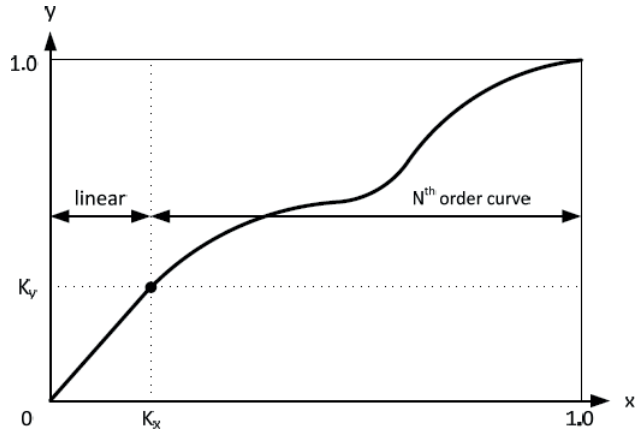


Figure 7 - Tone Mapping Function Based on N[th] Bezier Curve. Taken from Reference [10].

Figure 8 - Static Tone Mapping Image.

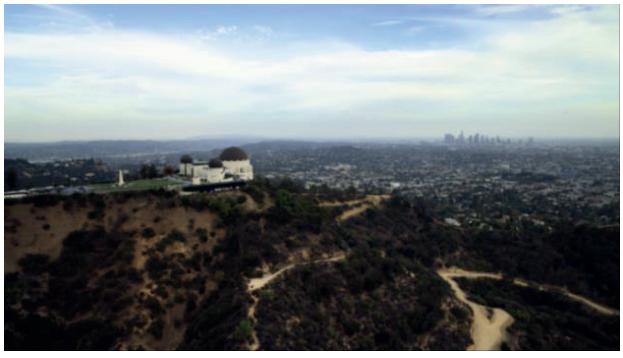


Figure 9 - Dynamic Tone Mapping Image.

IV. HDR10+ WORKFLOWS

The HDR10+ workflow is designed to have minimal impact on existing HDR capture, encode, distribution and display infrastructure. HDR10+ dynamic metadata is generated on captured or mastered HDR content and stored in the video stream during or after encoding through low complexity SEI (Supplemental Enhancement Information) messages and is backwards compatible with HDR10 capable devices.

HDR10+ is also supported in several commercial hardware and software encoders for use in both live and off-line workflows. This flexibility allows content owners to update their catalogs easily and rapidly to the latest dynamic metadata standard and take full advantage of dynamic metadata in their live broadcasts.

Tone mapping is applied frame by frame according to video content features and display capabilities allowing it to be fully capable of portraying optimized visuals with consistent color saturation and detail improving user's experience.

These concepts are detailed in the following topics.

A. Live Workflow

HDR10+ live encoding is easily executed by encoders, which follows workflow shown in Figure 10.

Real time broadcast operations are supported at the point of transmission enabling the flexibility to transmit the video stream either with no metadata or with dynamic metadata. In the second case, metadata is generated and inserted directly into the video stream in parallel with video encoding adding no additional latency to the overall system

The HDR10+ capable encoder will calculate the statistics of the input frame, create the tone curve for the current one, injects into the encoded frame and it can be output as an ES (elementary stream). Such ES is transmitted over terrestrial networks, as shown in Figure 11, and the embedded HDR10+ dynamic metadata is extracted and interpreted in a compatible display to provide scene-by-scene tone mapping to the HDR signal.

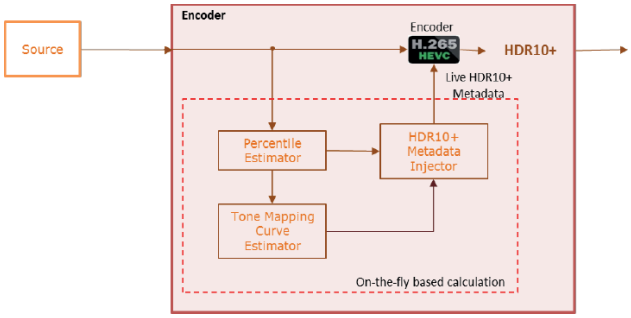


Figure 10 - HDR10+ Encoder for Live Broadcast Environment. Taken from Reference [4].

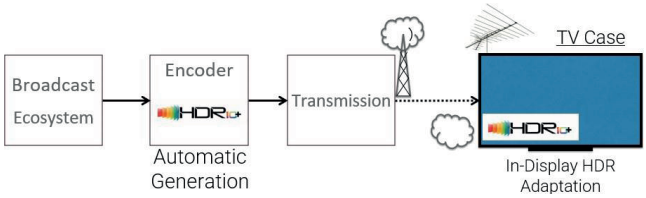


Figure 11 - HDR10+ Live Broadcast Workflow.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.2. Web Link: http://dx.doi.org/10.18580/setijbe.2021.2

B. *Off-line Workflow*

In an off-line workflow, HDR10+ metadata is generated on HDR mastered content. For UHD Blu-ray discs, the HDR10+ metadata is inserted in the encoded file during disc authoring. When inserted in an HDR10+ capable UHD Blu-ray player, the metadata is then extracted from the encoded stream and transmitted through the HDMI channel with Vendor Specific Info-frames to the HDR10+ capable display, where the dynamic tone mapping is then applied to the input signal, as depicted Figure 12.

The same process is followed for streaming content or if the TV set receives HDR10+ content directly. The TV detects the presence of HDR10+ metadata in the SEI message of the input stream and applies the tone mapping automatically.

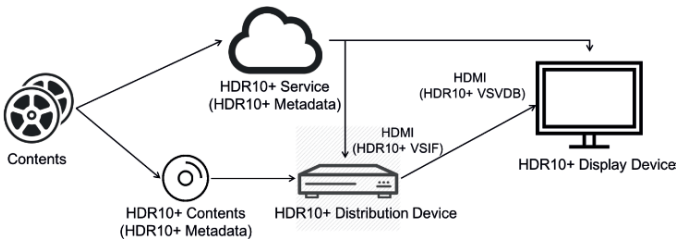


Figure 12 – HDR10+ Off-line Workflow. Taken from Reference [4].

C. *Ecosystem Deployment*

Furthermore, HDR10+ is an open-source, royalty free technology allowing it to be easily adopted and deployed across all devices in an HDR10 ecosystem with very low overhead and cost [4].

The HDR10+ ecosystem is rapidly growing with continued support from major device manufacturers and content creators. Since the creation of the HDR10+ Display Certification Program in 2018 until today, a dozen global display manufacturers have already produced over 4,700 different models of HDR10+ certified displays available to consumers worldwide (more details in section V).

Along with displays, there are also over 140 certified mobile devices, both tablets and smart phones, that support HDR10+ playback and in some cases HDR10+ capture as well allowing for rapid proliferation of HDR10+ content generated by amateurs and enthusiasts.

Professional/theatrical content is abundant and readily available from 5 major Hollywood studios on both UHD Blu-ray discs and through streaming OTT (Over-the-top) services with thousands of hours of episodic and theatrical pieces utilizing HDR10+ dynamic metadata for consumers to enjoy.

D. *HDR10+/HDR10 Backwards Compatibility*

HDR systems have been deployed since 2015 [2], when studios mastered video contents using PQ curve. HDR10 spec arose as one of the firsts systems, which use this transfer function and static metadata to improve image quality. ST2094-40 dynamic tone-mapping technology is advanced from HDR10, adding dynamic metadata for the same purpose. Such evolution concept is synthetized in the Figure 13.

As part of the HDR10 ecosystem, HDR10+ content works seamlessly with both HDR10 and HDR10+ devices. This content will display on HDR10 devices, which simply ignore the dynamic metadata while HDR10+ capable devices can take advantage of the dynamic metadata. HDR10 contents are totally supported by HDR10+ devices, as Figure 14.

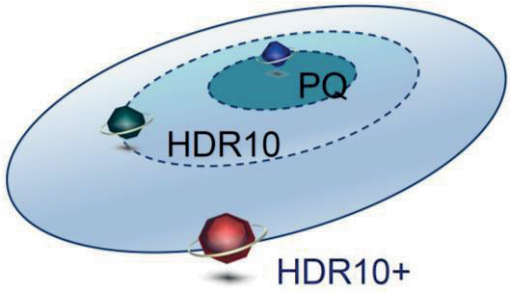


Figure 13 – HDR evolution from PQ to HDR10+.

Figure 14 – HDR10+/HDR10 Backwards Compatibility- Taken from Reference [4].

E. *Device Tone Mapping*

HDR10+ displays will process the video, frame by frame, with HDR10+ metadata to apply the best tone mapping for the content on the device, as the diagram of Figure 15.

Tone mapping happens when the source content peak luminance is higher than the display device peak luminance. Once HDR10+ content is delivered to a display device, the decoder will parse the HDR10+ metadata and video essence and process through the video pipeline. If a display doesn't support the HDR10+ metadata, it will be simply ignored as any other optional ITU-T T.35 metadata and the display reverts to show the strict version of HDR10.

The existing HDR technology, HDR10, leads to inconsistent reproduction of HDR content from one display device to another as only limited static metadata for content can be provided.

HDR10+ provides articulated scene based statistical data and optional guided tone mapping information to the display. This data enables a consistent reproduction of the source master content across displays of varying capability. Additionally, when a knee point is included with HDR10+ metadata, displays maintain the content's original look in shadow detail as no tone mapping happens below the knee-point [4].

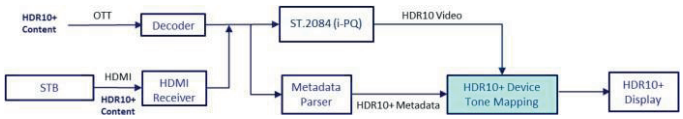


Figure 15 - HDR10+ Device Tone Mapping. Taken from Reference [4]

V. HDR10+ IN PRACTICE

This section is composed by three different sub-sections. First, introduces certification program promoted by HDR10+ Technologies LLC. Second, presents the standardization efforts by international SDOs (Standards Developing Organizations). The last sub-section brings the results of subjective evaluation tests made by a third-party lab.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.2. Web Link: http://dx.doi.org/10.18580/setijbe.2021.2

A. *Certification Program Available*

In order to promote HDR10+ technology, HDR10+ Technologies LLC [4] was created. The entity administers the license and certification program for products that want to adopt the technology, and provides the technical and test specifications, and certified logo, shown in Figure 16.

Royalty-free SW development kits and cloud CLI tools available for implementing this solution.

Nowadays, on the content production and distribution side, several studios, content distributors and equipments vendor adopted the technology. A current list of HDR10+ adopters can be found on the HDR10+ website, https://hdr10plus.org/adopters/.

Figure 16 - HDR10+ Logo. Taken from Reference [4].

B. *Standardization*

Besides SMPTE 2094-40 spec, HDR10+ technology, from metadata generation to decoding, is fully standardized through multiple SDOs, adopted by DVB and ATSC 3.0 DTV systems, being available to be included in any broadcasting plan as below.

- SMPTE ST 2094-40:2020 [3] – describes how the technology works.
- CTA-861-H (2020) [12] – describes how digital audio/video signals can be sent from one device to another.
- CTA-5001-C (2020) [12] – deals with encoding and packaging of segmented media for delivery and decoding on end user devices in adaptive multimedia presentations.
- SCTE 215-1-1 2020b (Cable TV) [14] – specifies how HDR10+ is used in cable video services applications.
- DVB Blue Book (version A001r17) [15] – specifies how to use ST 2094-40 in DVB system.
- ATSC 3.0 A/341 [15] – specifies how to use ST 2094-40 in ATSC 3.0 system.

C. *Subjective Evaluation*

Following ITU-BT.500 recommendation [16], a subjective performance evaluation [17] of 2094-40 dynamic tone mapping vs BT.2390 [18] representing tone mapping with static HDR metadata was performed by independent third party entity with the following three goals:

1. 2094-40 dynamic tone mapping and static tone mapping based on BT.2390 adaptation were each compared to the original HDR content on the same professional monitor.
2. The subjective overall quality of dynamic tone mapping using 2094-40 is compared against static tone mapping using BT.2390 adaptation, to verify the quality benefit of dynamic tone mapping on the same professional monitor.
3. The subjective overall quality of dynamic tone mapping using 2094-40 is compared against static tone mapping using BT.2390 adaptation, to verify the quality benefit of dynamic tone mapping on the same consumer monitor.

A total of 34 unique video clips with the following variants were used:

A. Original source content in PQ, mastered at 1000 nits.

B. Original source content in PQ, mastered at 4000 nits.

C. HDR clips with static metadata and a curve applied from BT.2390, tone mapping from a 1000 nit source to 400 nits.

D. HDR clips with dynamic metadata and a curve applied from 2094-40, tone mapping from a 1000 nits source to 400 nits.

E. HDR clips with static metadata and a curve applied from BT.2390, tone mapping from 4000 nits source to 400 nits.

F. HDR clips with dynamic metadata and a curve applied from 2094-40, tone mapping from a 4000 nits source to 400 nits.

Clips share the properties indicated in Table I:

TABLE I
CLIP PROPERTIES.

Property Value	Property Value
Resolution	3840x2160
Frame rate	24000/1001 Hz
Chroma sampling	4:2:2
Color bit depth	10-bits per color
EOTF	PQ
Color model	YcbCr
Encoding type	ProRes 4444 HQ
Color space	BT.2020 container (limited to P3)
Container	Quicktime
Range	Legal

Taken from Reference [17].

Three test scenarios were identified in order to effectively evaluate static HDR tone mapping using BT.2390 vs 2094-40 dynamic HDR tone mapping, as below:

- Test Scenario I - Evaluation of HDR tone mapping subjective quality (HDR artistic intent preservation): comparing original vs tone mapped result with two identical reference displays (Sony X300).
- Test Scenario II - Evaluation of HDR tone mapping subjective quality for 4000 nits content on two identical reference displays (Sony X300).
- Test Scenario III - Evaluation of HDR tone mapping subjective quality for 4000 nits content on two identical and limited performance consumer displays (Samsung NU8000).

C.1 *Test Scenario I*

Static and dynamic HDR systems are evaluated: both tones mapping a 1000 nits source image to 400 nits.

Viewers judge the visual quality difference between the source reference image (A) and the tone mapped content (C and D) as follows:

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.2. Web Link: http://dx.doi.org/10.18580/setijbe.2021.2

- Original graded source content vs. (C) BT.2390.
- Original graded source content vs. (D) 2094-40.

Test result aims to identify which (if either of the two) tone mapping methods best preserves the appearance of the original source content.

Video Clips used are identified in Table II:

TABLE II
TEST SCENARIO I SEQUENCE – 1000 NIT SOURCE TONE MAPPED TO 400 NITS

Clip #	Clip Name	Standard
1	Seattle Waterfront	BT. 2390
2	Cakes	BT. 2390
3	Rock climbers 1	2094-40
4	Seattle Waterfront	Identical – Control Clip
5	Cactus	2094-40
6	Boys Face/helmet	2094-40
7	Rock climbers 1	BT.2390
8	Woman Portrait	2094-40
9	Cactus	BT.2390
10	Cakes	2094-40
11	Woman Portrait	BT.2390
12	Boys Face/helmet	BT.2390

Taken from Reference [17].

- *Results*

The Table III and Figure 17 present the mean opinion score (MOS) and the 95% confidence interval (CI) for each clip observers were presented with. The control clip (identical on both displays) was voted highly and significantly higher than the BT.2390 rendition for the Seattle Waterfront clip which indicated the observer's ability to identify the difference. Results for 2094-40 were higher in each clip over BT.2390.

The Figure 18 illustrates the density of observers scoring as a single combined score sheet. Larger and darker circles denote greater frequency of observer selection.

TABLE III
AVERAGE MEAN OPINION SCORE (MOS) RESULTS OF TEST SCENARIO I WITH AVERAGE 95% CONFIDENCE INTERVAL (CI)

	Average MOS	Average CI
BT. 2390	47.94	9.81
2094-40	78.67	7.18

Taken from Reference [17].

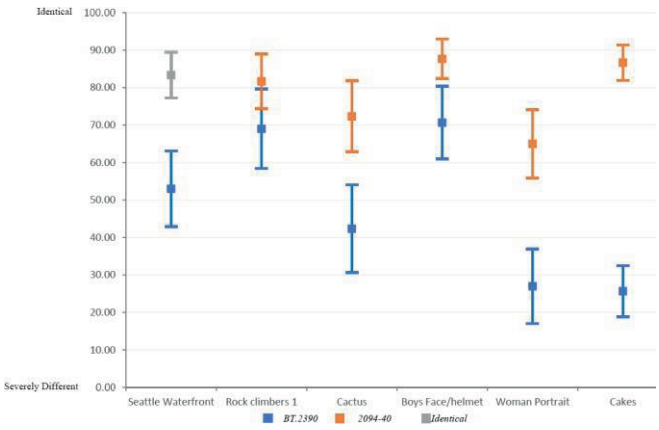


Figure 17 - Mean Opinion Score (MOS) Results of Test Scenario I with 95% Confidence Interval (CI). Taken from Reference [17].

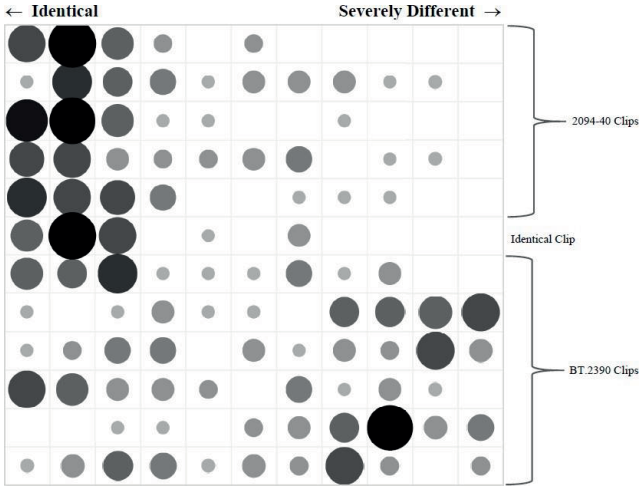


Figure 18 - Mean Opinion Score (MOS) Results of Test scenario I - Observer Scoring Sheet. Taken from Reference [17].

C.2 Test Scenario II

It is compared static vs. dynamic tone mapping of a 4000 nits source image (B) to 400 nits using identical reference monitors and judge which of the two has the better overall visual quality:

- (E) BT.2390 vs (F) 2094-40.

Test result aims to identify which (if either of the two) tone mapping methods has the better subjective overall picture quality.

Video Clips used are identified in Table IV:

TABLE IV
TEST SCENARIO II SEQUENCE – 4000 NIT SOURCE TONE MAPPED TO 400 NITS

Clip #	Clip Name	Left Standard	Right Standard
1	Airplane takeoff	BT. 2390	2094-40
2	Donkeys	BT. 2390	2094-40
3	Rock climbers 2	2094-40	BT.2390
4	Alpacas	2094-40	BT.2390
5	Water Tower	2094-40	BT.2390
6	Canyon Village	BT. 2390	2094-40
7	Fire Breather 1	2094-40	BT.2390
8	Swamps	BT. 2390	2094-40
9	Inca Ruins	2094-40	BT.2390
10	Village Docks	Identical	Identical
11	Girl Swimming	BT. 2390	2094-40

Taken from Reference [17].

- *Results*

The Table V and Figure 19 detail MOS and 95% CI for each clip observers were presented with. The control clip (identical on both displays) "Village Docks" was clearly identified as the "same" in terms of picture quality in observer voting. The 2094-40 renditions were consistently voted higher than BT.2390. Some 2094-40 renditions were voted significantly higher than their BT.2390 counterparts and others were voted with a smaller improvement.

The Figure 20 illustrates the density of observers scoring as a single combined score sheet, showing a favoring of the 2094-40 renditions along with a clear identification of the control clip.

TABLE V
MOS AVERAGE OF TEST SCENARIO II AND 95% CI ACROSS ALL CLIPS

Average MOS	Average CI
26.55	5.52

Taken from Reference [17].

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.2. Web Link: http://dx.doi.org/10.18580/setijbe.2021.2

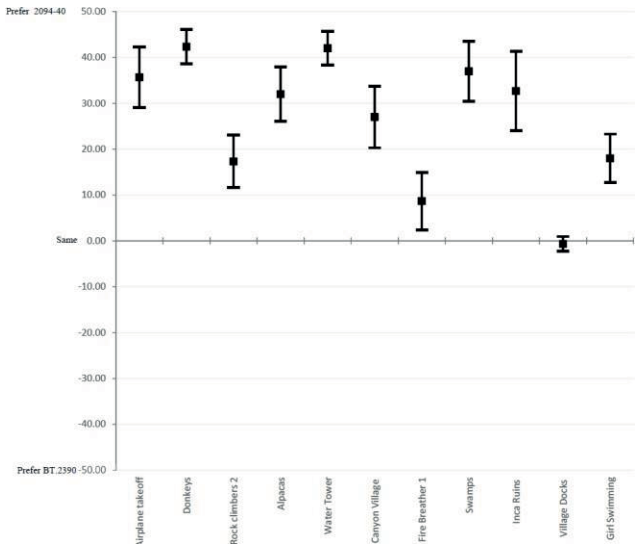


Figure 19 - MOS Results of Test Scenario II with 95% CI. Taken from Reference [17].

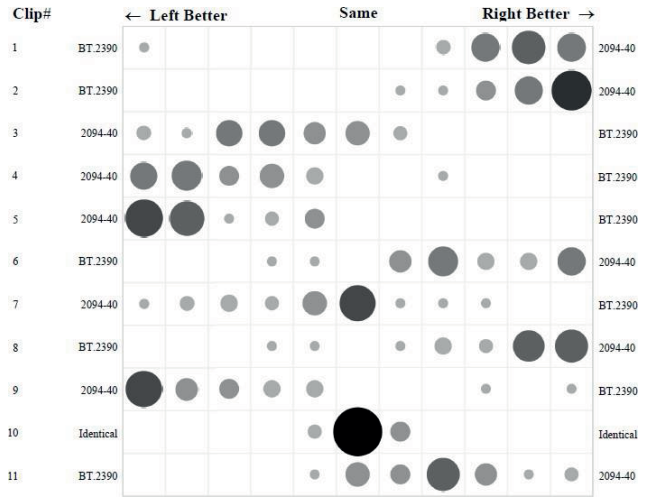


Figure 20 - MOS Results of Test Scenario II - Observer Scoring Sheet. Taken from Reference [17].

C.3 Test Scenario III

It is compared static vs. dynamic tone mapping of a 4000 nits source image (B) to 400 nits and judge which of the two has the better overall visual quality using limited performance consumer TV:

- (E) BT.2390 vs (F) 2094-40.

Test result aims to identify which (if either of the two) tone mapping methods has the better overall picture quality using limited performance consumer displays

Video Clips used are identified in Table VI:

TABLE VI
TEST SCENARIO III SEQUENCE – 4000 NIT SOURCE TONE MAPPED TO 400 NITS

Clip #	Clip Name	Left Standard	Right Standard
1	Girl Swimming	2094-40	BT. 2390
2	Inca Ruins	BT. 2390	2094-40
3	Rock climbers 2	2094-40	BT. 2390
4	Water Tower	2094-40	BT. 2390
5	Airplane takeoff	BT. 2390	2094-40
6	Canyon Village	2094-40	BT. 2390
7	Swamps	BT. 2390	2094-40
8	Fire Breather 1	BT. 2390	2094-40
9	Donkeys	BT. 2390	2094-40
10	Village Docks	Identical	Identical
11	Alpacas	2094-40	BT. 2390

Taken from Reference [17].

- *Results*

The Figure 21 and Table VII detail the mean MOS and the 95% CI for each clip observers were presented. Like test scenario II the control clip (identical on both displays) "Village Docks" was clearly identified as the "same" in terms of picture quality in observer voting as well as the 2094-40 renditions being consistently voted higher than BT.2390.

The Figure 22 illustrates the density of observers scoring as a single combined score sheet, showing a favoring of the 2094-40 along with a clear identification of the control clip in the center.

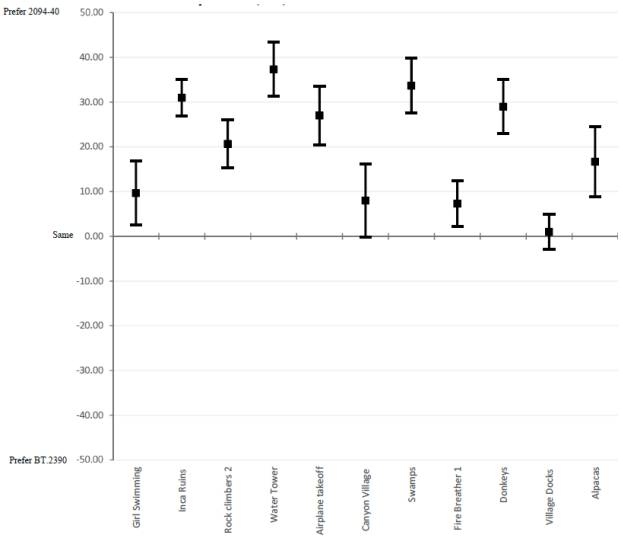


Figure 21- MOS Results of Test Scenario III with 95% CI. Taken from Reference [17].

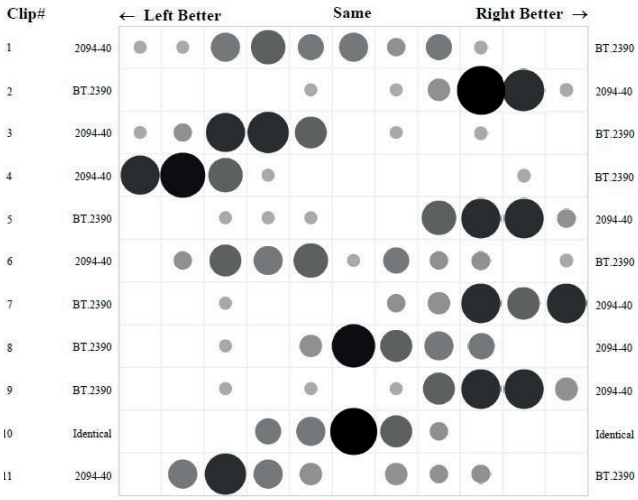


Figure 22- MOS Results of Test Scenario III - Observer Scoring Sheet. Taken from Reference [17].

TABLE VII
MOS AVERAGE OF TEST SCENARIO III AND 95% CI ACROSS ALL CLIPS

Average MOS	Average CI
20.12	6.02

Taken from Reference [17].

C.4 Test Scenario II and III Comparison

The Figure 23 shows MOS comparison between test scenarios II and III, indicating a clear favoring of 2094-40 HDR format.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.2. Web Link: http://dx.doi.org/10.18580/setijbe.2021.2

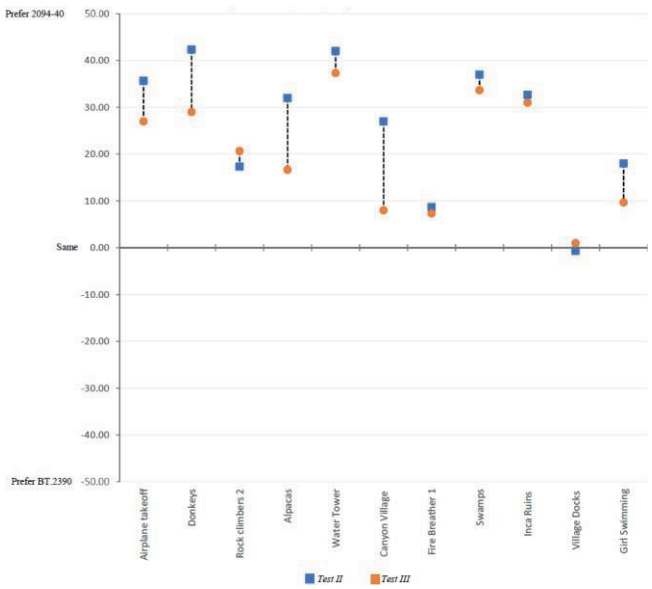


Figure 23 – MOS Comparison of Tests II and III Results. Taken from Reference [17].

VI. HDR10+ BENEFITS AND ADVANTAGES FOR TV 3.0

HDR10+ technology enhances image quality, providing a better user experience on consumer receiving displays. It complies with TV 3.0 requirements for video sub-component HDR Dynamic Mapping Codecs, and embraces the following characteristics [1], [3], [4]:

- As resolution agnostic technology, it can support any resolution used in TV 3.0 (2K/4K/8K, etc.).

- Able to be supported on any NAL based codec with support already in place for HEVC, VP9 and AV1.

- Bit representation: 10-bit or more (up to 16-bit), more smoothly showing color transition. TV 3.0 requires 10-bit as bit depth.

- Color space: WCG (ITU-R BT.2100/BT.2020) [6], [7], being able to reproduce more colors.

- 60 fps and lower frame rates demanded in TV 3.0 CfP. The 120 fps support is under development. The higher the frame rate, the smaller the blur presented in action content.

- Pixel representation: 10,000 nits per R/G/B, taking full advantage of PQ curve.

- Can be deployed over terrestrial broadcasting systems, as well as in broadband delivery, as expected by TV 3.0. Therefore, improving user's experience in the different ways TV content can be delivered.

- Its dynamic metadata are generated automatically (i.e., without operator intervention) and have been implemented in Live TV and VoD applications, being ready for all TV 3.0 signal transmission possibilities.

Moreover, HD10+ provides powerful advantages like [1] [3], [4]:

- Robust standardization: HDR10+ has been established as SMPTE standard ST 2094-40, broadly standardized by many international associations, and work continues on additional standards. Ratifying HDR10+ Technologies LLC commitment in continuously support HDR10+ technology.

- Well-established deployment across HDR ecosystem: streaming platforms, TV manufacturers, mobile device manufacturers, streaming devices, Blu-ray disc players, A/V receivers, SoC makers, professional tools, international video distributors, OTT providers and others, facilitating user's access to this technology.

- Royalty Free: it is open and royalty free solution, facilitating interoperability and broader adoption, which helps industry improve the technology. The only one free of various other dynamic tone-mapping technologies.

- 100% backward compatible with HDR10, allowing user to enjoy HDR experience even if it has only a HDR10 compatible receiver.

- To deliver all these technical benefits as a consistent user experience, HDR10+ Technologies, LLC was created. Since June 2018, the organization has been certifying HDR10+ compatible content, devices, tools, and services. This also helps promote consumer awareness.

- HDR10+ is an open standard that any manufacturer or content producer can embrace without paying hefty licensing fees.

A comparison among HDR10+ and other HDR technologies are shown in Table VIII:

TABLE VIII
HDR SYSTEMS TECHNICAL FEATURES COMPARISON

System	Transfer Function	Metadata	Bit Depth	Color Gamut Limit	Max. Range	Royalty Paid
HLG (Hybrid Log Gamma)	HLG (BT.2100)	No	8-bit, 10-bit	BT. 2020	100%	No
HDR10	PQ (ST2084/ BT.2100)	Static (ST2086)	10-bit	BT. 2020	10,000 nits	No
HDR10+	PQ (ST2084/ BT.2100)	Static (ST2086)/ Dynamic (ST2094-40)	10-bit up to 16-bit	BT. 2020	10,000 nits	No
Dolby Vision	PQ (ST2084/ BT.2100)	Static (ST2086)/ Dynamic (ST2094-10)	8-bit up to 12-bit	BT. 2020	10,000 nits	Yes
SL-HDR	HLG/PQ (BT.2100)	Static (ST2086)/ Dynamic (ST2094-20/30)	8-bit, 10-bit	BT. 2020	10,000 nits	Yes

The technology may offer essential benefits to all TV 3.0 ecosystem, as below [1], [3], [4].

A. Public (Consumers)

- HDR10+ is the most advanced dynamic metadata technology, offering enhanced video quality, allowing improved user experience on consumer displays.

- Subjective evaluation results indicate that HDR content tone mapped using 2094-40 can produce a closer reproduction of the reference image and better perceptible

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.2. Web Link: http://dx.doi.org/10.18580/setijbe.2021.2

overall image quality compared to HDR content tone mapped with BT.2390 [17].

- In the test scenario I results shows a consistently measured opinion that 2094-40 better represented the source image appearance to a lay viewer and is able to more closely maintain a perceptible visual similarity to the original source content than had it been tone mapped with BT.2390 [17].

- This is a royalty free technology, allowing the user to take advantage of its benefits, without additional cost to their equipment. It brings excellent scale representation for brightness and darks. HDR10+ can enable the creative intent to be expressed precisely, resulting in a consistent content reproduction with a more realistic, exciting, captivating entertainment experience.

- Growing industry support. Over 100 companies have already signed on as HDR10+ adopters, and the support from movie studios, streaming companies and list of device manufacturers is growing. Real opportunity for the user to appreciate its favorite content with enhanced video quality.

B. Broadcasters

- Simple program production. content creators, such as broadcasters, can focus on making the best HDR content, knowing that HDR10+ consumer devices will provide optimum performance. HDR10+ mastering is straightforward and is supported by professional production tools, which also make it simple to upgrade previously produced HDR10 content to HDR10+. In summary, HDR10+ workflow is simple, offering no significant impact to broadcasters.

- Provides better qualified metadata than simply HDR10 and total system compatibility.

- HDR10+ identifies the most important areas of each scene to improve the reproduction of those areas, retaining details in the highlights and shadows across all types of display capabilities.

- Expanding list of HDR10+ content available to consumers from major content providers on both UHD Blu-ray discs and streaming services.

- Allows broadcaster to improve audience experience, offering latest video content with enhanced video quality to a huge number of displays, which supports the technology.

C. Industry

- Technology well established in the marketplace, which improves the reliability and longevity of the technology.

- Certification program launched by HDR10+ LLC, which provides further quality assurance, allowing certification logo only for products that pass the rigorous testing requirements leading to playback compliance.

- Adoption based on open technology. Ease of access to anyone who wants to implement the technology. Also, popular professional and open-source tools available for implementing HDR10+ solution (mastering, postproduction, encoding, authoring) on content and devices.

- Ease of implementation & cost effective for operation. HDR10+ works with a range of video codecs, requires no licensing fees to take full advantage of the brightness of each specific display technology.

- HDR10+ technology is standardized by international standardization bodies recognized by SBTVD Forum [1].

VII. CONCLUSIONS

Brazil is at the verge of starting the development of the next generation TV Broadcasting system (TV3.0), a disruptive system expected to go ‘on the air’ in upcoming years. The elaboration of standards is scheduled for 2022 and HDR Dynamic Mapping has been a key-point of TV 3.0, established to improve video technology, adequate to consumer electronics display evolution. Therefore, this paper initiates providing an overview of five elements with huge influence over image quality, HDR importance, working principals, main features and benefits.

Differences between static and dynamic tone mapping are also discussed.

HDR10+ technology has minimum impact for broadcaster workflows and can offer great advantages to the end user, improving image quality and user’s experience while viewing video content.

It is royalty and license free technology and is facing a rapid expansion making the technology available to SoCs, receiving display devices vendors and all broadcaster ecosystem (generation, distribution, and reception). This represents a technology easily accessible to adopters, and a growing commitment to delivering a premium HDR experience.

As a developing technology, ST2094-40 can be expanded to technologies in the future, increasing and improving the solutions offered to all workflow, developing tools, compatible displays, and contents catalog, therefore offering more HDR10+ video contents to final user, improving its experience.

HDR10+ Technologies LLC entity was introduced in this paper as responsible to promote the technology and administrates the certification program. As result, several companies have already adopted this technology and, huge gamma of encoding equipments, receiving devices, video codecs, content production, and SW tools support it.

ST 2094-40 metadata offers improved video quality, compared to static metadata tone mapping, as indicated by subjective evaluation tests executed by third part labs.

This paper also listed some advantages and benefits provided by HDR10+ for both: viewers, broadcasters, and equipment vendors.

To conclude, besides other solutions to be adopted for all the components of TV 3.0 architecture taking the DTV system to the next level, the HDR10+ incorporation may offer enhanced feature to broadcast television, enabling the viewer’s TV sets to advance seamlessly with the latest technology. HDR10+ is able to deliver the optimal viewing experience across a variety of TV models.

ACKNOWLEDGMENT

The authors would like to thank SIDIA DTV Lab and Samsung SRA for the support and opportunity.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.2. Web Link: http://dx.doi.org/10.18580/setijbe.2021.2

VIII. REFERENCES

[1] TV 3.0 Call for Proposal at Forum SBTVD website https://forumsbtvd.org.br/tv3_0/, access on Aug 30th, 2021.
[2] T. Ogura and P. Espinosa, "4K HDR Workflow: from Capture to Display," *2018 IEEE Broadcast Symposium (BTS)*, 2018, pp. 1-9, Oct 2018.
[3] *Dynamic Metadata for Color Volume Transform - Application #4*, document SMPTE Standard ST 2094-40:2020, pp.1-29, 16 May 2020.
[4] HDR10+ Technologies website https://hdr10plus.org/; access on Aug 30th, 2021.
[5] Vaz, R. A.; Alves, L. G. P.; Akamine, C., "Video Scalability Advantages for the next Brazilian Terrestrial Digital Television Generation (TV 3.0). *SET International Journal of Broadcast Engineering,* Dec 2020.
[6] *Parameter values for ultra-high definition television systems for production and international programme exchange*, document ITU-R BT.2020-2., 2015.
[7] *Image Parameter values for ultra-high definition television systems for use in production and international programme exchange*, document ITU-R BT.2100-2., 2018.
[8] *High Dynamic Range Electro-Optical Transfer Function of Mastering Reference Displays*, document SMPTE Standard ST 2084:2014, pp.1- 14, 29 Aug. 2014
[9] *Mastering Display Color Volume Metadata Supporting High Luminance and Wide Color Gamut Images,* document SMPT StandardST 2086:2018, pp.1-8, Apr 2018.
[10] Ikizyan, Ike, "HDR Dynamic Tone Mapping with Enhanced Rendering Control", *SID Symposium Digest of Technical Papers*, vol. 50, pp. 303- 306.
[11] *A DTV Profile for Uncompressed High Speed Digital Interfaces*, document CTA-861-H (2021), Jan 2021.
[12] *Web Application Video Ecosystem – Content Specification*, document CTA-5001-C, Dec 2020.
[13] *HEVC Video Constraints for Cable Television Part 1-1 HDR10 Coding*, document ANSI/SCTE 215-1 2020b, 2020
[14] *Specification for the use of Video and Audio Coding in Broadcast andBroadband Applications*, document DVB BlueBook A001r17, Feb 2021.
[15] *A/341:2019 Amendment No. 2, ST 2094-40*, ATSC A/341 Video – HEVC, Sep 2021.
[16] *Methodology for The Subjective Assessment of The Quality of Television Pictures*, Document Recommendation BT.500, Oct 2019.
[17] *Subjective Visual Evaluation of Dynamic HDR metadata system (2094-40) vs static HDR metadata system Verification Tests Report*, document BluFocus Quality Assurance Evaluation Report, Nov. 2018.
[18] *High Dynamic Range Television for Production and International Programme Exchange*, Document Recommendation BT.2390, Apr 2018.

Rodrigo Admir Vaz received the B.S., and M.S. degrees in telecommunications engineering from University of São Paulo (USP), Brazil, in 2003, and 2007, respectively. Currently, he is studying the Ph.D. degree in electrical engineering at Mackenzie Presbyterian University. He has been with SIDIA DTV Lab / SAMSUNG Visual Display (VD) Group since 2007, where he is currently senior engineer. His research interests include next generation terrestrial broadcasting system, High Dynamic Range (HDR), video scalability, broadband and broadcast integration.

Luiz Gustavo Pacola Alves received his B.Sc. and M.Sc. degrees in Electrical Engineering with emphasis in Computer from University of São Paulo (USP), São Paulo, Brazil, in 2004 and 2008, respectively. He has been working at SIDIA DTV Lab / SAMSUNG Visual Display (VD) Group since 2009, where he is currently Technical coordinator. He has experience in Standardization & DTT planning focused on Latin America technical harmonization of Digital TV development in regional standardization bodies. He also has experience in Technical Business Development related to pay TV and live streaming. He is member of ISDB-T International and Technical Module of the Brazilian Digital Terrestrial Television Forum (SBTVD Forum) with active participation at WGs level representing SAMSUNG, currently engaged on next generation terrestrial broadcasting system with emphasis on High Dynamic Range (HDR).

Steve Larson received the B.S. in Electrical and Computer Engineering with an emphasis in Communications Systems from the University of California, San Diego, in 2010. He works as a Technical Program Manager for Samsung Research America (SRA) in the Digital Media Solutions Lab. His areas of focus are high dynamic range (HDR), respective applications to streaming, gaming, live and cinematic content and with the 8K video ecosystem working with the 8K Association in the Technical Working Group to establish specifications and best practices around 8K video formats.

Received in 2021-08-31 | Approved in 2021-12-07

MPEG-H Audio System for SBTVD TV 3.0 Call for Proposals

Adrian Murtaza
Stefan Meltzer
Yannik Grewe
Nicolas Faecks
Mickael Raulet
Lucas Gregory

CITE THIS ARTICLE
Murtaza, Adrian; Meltzer, Stefan; Grewe, Yannik; Faecks, Nicolas; Raulet, Mickael; Gregory, Lucas; 2021. MPEG-H Audio System for SBTVD TV 3.0 Call for Proposals. SET INTERNATIONAL JOURNAL OF BROADCAST ENGINEERING. ISSN Print: 2446-9246 ISSN Online: 2446-9432. doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

SET INTERNATIONAL JOURNAL OF BROADCAST ENGINNERING - SET IJBE V.7, 2021, Article 3, 16p.
 / ISSN Print: 2446-9246 | ISSN Online: 2446-9432

MPEG-H Audio System for SBTVD TV 3.0 Call for Proposals

Adrian Murtaza[(1)], Stefan Meltzer[(1)], Yannik Grewe[(1)], Nicolas Faecks[(1)],
Mickael Raulet[(2)], Lucas Gregory [(2)]
[(1)] *Fraunhofer Institute for Integrated Circuits (IIS),* [(2)]*ATEME*

***Abstract*— Under the name "TV 3.0 Project", the Brazilian Terrestrial Television System Forum (SBTVD) has issued the Call for Proposals (CfP) for a next generation Brazilian digital TV system, in July 2020. The MPEG-H Audio system, based on the open international standard ISO/IEC 23008-3, has been proposed by Fraunhofer IIS, ATEME, the Digital Broadcasting Experts Group (DiBEG) and the Advanced Television Systems Committee (ATSC). This paper provides an overview of the MPEG-H Audio system and the TV 3.0 Project requirements for the audio component. The TV 3.0 Project specifies a detailed test and evaluation procedure for verifying the fulfillment of the requirements. With wide industry support, the MPEG-H Audio system brings immersive sound, advanced interactivity, and accessibility options, as well as advanced features like hybrid delivery, consistent loudness after user interaction, connectivity options for external sound devices and seamless configuration changes. The MPEG-H Audio proponents have submitted a complete production and broadcast real-time chain to the SBTVD Forum which demonstrates the most advanced features.**

***Index Terms* — 3D and Immersive Audio, Accessibility, ATSC 3.0, Audio Coding, Broadcast, Broadband, Emergency warning system, Hybrid, Immersive Sound, MPEG-H Audio, Next Generation Audio, Object-based broadcasting, Personalized Sound, SBTVD TV 3.0, Streaming, Virtual Reality, Augmented Reality**

I. Introduction

THE Brazilian Digital Terrestrial Television System Forum (SBTVD) issued, in July of 2020, a Call for Proposals (CfP) seeking input for Brazil's next generation Digital TV system under the name "TV 3.0 Project" [1]. Without the constraints of a backward compatibility requirement, the TV 3.0 Project is paving the way for an advanced and modern next-generation television system in Brazil. The SBTVD Forum has established a set of TV 3.0 requirements and use cases, covering six system components (Over-the-air Physical Layer, Transport Layer, Video Coding, Audio Coding, Captions, and Application Coding). The CfP was divided into two phases: Phase 1 required an initial submission from proponents identifying the candidate technology and providing basic information, while during Phase 2, the proponents were expected to submit a full specification of the candidate technology as well as hardware and software solutions for the feature evaluation.

In response to the SBTVD TV 3.0 Call for Proposals, Fraunhofer IIS, ATEME, DiBEG, and ATSC have proposed the MPEG-H Audio system for the audio component [1] and have provided a complete production and broadcast chain to the SBTVD Forum for Phase 2 feature evaluation. The MPEG-H Audio system is fulfilling all TV 3.0 requirements listed in the CfP and provides the most advanced feature set and use cases as detailed in this document.

This paper describes a snapshot of the MPEG-H Audio proposal to SBTVD TV 3.0 Project. It is structured as follows: given that MPEG-H Audio is an open international ISO/IEC standard, the MPEG standardization process and adoption in various worldwide application standards is briefly introduced. Then, existing production workflows using MPEG-H Audio for live broadcast and post-production are outlined. Finally, we describe how the MPEG-H Audio system fulfils the most challenging TV 3.0 requirements and ensure an easy transition from the existing ISDB-Tb broadcast system to the future based TV 3.0 system.

II. MPEG-H Audio System Introduction

MPEG-H Audio is the most advanced Next Generation Audio (NGA) system and based on an open international standard: ISO/IEC 23008-3, MPEG-H 3D Audio [2].

The MPEG-H Audio system provides more realism through sound from above and below as well as around the listener and an unprecedented degree of freedom to consumers for personalizing the audio experience. With its unique interactivity features, MPEG-H Audio offers viewers flexibility to actively engage with the content and adapt it to their own preferences. The easiest way to interact with the content is to select one of several predefined audio presentations, called Presets. Those are complete audio mixes with a descriptive label attached to them, for example "Default TV mix", "Dialog enhanced audio" or "Venue sound".

Furthermore, simple adjustments are possible, such as increasing the dialogue prominence in relation to other audio elements. Interested viewers can dive deeply into advanced scenarios, select certain audio elements of the audio mix and adjust these elements in level and/or position.

A menu displaying all personalization options is available on MPEG-H Audio enabled devices or applications for viewers to personalize their content using, for example, the remote control of the TV or the touch screen on a mobile device. With its innovative system design, the MPEG-H menu will automatically adapt to the content creator's intentions and only display the interactivity options currently

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

available.

MPEG-H Audio opens an entire new level of sound going beyond stereo and surround. With sound coming from above or below, a third dimension is added to the audio experience and lets listeners experience sound in a more realistic and natural way. Depending on the playback system, the soundscape can be extended with sounds from below like footsteps down on the floor completing the immersive experience.

Another unique feature of MPEG-H is its capability to adapt the playback of content to the capabilities of the playback device. With the built-in renderer and advanced dynamic range and loudness management, the content will always be reproduced in the highest quality and with the best user experience achievable on the device in use.

This feature eases the content creation process, as a single MPEG-H stream can deliver the content to all kinds of receiver and playback devices, from headphones to sound bars and discrete loudspeaker systems providing the best quality possible.

A. MPEG Standardization

Finalized in 2015, MPEG-H 3D Audio (ISO/IEC 23008-3 [2]) is the latest audio compression standard developed by Moving Pictures Experts Group (MPEG), following a defined, competitive and collaborative process with the participation of the world's leading experts in the field of audio coding technology. MPEG is a joint working group of the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC).

From the start of the MPEG-H Audio development, the goal was set to deliver the best possible experience with immersive sound as well as enabling advanced accessibility, interaction and personalization features in one solution, taking audio to the next level. The combination of highly efficient encoding technologies, the ability to represent audio content using three different formats (see below), as well as advanced loudness and Dynamic Range Control (DRC) management is the base of the MPEG-H Audio system. The core codec represents the latest evolution of the AAC codec family with the highest coding efficiency and flexibility in computational complexity. The ability to represent audio content in three formats — channel-based, object-based, and scene-based (HOA) – or even as a simultaneous mix of all three, enables maximum flexibility for content producers and allows the usage for all kind of program material and applications including VR/AR.

The use of audio objects introduces the option of interactivity and personalization for the end user into the creative's toolbox and enhances the user experience in an unprecedented way. At the same time, these new elements do not necessarily require large upgrades to the broadcasting infrastructure. The flexibility, the rich set of features, and the coding efficiency are also the reasons why MPEG has selected MPEG-H as the core audio codec for the next generation of audio standards currently under development in the MPEG-I Immersive Audio Call for Proposals [3].

MPEG standards are published by the International Organization for Standardization (ISO) and are publicly available. MPEG is also maintaining the standards and provides updates which bring corrections and clarifications of the standard's document as well as new technologies and adaptations to new market developments. These facts are key for broad adoption of the standards in the market and ensure that multiple implementations provide a competitive environment and bring forward the best products to serve market needs. To support the development of such products, MPEG provides compliance bit streams, and MPEG members help to create test suites for certain applications, e.g., the MPEG-4 AAC/HE-AAC test suites for ISDB-T receivers in Brazil. A detailed description of the newly introduced coding tools in MPEG-H Audio can be found in [6].

B. Profiles and Levels

MPEG standards are defined as a toolbox, including a wide range of tools. Different profiles are defined based on use cases and applications by selecting only the tools fitting best for the targeted application space. This is also the case for the MPEG-H 3D Audio standard and its profiles.

While profiles establish a subset of the tools, levels put additional constraints on the parameters of these tools, allowing a finer adaptation on certain use cases. The combination of profile and level finally determines the necessary processing power and memory requirements for a specific application.

The ISO/IEC 23008-3 3D Audio standard, defines the following audio profiles:

1) The High Profile – includes all tools of the standard and provides a complete set of features. The High Profile is a theoretical profile and a superset of the Low Complexity and Baseline Profiles.

2) The Low Complexity Profile – provides features for broadcasting, VR/AR and streaming applications (ISO/IEC 23008-3, subclause 4.8.2.1[2]).

3) The Baseline Profile – provides features for broadcasting and streaming applications (ISO/IEC 23008-3, subclause 4.8.2.5 [2]). The Baseline Profile is a subset of the Low Complexity Profile.

TABLE I
LEVELS AND THEIR CORRESPONDING RESTRICTION FOR BASELINE AND LOW COMPLEXITY PROFILES

Level	1	2	3	4	5
Max. sampling rate [kHz]	48	48	48	48	96
Max. number of core channels in compressed data stream	10	18	32	56	56
Max. number of decoder-processed core channels	5	9	16(*)	28	28
Max. number of channels in referenceLayout	5	9	16(*)	24	24

(*) The Baseline profile supports in Level 3 up to 24 objects if the additional complexity restrictions given in ISO/IEC 23008-3, subclause 4.8.2.5.2 [2]) are applied on the encoding process.

Table 1 provides an overview of the levels and their corresponding characteristics for Low Complexity and Baseline Profiles. A complete description can be found in ISO/IEC 23008-3, subclause 4.8.2 [2].

1) Low Complexity Profile

The Low Complexity Profile is a superset of the Baseline

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

Profile. It includes two additional coding tools for Higher-Order Ambisonics (HOA) and Linear Prediction Domain (LPD), which is irrelevant for the majority of Next Generation Audio broadcast and streaming applications.

The HOA path of the Low Complexity Profile uses tools for decoding and rendering the HOA signals, in addition to the channel and object signals. Implementation of these tools can be quite complex and can lead to doubling the implementation effort in comparison to the Baseline Profile. Similarly, the testing effort also increases as higher number of conformance tests streams and test cases need to be verified for compliance. The increased implementation and testing effort for the Low Complexity Profile is justified for applications that benefit from having the HOA format available, such as VR/AR use cases.

2) Baseline Profile

The Baseline Profile is tailored especially for today's broadcast and streaming use cases. Without the support for HOA and the Linear Prediction Domain (LPD) tools, it offers significantly reduced implementation and testing effort without limiting the capabilities for all significant broadcast and streaming applications. This makes the Baseline Profile the natural choice for Consumer Electronics (CE) manufacturers.

As the Low Complexity and Baseline Profile share the majority of the tools, the MPEG-H 3D Audio standard specifies a compatibility signaling mechanism (see ISO/IEC 23008-3, subclause 4.8.2.7 and Annex P [2]), which ensures that Baseline Profile bit streams can also be decoded by Low Complexity Profile decoders and vice versa. The signaling maximizes the amount of content, which could be handled by a decoder to the benefit of the users.

The MPEG-H 3D Audio Baseline Profile fulfils all mandatory requirements for SBTVD TV 3.0 and is also proposed for next-generation DTTB in Japan.

C. Subjective Quality Evaluation

The performance of the MPEG-H Audio system was carefully evaluated by MPEG and documented in two MPEG Verification Test Reports [4][5].

With more than 1 million subjective ratings, from 341 expert listeners at nine prestigious independent test labs around the world (Fraunhofer IIS, Sony, NHK, Gaudio, Nokia, Orange, Qualcomm, Dolby and ETRI), the MPEG-H 3D Audio standard is the most thoroughly tested NGA codec.

The "MPEG-H 3D Audio Baseline Profile Verification Test Report" [4] includes five listening tests assessing the performance of the Baseline Profile. The tests cover a wide range of bit rates as well as an exhaustive range of use cases (i.e., from 22.2 down to 2.0 channel presentations).

The statistical analysis of the test data resulted in the following conclusions:

1) Test 1: Ultra-HD Broadcast

The "Ultra-HD Broadcast" use case was evaluated using highly immersive audio material coded at 768 kb/s and was presented on 22.2 or 7.1+4H channel loudspeaker layouts. The results showed that the Baseline Profile easily achieves "ITU-R High-Quality Emission" quality at the tested bit rate for broadcast applications.

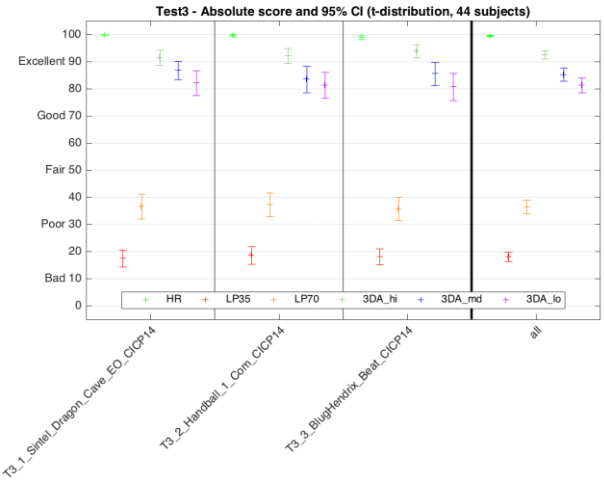


Fig. 1. Test 3 performance for 5.1+2H layout (CICP 14) immersive content at three different bit rate levels: low ("3DA_lo" - 144 kbps), mid ("3DA_md" - 192 kbps) and high ("3DA_hi" - 256 kbps). [4]

2) Test 2: HD Broadcast

The "HD Broadcast" or "A/V Streaming" use case was evaluated using immersive audio material coded at three different bit rates: 512 kb/s, 384 kb/s and 256 kb/s and presented on 7.1+4H or 5.1+2H channel loudspeaker layouts. The test showed that for all bit rates, Baseline Profile achieved a quality in the range of "Excellent" on the MUSHRA subjective quality scale.

3) Test 3: High Efficiency Broadcast

The performance for the "High Efficiency Broadcast" use case was evaluated using audio material coded at three different bit rates, with specific bit rates depending on the number of channels in the material. Bit rates ranged from 256 kb/s (5.1+2H) to 48 kb/s (stereo). The test showed that for all bit rates, the Baseline Profile achieved a quality in the range of "Excellent" on the MUSHRA subjective quality scale.

4) Test 4: Mobile

The performance for the "Mobile" use case was evaluated using immersive audio material coded at 384 kb/s and presented via headphones. The test showed that at 384 kb/s, Baseline Profile with binauralization achieved a quality in the range of "Excellent" on the MUSHRA subjective quality scale.

5) Test 5: High Quality Immersive Music Delivery

The "High Quality Immersive Music Delivery" use case requires delivery of object based immersive music to the receiver with up to 24 objects at high per object bit rates. This test used 11.1 (as 7.1+4H) as a presentation format, with material coded at a rate of 1536 kb/s. The test showed that at the bit rate of 1536 kb/s, Baseline Profile easily achieves "ITU-R High-Quality Emission" quality for high quality music delivery applications.

Fig. 1 shows the performance for 5.1+2H layout (CICP 14) immersive content at three different bit rate levels: low ("3DA_lo" - 144 kbps), mid ("3DA_md" - 192 kbps) and high ("3DA_hi" - 256 kbps).

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

The performance of the Low Complexity Profile of MPEG-H 3D Audio was assessed in the "MPEG-H 3D Audio Verification Test Report" [5]. Since Low Complexity Profile is a superset of the Baseline Profile, the BL verification test results [4] for channel- and object-based described above apply also to Low Complexity Profile. Additionally, the report [5] includes scores for HOA content in Tests 1 – 4.

The efficiency of the MPEG-H Audio codec allows carrying better quality and/or more channels with the same bit budget as currently used to carry only 5.1 channels. Thus, with the current commonly used broadcast audio data rate of 192 kbps, 5.1 surround channels with four additional height channels (i.e., 5.1+4H) can be delivered and that with improved subjective quality.

D. International Standards and Adoption

MPEG-H Audio has been widely adopted and included in various regional and international standards worldwide. Currently, MPEG-H Audio is the only Next Generation Audio system standardized for broadcast, streaming, hybrid and VR/AR/360-degree video streaming applications within 3GPP. The most notable application standards specifying MPEG-H Audio are:

- **ATSC:** The Advanced Television Systems Committee has successfully included MPEG-H Audio in its ATSC 3.0 suite of standards as ATSC Standard A/342-3 [7]. The corresponding transport layer signaling is specified in ATSC A/331 [8].
- **TTA (South Korea)**: The Telecommunications Technology Association (TTA) has selected MPEG-H Audio as the sole audio system for ATSC 3.0 in South Korea, as specified in TTAK-KO-07.0127 [9].
- **SBTVD (Brazil)**: The SBTVD Forum has adopted MPEG-H Audio [10] as part of the ABNT specification for TV 2.5 in Brazil, ABNT NBR 15602-2 [11]. The additional signaling for transport layer is specified in ABNT NBR 15603 [12] and the MPEG-H Audio receiver specification is provided in ABNT NBR 15604 [13].
- **3GPP**: 3GPP has selected MPEG-H Audio as the only audio format for 360° video streaming services over 5G within Release-15 of the specifications, TS 26.118 3GPP Virtual reality profiles for streaming applications [14].
- **DVB**: DVB has also selected and included MPEG-H Audio in the specification ETSI TS 101 154 v2.3.1 [15] defining the usage of audio and video codecs for DVB systems. The proper signaling for MPEG-2 TS DVB systems was specified in ETSI EN 300 468 [16].
- **ITU**: International Telecommunications Union (ITU) issued the recommendation ITU-R BS.1196-7 (01/2019) for Audio coding for digital broadcasting [17].

Additionally, all major OTT Specifications have adopted MPEG-H Audio, including MPEG CMAF, CTA WAVE, HbbTV, DASH-IF or DVB DASH.

The MPEG-H Audio system was the first Next Generation Audio codec worldwide to go on air 24/7 as South Korea launched its 4K UHD TV services using the ATSC 3.0 standard on May 31, 2017 [18]. The Korean standard [9] for this service mandates the MPEG-H Audio system as the only audio codec for delivery of immersive and personalized sound in South Korea. The Korean government timeline requires the percentage of native UHD content to increase in steps from 5% 2017 to 50% in 2025 and finally to 100% in 2027. In 2027, the HD service based on ATSC 1.0 and currently operated in a simulcast, will be completely switched off.

Fig. 2. UHD TV receiving MPEG-H Audio signal over ATSC 3.0 RF input and displaying native user interface for user interactivity

Professional equipment from encoders, metadata authoring and monitoring units, as well as test receivers are available. On the consumer side, TV sets are available on the market supporting the full feature set of the MPEG-H Audio system (see Fig. 2). Additionally, MPEG-H Audio is available for several years in immersive soundbars and AVRs. Thus, the complete end-to-end chain is available allowing broadcasters to make full use of the advanced features. Lessons learned during live broadcasts of major events using MPEG-H Audio are detailed in [19].

In Brazil, MPEG-H Audio was adopted as part of the TV 2.5 Project to enhance the audio experience over ISDB-Tb with immersive and personalized sound and is fully specified in the ABNT standards [10][11][12][13]. Fraunhofer IIS and its technology partners ATEME, Telos Alliance, SSL, EiTV and Sennheiser have showcased for the first time at the 2019 SET Expo trade show, a live ISDB-Tb broadcast local transmission using MPEG-H Audio in Brazil according to the TV 2.5 suite of standards.

Moreover, during one of the world's biggest music festivals, Rock in Rio [20], Rede Globo has successfully used MPEG-H Audio for a live terrestrial broadcast over the ISDB-Tb system. The musical performances on the two main stages, "Mundo" and "Sunset," were delivered over the air with MPEG-H immersive and personalized sound in the Rio de Janeiro area. The same MPEG-H Audio production was simultaneously delivered over multiple distribution platforms. Besides the ISDB-Tb terrestrial broadcast, an HLS streaming service using MPEG-H Audio was offered and with the support of Rohde & Schwarz, MPEG-H Audio was embedded in an experimental broadcasting UHF channel for the first 5G broadcast transmission field test in Brazil. ATEME's TITAN Live encoder created all three services in parallel from the same input.

Following the successful educational program of Fraunhofer IIS in South Korea and China, the first MPEG-H training center in Sao Paulo opened in February 2021 [21]. Fraunhofer has teamed up with Cinecolor Brazil to offer

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

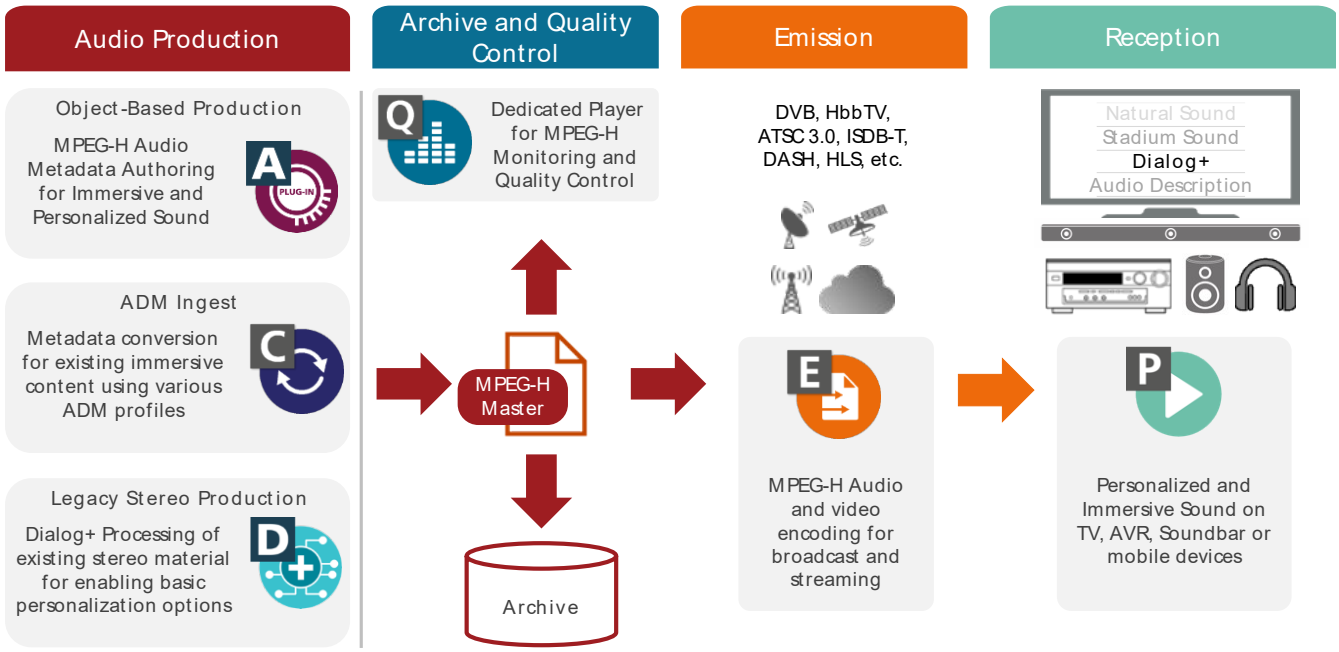


Fig. 3. MPEG-H Master usage in production workflows (simplified)

broadcasters access to the most advanced immersive audio technology.

III. Production Workflows

The production and transmission of MPEG-H Audio introduces new concepts compared to legacy production. The MPEG-H Audio system has been designed specially to explore this new creative options. Besides immersive 3D Audio, content creators can prepare mixes (including the default or main mix of the program) using authoring tools that specify an ensemble of gain and position settings for objects to create preset mix selections that can be presented on a simple menu to the user. All interactivity features offered to the users are strictly defined by the broadcaster during metadata creation. This process of generating metadata is called "authoring" and is the most important difference in production of MPEG-H Audio content compared to a legacy production.

Additionally, a different handling of audio elements is required when using immersive channel-beds and audio objects in production. A 3D-Audio bus structure needs to be available in the production tools, such as Digital Audio Workstation (DAW), broadcast mixing desk and monitoring paths. The additional audio objects must be kept separated from the other components such as *Music & Effects (M&E)* mixes – also called the channel bed – up to the authoring stage, where metadata is generated, including:

- Position information about object reproduction position in 3D space,
- Interactivity limitations for audio objects and presets,
- Loudness information about each component and preset,
- Text labels for presets and audio objects (also in multiple languages),
- Reference and target loudspeaker layouts and
- Many more.

In the following, an introduction to MPEG-H Audio production formats and workflows for live- and post-productions is provided.

A. 3D Audio Studio Recommendations

Jointly with leading industry experts in live production for sports and other major events, Fraunhofer IIS has published 3D Audio Studio Recommendations [22] specifying the main structural requirements and technical specifications for a 3D-Audio production environment. It details best practice for mixing and reproduction in a flexible manner for loudspeaker reproduction systems, including the most common setups such as 5.1+4H and 7.1+4H. The Studio Recommendations support the sound producers with additional guidance for room geometry and room acoustics, loudspeaker positioning and electroacoustic performance, 3D-Audio monitoring and mixing capabilities and provide recommendations for related literature.

B. MPEG-H Master

In the MPEG-H Audio System, all metadata is tightly coupled to the audio essence, ensuring the integrity of the transmitted or stored audio scene. This is achieved by using the "MPEG-H Master" which is a bundle of metadata and the audio content. The MPEG-H Master can be exported as either Broadcast Wave Format File carrying Audio Definition Model (ADM) metadata or MPEG-H Production Format (MPF) file including the metadata as MPEG-H Control Track (see below). Fig. 3 provides a high-level overview of production workflows using the MPEG-H Master format for new object-based productions, ingest of existing ADM metadata and legacy stereo production that can be enhanced with Dialog+ capabilities.

C. MPEG-H Production Format (MPF)

An MPEG-H Production Format (MPF) file is a multi-channel wave file which contains all the audio and metadata of the MPEG-H Audio scene. The metadata is modulated into a regular audio channel called "Control Track" (CT). This is a "time-code like" signal and has been introduced for efficient and robust usage of MPEG-H Audio in SDI-based workflows.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

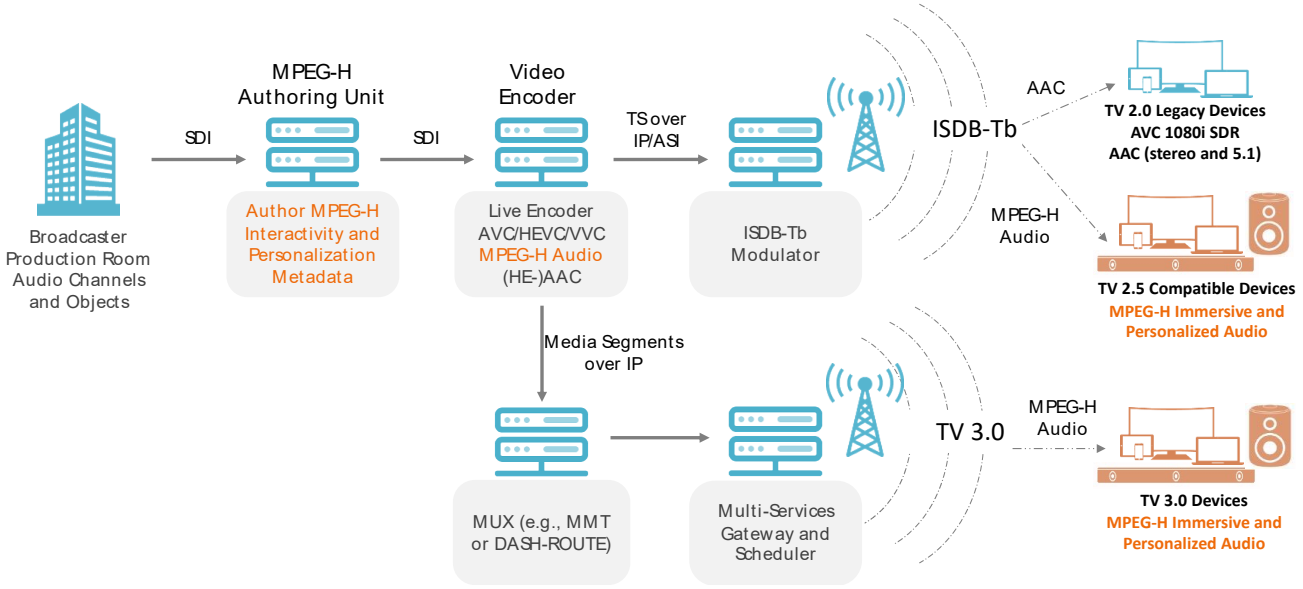


Fig. 4. MPEG-H Live Broadcast workflow (simplified)

The metadata for the audio signal is collected into packets synchronized with the video signal and is modulated with analog channel modem techniques into the Control Track, fitting in the audio channel bandwidth. This signal is unaffected by typical filtering, resampling, or scaling operations in the audio sections of broadcast equipment and ensures the synchronization of metadata with the corresponding audio and video signal. As the metadata contained in the Control Track is aligned to the audio and video data, any configuration change in live or post-production can be applied at every video frame boundary. Typically, the CT is carried on channel 16 within an SDI framework for live broadcast applications or on channel 16 of a multichannel wave-file.

The MPEG-H Control Track does not force audio equipment to be put into data mode or non-audio mode to pass through.

D. Audio Definition Model (ADM)

The Audio Definition Model (ADM) according to ITU-R BS.2076 [23] defines an open metadata format for production, exchange and archiving of Next Generation Audio (NGA) content in file-based workflows. Its comprehensive metadata syntax allows describing many types of audio content including channel-, object-, and scene-based representations for immersive and interactive audio experiences. A serial representation of the Audio Definition Model (S-ADM) specified in ITU-R BS.2125 [24] defines a segmentation of the original ADM for use in linear workflows such as real-time production for broadcasting and streaming applications.

It is acknowledged by ADM experts that application-specific ADM profiles are needed to achieve interoperability in ADM-based content ecosystems. Those ADM profiles incorporate the specific requirements for production, distribution and emission. To ensure interoperability with existing NGA workflows, applications adopting the ADM format should be able to convert native metadata formats to ADM metadata and vice versa such that artistic intent is preserved in a transparent way.

The MPEG-H ADM Profile [27] defines constraints on ITU-R BS.2076 [23] and ITU-R BS.2125 [24]. Those enable interoperability with established NGA content production and distribution systems for MPEG-H 3D Audio as defined in ISO/IEC 23008-3 [2].

E. Post-Production workflows

Typically, DAWs are used for audio post-production. Most common DAWs support the integration of AAX, VST or AU based plugins, which extend the capabilities of the host. Plugins such as the MPEG-H Authoring Plugin – part of the freely available MPEG-H Authoring Suite [25] – or the Spatial Audio Designer (SAD) [26] by New Audio Technology can be integrated seamlessly into the existing post-production workflows for MPEG-H audio content creation.

Different audio tracks from the DAW need to be arranged for the channel-based bed using a 3D panner. Furthermore, selected tracks are configured as audio objects. Compared to real time panning, post-produced content allows for more advanced object movements because automation can be edited and retrieved. The number of objects that move at the same time can be higher for the same reasons. Another benefit of using MPEG-H plugins is that they can overcome the DAWs bus width limitation, e.g., enabling a high number of objects to loudspeaker layouts not natively supported in the DAW, such as 5.1+2H, 5.1+4H or 7.1+4H.

As a next step during post-production, the MPEG-H metadata needs to be created. To monitor the different presets or switch groups that have been defined, a full MPEG-H renderer is included in the authoring tools. When the mix is finished and all metadata entries are authored, the session can be exported to an MPEG-H Master. During this stage, the loudness of all components and presets is measured and embedded into the corresponding metadata fields.

For the case that pre-existing content, e.g., a stereo or 5.1 audio mix, should be prepared for MPEG-H broadcast, the mix does not need to be touched. Only metadata need to be generated. For this, stand-alone tools such as the MPEG-H Authoring Tool (MHAT) – part of the MPEG-H Authoring Suite – are available. It is also possible to monitor the created scene and presets and export the final MPEG-H Master.

To control created MPEG-H Masters with or without an accompanying picture, the MPEG-H Production Format Player – part of the MPEG-H Authoring Suite – can be used

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

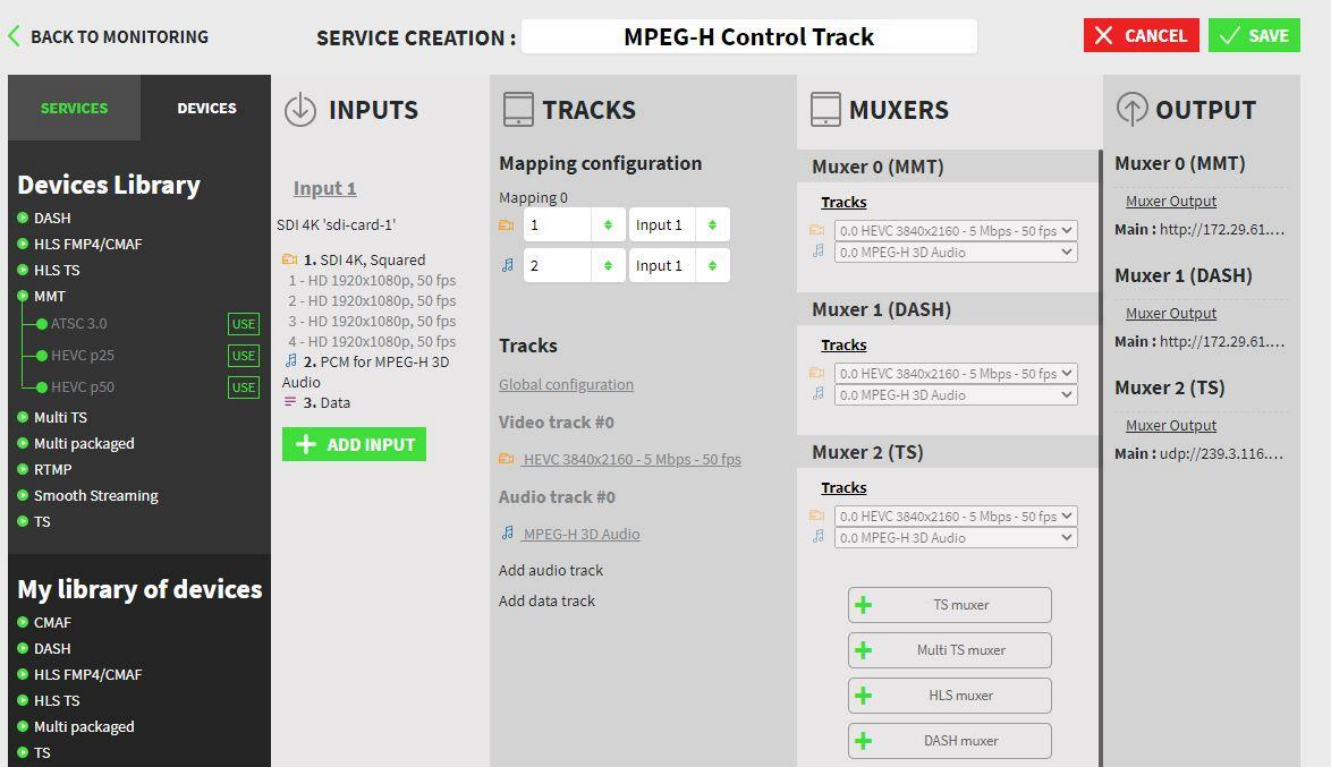


Fig. 5. ATEME Titan Encoder view, MPEG-H Audio single encoding for multiple outputs: MMT/DASH/TS

to ensure that the quality of the mix and the authoring is matching the expectations.

F. Live-Production workflows

The MPEG-H Audio system is designed to work with today's streaming and broadcast equipment using SDI-based workflows as well as with future IP-based infrastructure. In real-time scenarios, the authoring of MPEG-H Audio scenes and the metadata export is handled by a device class called "Authoring and Monitoring Unit" (AMAU).

AMAU systems are easy to integrate into the existing SDI, MADI or future IP based signal flows which are used for broadcast. Fig. 4 shows how AMAU units can be used in live broadcast infrastructure for existing ISDB-Tb TV 2.5 system in Brazil, where legacy receiving devices will decode the AAC signal, while newer receivers are able to provide an immersive and personalized MPEG-H experience. Additionally, the same AMAU can be used for enabling the future TV 3.0 broadcast in Brazil. With a single production and integrated audio and video broadcast encoders, MPEG-H can ensure a smooth transition from TV 2.5 to TV 3.0 with minimum investments for broadcasters.

Audio signals from the mixing console are fed into the AMAU using I/O converters. Within the AMAU, metadata creation and rendering takes place. The AMAU can be controlled using a web interface or hardware controller; typically, the output of the AMAU is an SDI signal including 15 channels of audio and the Control Track on channel 16. In a IP based production facility, the audio and metadata would be transmitted in a container over IP. A video signal can be passed through for visual reference and for synchronization purpose. Finally, the audio and video signals are provided to the emission encoder.

Currently, there are two AMAUs available and used for MPEG-H Audio production in live broadcast: the Multichannel Monitoring and Authoring system from Jünger Audio (MMA) and the Authoring and Monitoring System from Linear Acoustic (AMS). The use of AMAU systems allows productions to make use of all MPEG-H features without changing the entire workflow. Most of today's existing broadcast equipment can still be used.

Due to the advanced capabilities of the MPEG-H system, the monitoring stage during a production is important. Many different speaker layouts from stereo to 7.1+4H can be connected for 3D Audio playback and used for monitoring in an AMAU. Additionally, all interactivity options and the audio quality can be monitored during production using an emulation of end-user receivers with different reproduction configurations.

AMAU systems measure the loudness and true peak values of all channels, objects, output busses and formats, as well as every created Preset in real-time. With the resulting data, correction values are added to the metadata stream compliant with the applicable loudness regulation. The measurement of all generated DRC profiles and real-time loudness correction are also included. Additionally, AMAU production tools support the user with visualizations of all crucial measurement values.

As explained above, AMAU systems interface perfectly with currently deployed broadcast equipment, such as ATEME's Titan Live Encoder (see Fig. 5). This emission encoder includes Fraunhofer's MPEG-H 3D Audio encoder library, which can be either configured by the operator, using presets and encoding profiles up to 7.1+4H, or configured using a Control Track produced upstream by an AMAU and feed to the encoder through the SDI input. In the second scenario, a fallback configuration is also given to the MPEG-H Audio encoder and applied if no consistent Control Track is found in the input stream.

In contrast with the usual operation of Titan Live, or any other emission encoder, the use of the Control Track also permits to address advanced scenarios, such as dynamic

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

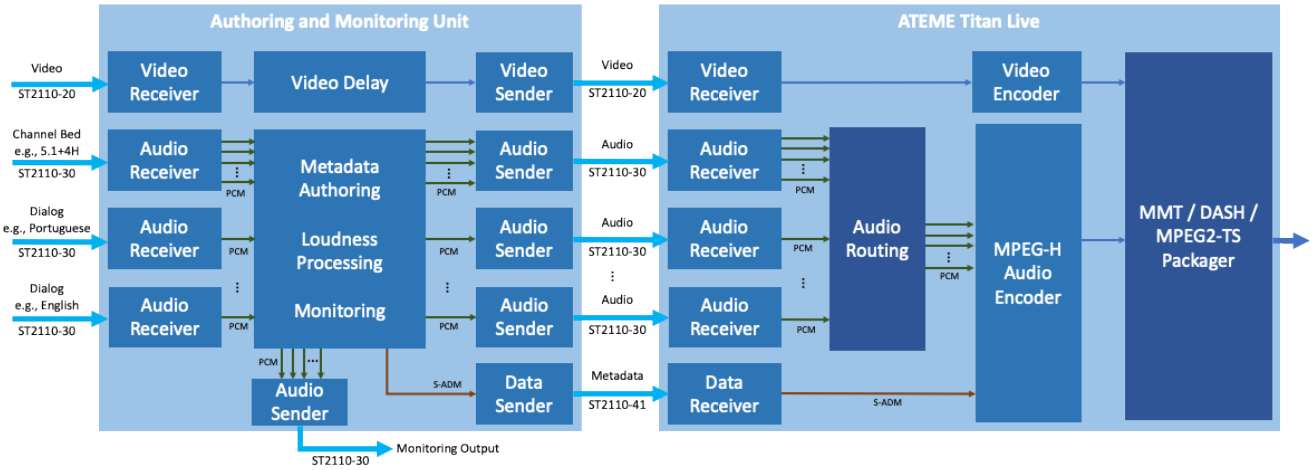


Fig. 6. MPEG-H Audio production workflow based on SMPTE ST 2110 (simplified)

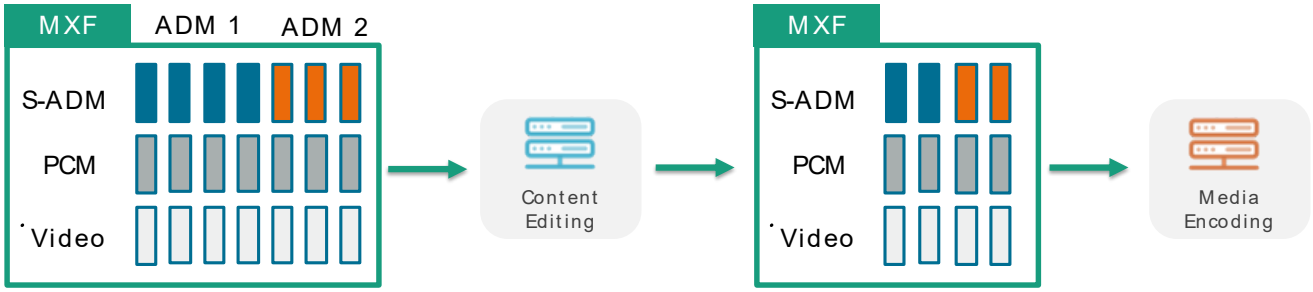


Fig. 7. Example MXF based workflow using ADM metadata

configuration changes for the audio encoder. For instance, the encoder can use a 5.1+4H channel-bed configuration while encoding a live sport event or a movie, and switch to a stereo layout during the commercial breaks. In this case, the configuration changes are triggered by the AMAU, and the MPEG-H Audio encoder seamlessly changes its configuration when a new configuration model is detected in the input Control Track.

Titan Live's muxer and packager unit adapts to these configuration changes by dynamically packaging the encoded audio stream. For DASH and MMT cases, fragmented MP4 (fMP4) segments have variable duration in order to ensure MPEG-H Audio Random Access Points (RAPs) keep aligned with the segments, and that a configuration change does not occur in the middle of a fMP4 segment. High-level signalization is also dynamically updated at configuration changes to reflect the encoded elementary stream's content. This is important for MPEG-H Audio since MPEG2-TS descriptors and fMP4 *SampleEntry* boxes contain part of the information needed to decode and present the content to the end-user. Such information may be used at receiver tune-in or startup and is essential that bitstream data is aligned to the information signaled on the transport layer for a high quality of experience on consumer side.

G. IP-based Production workflows

In legacy SDI and MADI based workflows, PCM audio essence is transmitted via audio channels or embedded audio channels of the SDI video signal and audio metadata is transmitted by means of the MPEG-H Control Track. In IP based workflows according to SMPTE ST 2110, media essences such as video, audio and metadata are transmitted over separate RTP connections and PTP (IEEE 1588, SMPTE ST 2059) is used to synchronize the different essence streams. The transmission of the PCM audio essence (SMPTE ST 2110-30) is already established whereas the standardization for the transport of metadata including serialized ADM metadata (SMPTE ST 2110-41) is still ongoing at the time of writing.

The serial representation of the Audio Definition Model (S-ADM) according to ITU-R BS.2125 defines a segmentation of the original ADM for use in linear workflows such as live production for broadcasting and streaming applications. Like the MPEG-H Control Track, one S-ADM frame contains a set of metadata describing at least the audio frame over the time period associated with that frame. S-ADM has the same structure, attributes and elements as those of ADM, as well as additional attributes to specify the frame format. The S-ADM frames are non-overlapping and contiguous with a specified duration and start time.

Fraunhofer is actively participating in the standardization of live production workflows based on S-ADM and SMPTE ST 2110 in following international standardization organizations: ITU-R, EBU, AES and SMPTE. As soon as the standardization is complete, Fraunhofer and its technology partners will provide and deploy complete solutions for authoring and monitoring of MPEG-H Audio for workflows based on S-ADM and SMPTE ST 2110 in IP-based production environments. For illustration purposes, a simplified ST 2110 Audio and Metadata over IP workflow for MPEG-H Audio is depicted in Fig. 6.

For storage and playout of S-ADM based content, Fraunhofer IIS is actively participating in the standardization of the transport of PCM audio essence and S-ADM metadata inside the Material Exchange Format (MXF, SMPTE ST 377-1), which is underway at the time of writing. The MXF Format is optimized for content interchange or archiving by creators and/or distributors and provides a complete

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

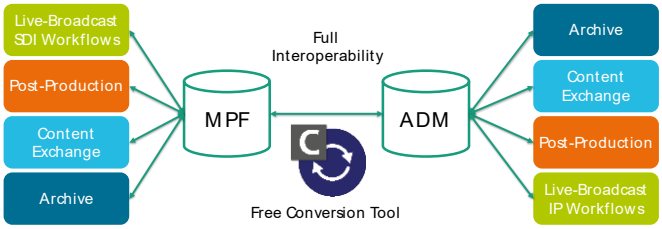


Fig. 8. Full Interoperability between MPEG-H Masters and ADM-based workflows

framework for the transport of NGA. An example application of MXF in practice is shown in Fig. 7.

H. Interoperability

The MPEG-H Conversion Tool [25] provides a lossless conversion between MPF and ADM based production formats. As shown in Fig. 8, the MPF format is a robust and reliable solution for existing SDI-based live workflows and it may be used for all other purposes such as post-production, content exchange and archiving. On the other hand, the ADM format is ideal for content storage, content exchange using MXF based workflows, and with its serialized option for IP-based workflows according to SMPTE ST 2110 suite of standards.

In the future, Fraunhofer IIS and its technology partners will provide tools for the lossless conversion between SDI and ST 2110 interfaces including MPEG-H Control Track and/or S-ADM metadata.

IV. SBTVD TV 3.0 Requirements

The SBTVD TV 3.0 Call for Proposals [1] specifies detailed requirements for each component or sub-component as well as detailed test procedures for verification and validation of the features for each candidate technology. The TV 3.0 audio coding requires – amongst others – support for:

- immersive and interactive audio,
- state-of-the-art coding efficiency,
- live (real-time) encoding with minimum end-to-end latency,
- audio description delivery in the same stream as the main audio,
- emergency warning information audio description,
- seamless and frame-accurate configuration changes or ad-insertion at any time instance.

Additionally, the audio coding shall enable a single delivery format for multiple audio playback configurations, consistent loudness across programs and inside the same program, frame-accurate audio/video synchronization, new immersive audio services, such as VR / AR / XR / 3DoF / 6DoF and extensibility. In the following sub-sections, few of these features are explained in context of the MPEG-H Audio system.

A. Immersive Sound

In MPEG-H, immersive sound can be carried in three primary ways: traditional channel-based sound where each transmission channel is associated with a loudspeaker position; sound carried through audio objects, which are positioned in three dimensions independently of loudspeaker positions; and scene-based (or Ambisonics), where a sound scene is represented by a set of coefficient signals that are the linear weights of spatial orthogonal spherical harmonics base functions.

As the only audio format natively supporting HOA, MPEG-H Audio can provide immersive sound using any combination of these three well-established audio formats.

B. Interactivity and Personalization

The MPEG-H Audio metadata, defined during authoring, carries all the information needed to allow viewers to change the properties of audio objects by attenuating or increasing their level, disabling them, or changing their position in the three-dimensional space. Additionally, the MPEG-H Audio metadata structures empower broadcasters to enable or disable interactivity options and to strictly set the limits to which extent a user can interact with the content.

The simplest use case is probably the most desired and powerful one, for which the dialog or commentary for a program is sent as an object. This allows the viewer to adjust the relative volume or "prominence" of the dialog relative to the rest of the audio elements in the program. In general, broadcasters attempt to mix the sound as a good compromise between dialog, natural sound, music, and sound effects. Viewer's preferences may vary, particularly as the (immersive) sound mix becomes more complex, such as in sports events or action dramas. Additionally, the reproduction setups at home are ranging from tiny building speaker to elaborated AVR controlled multi-speaker setups resulting in a huge influence on the user's experience. This simple case can be extended in offering two or more dialog objects for different languages or commentary oriented to each of the teams in a sporting event.

Moreover, the MPEG-H Audio metadata enables broadcasters to provide several versions of the content, as so-called "presets/preselections", which describe how all channels and objects signals are mixed together and presented to the viewer. Choosing between different presets is the simplest way to interact with the content. Advanced interactivity settings can be offered to more experienced users for manipulating objects individually.

C. Metadata and Broadcaster Control

The MPEG-H Audio system standardizes a rich metadata set to define an audio scene, the "Metadata Audio Elements" (MAE) as specified in ISO/IEC 23008-3, Clause 15 [2]. Each audio track with accompanying metadata is called an "audio element". This enables the most advanced and flexible end-user interactivity and personalization experience, while still offering full control over these features to the broadcasters. This is achieved with standardized metadata for controlling the personalization options such as setting the limits in which the user can interact with the content. The set of MAE metadata consists of:

1) Descriptive metadata: Information about the existence of objects inside the bit stream and high-level properties of audio elements, e.g., textual descriptions by labels, content kind and content language.

2) Control metadata: Information of how interaction is possible or enabled by the content creator.

3) Playback-related metadata: Information about special playback options.

4) Structural metadata: Grouping and combination of elements.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

Audio objects are associated with metadata that contains all information necessary for personalization, interactive reproduction, and rendering in flexible reproduction layouts. The metadata (MAE) is structured in several hierarchy levels. The top-level element of MAE is the "AudioSceneInfo". Sub-structures of the Audio Scene Info contain: "Groups", "Switch Groups" and "Presets".

1) Groups of Elements

The concept of an element group enables arranging related element signals that are to be treated together as a unit, e.g., for interactivity in common or for simultaneous rendering. A use case for groups of elements is the definition of channel-based recordings as audio elements (e.g., a stereo recording in which the two signals should only be manipulated as a pair). Grouping of elements allows for signaling of stems and sub-mixes by collecting the included element signals into groups that then can be treated as a single component.

2) Switch Groups of Elements

The concept of a switch group describes a grouping of components that are mutually exclusive with one another. It can be used to ensure that exactly one of the switch group members is enabled at a time. This allows for switching between, e.g., different language tracks or different commentators, when it is not desired to simultaneously enable multiple language tracks.

3) Presets

Presets can be used to offer combinations of groups and objects for more convenient user selection. Properties of the groups, like default gain or position can be set differently for each preset. It is not necessary to include all groups and objects in a preset.

4) Personalization and Interactive Control

Using the information in the MAE, the MPEG-H Audio system offers listeners the ability to interactively control and adjust various elements of an audio scene within limits set by broadcasters (e.g., to adjust the relative level of Dialog only in a range specified within the AudioSceneInfo structure).

The metadata allows for the definition of different categories of user interactivity as listed below:

- ***On-Off Interactivity***: The content of the referred group is either played back or discarded.
- ***Gain Interactivity****:* The overall loudness of the current audio scene will be preserved but the prominence of the referred signal will be increased or decreased.
- ***Positional Interactivity***: The position of a group of objects can interactively be changed. The ranges for azimuth and elevation offset, as well as a distance change factor can be restricted by metadata.

In order to reflect the content creator's intention to what extent their artistic intent may be modified, the interactivity definitions include minimum and maximum ranges for each parameter (e.g., the position can only be changed in a range between an offset of -30° and 30° azimuth).

D. Advanced Accessibility Options

Using object-based audio, MPEG-H Audio offers advanced and improved accessibility services (i.e., Dialog Enhancement and Audio Description) allowing hearing and visual impaired audience to experience at a new quality level.

1) Dialog Enhancement

MPEG-H Audio includes Dialog Enhancement (DE) for automatic device selection (prioritization) as well as for user manipulation. As an additional feature, MPEG-H Audio supports the personalization of Dialog Enhancement through a user interface offering the direct adjustment of the enhancement level, inside the range defined by the broadcaster. This range (e.g., minimum and maximum values) can be set differently for each audio object of the content.

Furthermore, MPEG-H Audio can also enable DE functionality for legacy stereo content (e.g., archive material stored in stereo without the original stems). This ensures a consistent user experience since the same functionality can be offered not only for object-based productions but also for existing stereo archive material.

The first and probably most important step is to obtain a "clean Dialog" version from the stereo content. This is achieved using a so-called “Dialog Separation” (DS) pre-processing technology. Several DS solutions are available on the market and can be used together with MPEG-H Audio. Relying on an open format such as BWF/ADM ensures interoperability of the MPEG-H system with existing and future solutions for Dialog Separation.

2) Audio Description

Similarly, the MPEG-H Audio system allows the delivery of Audio Description (AD) in multiple languages. AD services can be enabled by automatic device selection (prioritization) as well as by manual user selection. For each AD object, all advanced interactivity options are available and can be enabled by the broadcaster. The AD level can be adjusted independently and moreover, MPEG-H Audio is the only standardized audio system that allows the user to spatially move the Audio Description to a user selected position (e.g., to the left or right). This enables a spatial separation of main dialog and Audio Description, leading to a better intelligibility of the main dialog as well as of the Audio Description.

The system supports advanced connectivity use cases, where the Audio Description can be also provided via a Bluetooth channel for example to the headset of the person requiring the Audio Description. This person can now get same enhanced experience with the AD, while the rest of the family is experiencing the content without the AD. The standardized interfaces in MPEG-H Audio, ISO/IEC 23008-3 [2] can enable receivers and application layers to implement such advanced used cases.

3) Multi-language services

With existing audio codecs, multi-language programs are broadcasted as separate complete mixes in each language. Using one stream for each mix requires a high bit rate, directly proportional to the number of additional languages offered. Moreover, if Audio Description services have to be provided as additional complete mixes, the required bandwidth would significantly increase.

MPEG-H Audio enables a much more efficient way of offering accessibility and multi-language services by making use of object-based audio, similar to the DE feature, as described in previous sections. With a common channel bed and individual audio objects for each language dialog and audio description tracks, MPEG-H Audio requires a

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

Fig. 9. Example of multi-language labels (English – Upper side, French – Lower side).

significantly lower bit rate than legacy systems. For example, a 5.1 program is delivered in 5 different languages in a single stream using one audio object for each language. A legacy system would require transport of six complete 5.1 mixes in five different streams.

4) *Presentation of* services

The MPEG-H Audio metadata uses textual labels for describing the presets and the audio objects in multiple languages. The content creator can decide based on the regions where its content is distributed to author all labels in one or more languages. Based on the receiver's preferred language setting the correct labels will be displayed to the viewer. Fig. 9 shows the labels authored during a live broadcast trial in two languages: English and French.

Using the MPEG-H metadata, the content creators can ensure that their artistic intent and the various features they want to enable are correctly displayed to the user. In this way content creators are always in control of their content and the users will experience the content in the same way on all devices.

E. *Emergency Warning Information*

Delivery of Emergency Warning Information (EWI) caring emergency alerts, information and instructions to the TV viewers is a key feature of a next generation terrestrial broadcast system. The MPEG-H Audio system was defined including a flexible mechanism for emergency information messages. Multiple audio objects can be signaled as EWI using the MPEG-H Audio content type "mae_contentKind" == 12 ["emergency"], specified in ISO/IEC 23008-3 [2]. The EWI can be signaled as mandatory or optional, enabling the application layer in the receiver device to reproduce optional EWI messages based on the receiver settings or geo-location information. This way, according to the local requirements, the MPEG-H Audio system can be used to offer EWI Audio Description in different ways:

- **Mandatory emergency messages:** One audio object included in all presets and always active. The viewer cannot disable the EWI. Moreover, the EWI message may be delivered in a dedicated preset, replacing the main dialog or even replacing the complete mix
- **Optional emergency messages**: One audio object included in all presets and active based on receiver settings or geo-location of the receivers. These messages can be disabled by the viewer, and they are usually not critical alerts.

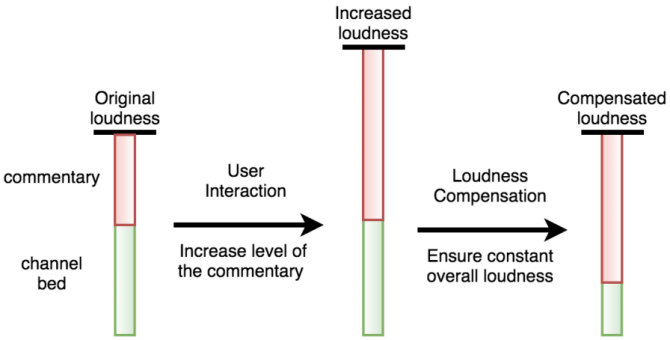


Fig. 10. Illustration of the loudness compensation concept for user interaction with the dialog object gain. The height of each bar corresponds to the loudness portion of the dialog object and the channel bed, respectively.

F. *Consistent Loudness*

The SBTVD TV 3.0 project requires a consistent loudness of the reproduced audio content. The MPEG-H Audio System accomplishes this automatically in two steps: the Loudness Normalization module aligns the loudness between program items to the target loudness of the decoder, while the Loudness Compensation module additionally compensates for loudness changes due to user interaction.

1) *Loudness Normalization*

The MPEG-H 3D Audio standard supports loudness information that is mandatorily included in the metadata of the MPEG-H Audio Stream. Various loudness measurement systems such as ITU-R BS.1770, EBU R128, ATSC A/85 are supported to fulfill applicable broadcast regulations and recommendations. The system allows to specify whether loudness information relates to the loudness of a full program or whether it refers to a specific anchor element of the program, such as the dialog or commentary.

Additionally, the system allows to input at encoding stage loudness information for each available preset separately. This enables immediate and automatic loudness control for interactive and personalized audio. For example, when the user switches between different presets the loudness normalization gain is instantaneously adjusted to ensure consistent playback loudness over all presets.

2) *Loudness Compensation*

The MPEG-H Audio system allows users to interact and control the rendering of individual audio elements which might result in an increase of the overall loudness of the resulting mix compared to the original preset authored in production. This behavior would interfere with the requirement of consistent loudness and preservation of signal headroom. Therefore, the MPEG-H Audio System includes a mandatory tool to compensate for loudness variations due to user interaction with individual audio elements (e.g., increase the dialog level compared to the rest of the mix).

The loudness compensation tool is based on metadata included in the audio stream that provides the loudness for each signal group or object that is part of the program mix. From these individual loudness values, a compensation gain is determined after any gain interaction done by the user, which is then applied together with the loudness normalization gain. The loudness compensation concept is illustrated in Fig. 10, for the example of a program consisting of a dialog object and a channel bed.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

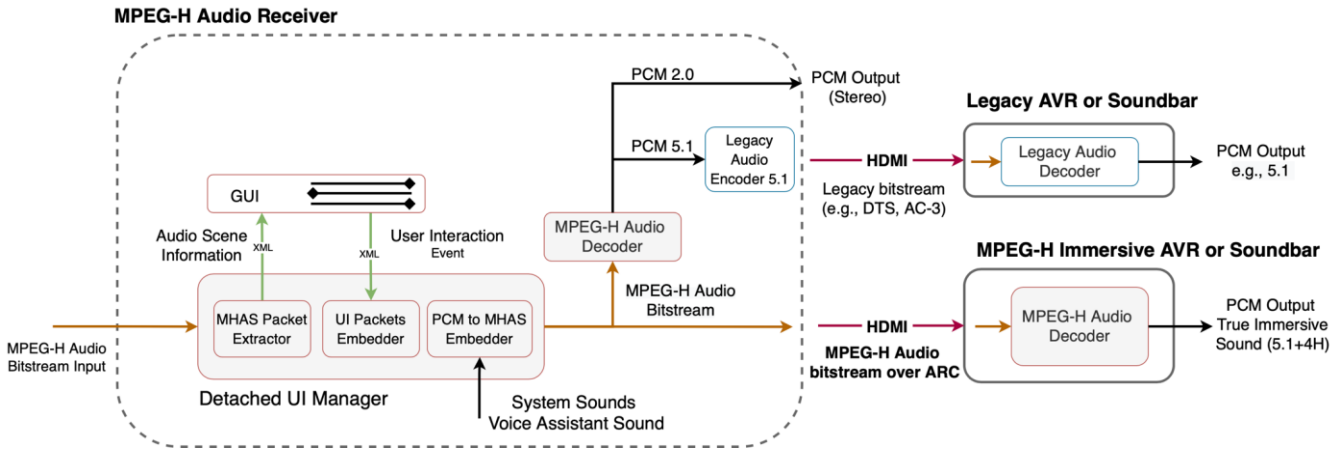


Fig. 11. Top level block diagram of a receiver using MPEG-H Audio

G. *Connectivity to external devices*

The TV 3.0 audio system has to enable the most advanced personalization options and at the same time reproduce the immersive sound, especially when connected to external sound devices. Therefore, the audio component requirements are evaluated by test labs assigned by the SBTVD using various connectivity options. Fig. 10 provides a high-level diagram of a receiver using MPEG-H Audio connected to external playback devices. The signal could be received over RF or IP. Using standardized metadata and interfaces for the systems/application layer, the MPEG-H Audio system enables extremely efficient ways to always provide the best audio experience while offering all its personalization features to the viewers.

Based on the available connections of the receiver (e.g., TV set), different outputs may be used:

- PCM Stereo output,
- HDMI bit stream output for legacy devices, or
- HDMI bit stream passthrough for immersive MPEG-H playback devices (e.g., AVRs and Soundbars).

All connectivity options, requirements and prioritization rules are specified in ABNT NBR 15604:2020, Section C.9 [13].

The MPEG-H Audio system enables a distributed architecture between metadata processing steps and the decoding step. This distributed architecture is enabled by the standardized MPEG-H Audio Stream (MHAS) packetized structure, as specified in ISO/IEC 23008-3, Clause 14 [2]. As seen in Fig. 11 the incoming bit stream is first preprocessed by the so-called "Detached UI Manager" before being sent to the MPEG-H Audio Decoder.

1) *User Interface on systems level*

When a receiver is connected to an external sound system, most of existing audio systems require full decoding, rendering, user interaction and re-encoding to a different audio format to be transmitted over HDMI to the external sound device. This transcoding process is computational complex and introduces additional delay. To avoid such unnecessary transcoding steps, the MPEG-H Audio system is capable to read the metadata required for user interactivity on the MHAS bit stream level without any decoding of the audio data.

As shown in Fig. 11, the Audio Scene Information is extracted from the MPEG-H Audio bit stream at systems level. The "MHAS Packet Extractor" parses the MHAS stream, extracts the MHAS PACTYP_AUDIOSCENEINFO packet and makes it available to the application for usage in a Graphical User Interface (GUI).

In return, the "UI Packets Embedder" accepts the user interactivity information from the application layer. The user interactivity information is encapsulated in the standardized MHAS PACTYP_USERINTERACTION packets, as defined in ISO/IEC 23008-3, Clause 14.4.9 [2]. The user interaction MHAS packets are then inserted "on-the-fly" back into the MHAS packet stream.

If the receiver is using its own loudspeakers or is connected over HDMI to a legacy AVR/Soundbar, the MHAS packet stream is fed into the MPEG-H Audio decoder, which decodes to stereo or 5.1 respectively.

If the receiver is connected over HDMI to an immersive MPEG-H AVR/Soundbar, the MHAS MPEG-H Audio decoder in the receiver is not used and the MHAS stream is provided over HDMI to the external playback device, which will perform the final audio decoding and playout of the audio. All user interactions are embedded into to MHAS stream and will be applied during the decoding in the external sound device.

Using such distributed architecture, the MPEG-H Audio system enables delivery of immersive sound to the end device without any compromise, while at the same time enabling the user interactivity in the receiving device. This unique solution:

- Significantly reduces computational complexity and the audio delay in the receiver,
- Offers the best audio quality for the immersive playback in the external sound system by avoiding any intermediate downmix or rendering in advance of the final playback device,
- Enables the viewer to use a single remote-control for all audio controls and personalization options,
- Provides a consistent UI design independent from the connected external device.

2) *System Sounds and Voice Assistant Sounds*

Providing audible feedback on a user's interaction or system status is important for enhanced accessibility user experience. These system sounds and voice assistant sounds are usually generated by the receiver during playback and guide the user

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

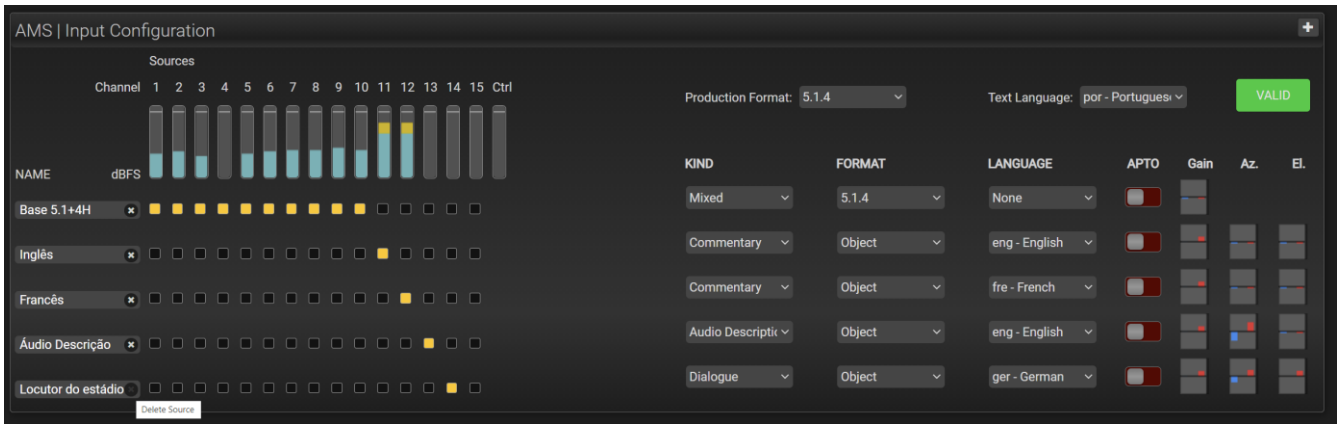


Fig. 12. Linear Acoustic AMS Control Interface – Remove the last audio object example

to navigate through the device options or interactions with the device.

When the receiver is connected to an external sound device these sounds have to be provided together with the main audio to the external device. This is usually achieved by decoding the audio data, mixing the audio data with the system sounds in the receiver and re-encoding for delivery over HDMI to the sound system. As previously described, such process is highly complex, introduces higher audio delay and may lead to compromising the audio quality by intermediate downmix and rendering process.

For ensuring the best possible audio experience, MPEG has specified a mechanism for embedding PCM samples of system sounds or voice assistant audio "on the fly" into an MPEG-H Audio bit stream without decoding the bit stream. The mechanism is based on the capabilities of the MPEG-H Audio system to handle Earcon sounds, the equivalent of visual icons in computer interfaces. Three MHAS packets have been defined for carriage of the PCM data and associated metadata:

- **PACTYP_EARCON** carrying configuration metadata, such as type, id, status (active/inactive), gain, position etc.
- **PACTYP_PCMCONFIG** carrying PCM configuration data such as sampling rate and frame size.
- **PACTYP_PCMDATA** carrying the uncompressed PCM samples.

Using these three standardized MHAS packets, the receiver is capable to embed the system sounds and voice assistant sounds into the received bit stream and deliver the bit stream to an external audio device (AVR, Soundbar) that supports MPEG-H Audio. The external device will decode and render the MPEG-H Audio bit stream and mix in the system sounds and/or voice assistant audio into the rendered audio scene as described in ISO/IEC 23008-3 clause 28.4 [2].

H. Seamless configuration changes in production

The SBTVD TV 3.0 Project was designed to enable the most advanced audio features in existing and future broadcast workflows. This requires seamless playback during changes in production. The broadcasters must be able to change various aspects of the production during live transmission based on their creative intent and offer the best experience to the viewer. Typical changes in a live broadcast will be tested and evaluated during the TV 3.0 Project, including:

- Change of the audio scene (objects, preselections, etc.),
- Enable/disable dialogs and Audio Description in multiple languages,
- Enable/disable interactivity options for one or more preselections,
- Change the interactivity options (min/max gain and position values) for one or more objects, or
- Change the textual labels for one or more objects or preselections.

The MPEG-H metadata used in production, allows alignment to the video frame rate and therefore any change in production (inside an authoring unit or an SDI-level switch) can be seamlessly applied without disturbing the playback on the consumer end devices. Configuration changes during production are translated seamlessly into configuration changes in the bit stream by the encoder and are seamlessly applied in the decoder.

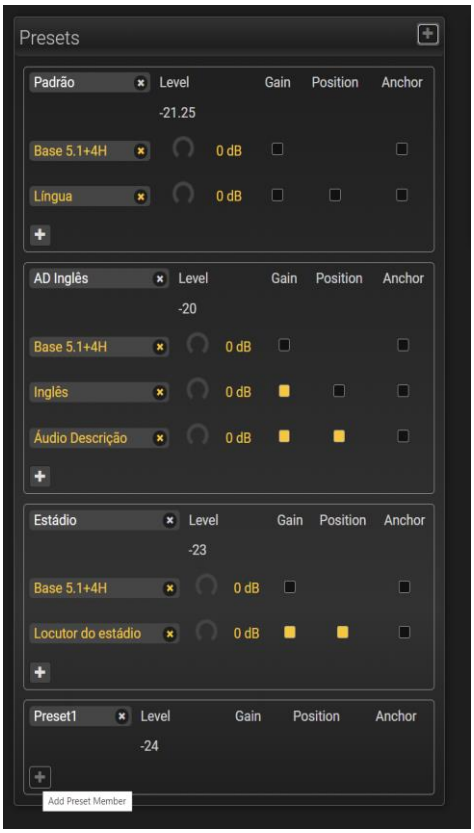


Fig. 13. Linear Acoustic AMS Interface – Add a component to the newly created example.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

Fig. 12 shows an example using the Linear Acoustic AMS authoring unit where the operator is able to remove one audio object in live broadcast. This could be needed in cases where a stadium announcer feed is not available any longer and the broadcaster would prefer to disable the option for the viewers to listen to this audio object. Similarly, the content producer could add an additional audio object, change interactivity options or add a completely new preset as illustrated in Fig. 13. When creating a new preset, the broadcaster can define a new textual label for it, decide which audio components will be part of the preset and what gain and position interactivity options should be allowed.

All these options can be enabled in production using the web control interface of the authoring unit by manually changing each entry or can be configured in advance of the live broadcast and simply uploaded when necessary. Either way, the MPEG-H Audio system will ensure a seamless update of the receiver's user interface.

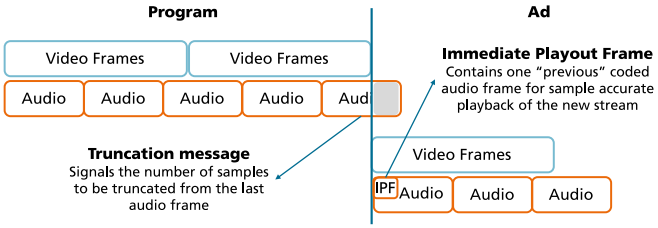


Fig. 14. Sample accurate ad-insertion example.

I. *Seamless ad-insertion and stream splicing*

For legacy systems, various methods have been used for efficiently enabling Ad-insertion. The TV 3.0 Project aims to achieve a seamless behavior during ad-insertion independent of the content used. With Next Generation Audio, it will most likely happen that the live content will use one configuration (e.g., immersive sound with several audio objects and personalization options), while the ad will use a different configuration (e.g., simple stereo without any interactivity options). In such a scenario, four aspects are important for the viewer:

- The transition to and from the ad should be seamless, meaning that no audio dropouts or glitches should occur,
- The ad-insertion can occur at any point in time at a video frame boundary, even in the middle of a coded audio frame,
- The overall perceived loudness level should be preserved before, during and after the ad, and
- The interactivity options displayed to the user should change accordingly, such that they perfectly match the audio content.
- The user shall not be able to change the language or increase the dialog level for example during the ad-break if the ad does not contain such options.
- The user settings should be restored after the ad-break.

For achieving this, the MPEG-H Audio system is using two new concepts: a frame truncation mechanism and an Immediate Playout Frame (IPF) Access Unit (AU).

The MHAS format offers the possibility to transmit truncation information via the AUDIOTRUNCATION packet. The truncation information is used to discard a certain number of audio samples from the beginning or end of a decoded AU. This can be used for alignment of decoded audio data to the video frame boundaries.

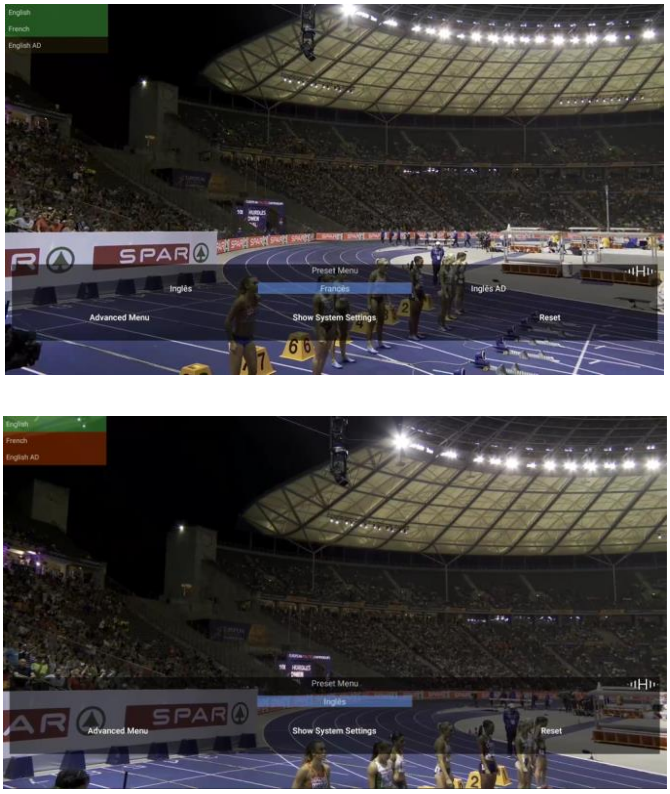


Fig. 15. MPEG-H Receiver Playback of main (broadcast) and two side streams (broadband) [upper picture] and the fallback to broadcast stream when broadband connection is not available [lower picture].

An IPF carries additional information of the "previous" audio frame and thus allows, beyond simple random-access operations, the sample accurate synchronization of audio and video streams. All modern audio codecs require at least two coded AUs for decoding valid audio samples and therefore it is essential to have access to the "previous frame" when changing to a different stream. Defining IPFs at the beginning of the ad or at the beginning of the stream after the ad allows for glitch-free and on-the-fly reconfiguration of the stream.

Fig. 14 illustrates such example, where the last frame of the stream (before the ad-insertion or configuration change) is truncated, and the new stream (after the configuration change) starts with an IPF. This way, the configuration change will be seamless and aligned to the video splicing point in a sample-accurate fashion.

J. *Hybrid delivery and extensibility*

With a forward-looking thinking, the SBTVD Forum has decided from the beginning to select an IP-based transport layer which will allow a smooth interaction between the broadcast and broadband paths. Such approach allows the delivery of a standard program over the broadcast available to all viewers and additional features can be enabled via broadband (e.g., premium commentators, different languages etc.). The TV 3.0 Project requires, for its audio system, such advanced capabilities where the additional features are presented to the user only when available over the broadband connection and fallback to the default settings from the broadcast stream in cases where the broadband connection is suddenly not available. Additionally, all user interactions should not disrupt the audio playback even when the audio components are delivered over the two different transmission paths.

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

Fig. 15 illustrates an example of the MPEG-H Audio receiver with hybrid reception using the test content for the TV 3.0 evaluation. In this example, the main stream contains a complete audio scene, meaning that it provides the complete experience (e.g., 5.1+4H immersive sound with English dialog) even without any broadband connection. The additional streams delivered via broadband will contain enhancements of the audio scene (e.g., the French dialog in the second stream and the Audio Description in the third stream).

The upper picture in Fig. 15 shows the available MPEG-H user interaction options when both connections are available, while the lower picture shows the fallback to the options available in the broadcast path only (the side streams are both marked in "red" and not available).

The MPEG-H multi-stream features enable the receiver to easily synchronize the streams and seamlessly switch between the streams. The playback always starts with the main stream (broadcast) but the MPEG-H Audio metadata is used to enable the receiver to display the available streams via broadband without disturbing the main broadcast feed. The user is able to switch between the available interactivity options in a seamless way. Based on the active preset the receiver is requesting the additional streams.

Besides the hybrid delivery which enables scalability of the broadcast system, the TV 3.0 Project requires easy extensibility of the audio system for enabling new applications in the future. The MPEG-H Audio system enables extensibility using well established mechanisms inherited from previous MPEG audio standards as well as through its state-of-the-art packetized bit stream structure: MHAS. Additional MHAS packets may be defined in a backwards compatible way, meaning that existing decoders will simply ignore the unknown newly defined MHAS packets while new decoders will be capable to read and process these newly defined MHAS packets for offering enhanced experiences.

Moreover, MPEG-H Audio was already selected by MPEG as the only audio codec for the future MPEG-I Audio system which will provide support 6 Degrees of Freedom (6DoF) audio playback. The MPEG-I Audio work item, currently under standardization, will define additional metadata which will be embedded in new MHAS packets.

MPEG-H features extension mechanisms on different layers that allow future extensions in well-proven, well-defined, clean, efficient, and backwards-compatible ways.

V. Conclusion

Fraunhofer IIS, ATEME, DiBEG, and ATSC have proposed the MPEG-H Audio system in response to the SBTVD TV 3.0 Call for Proposals. The TV 3.0 Project has been designed to offer the most advanced audio options to the viewers and requires advanced solutions for metadata and audio handling across the entire production and broadcast chain. With requirements for immersive sound, advanced interactivity and accessibility options, hybrid delivery, consistent loudness after user interaction, connectivity options for external sound devices and seamless configuration changes, to be evaluated in a real-time broadcast environment, the TV 3.0 Project has set the ground for the most detailed evaluation of the proposed audio systems.

The MPEG-H Audio system is the only fully standardized audio system fulfilling all TV 3.0 requirements listed in the Call for Proposals and provides the most advanced feature set and use cases as detailed in this document.

With hardware implementations already available and used in 24/7 broadcast, the MPEG-H Audio system can ensure an easy transition from the existing ISDB-Tb broadcast system to the future based TV 3.0 system.

References

[1] Brazilian Digital Terrestrial Television System Forum (SBTVD) TV 3.0 Call for Proposals, Available: https://forumsbtvd.org.br/tv3_0/

[2] ISO/IEC 23008-3:2019: "Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D audio", including ISO/IEC 23008-3:2019/AMD 1:2019 "Audio metadata enhancements" and ISO/IEC 23008-3:2019/AMD 2:2020, "3D Audio baseline profile, corrections and improvements".

[3] MPEG-I Immersive Audio Call for Proposals. Available: https://www.mpegstandards.org/wp-content/uploads/mpeg_meetings/134_OnLine/w20449.zip

[4] N19407, MPEG-H 3D Audio Baseline Profile Verification Test Report. Available: https://www.mpegstandards.org/wp-content/uploads/2020/07/w19407.zip

[5] N16584, MPEG-H 3D Audio Verification Test Report. Available: http://mpeg.chiariglione.org/standards/mpeg-h/3d-audio/mpeg-h-3d-audio-verification-test-report

[6] R. Bleidt et al. "Development of the MPEG-H TV Audio System for ATSC 3.0," in IEEE Transactions on Broadcasting, vol. 63, no. 1, March 2017. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7874294

[7] ATSC A/342-3:2021 "ATSC Standard, A/342 Part 3: MPEG-H System," Advanced Television Systems Committee, Washington, DC, 11 March 2021. Available: https://www.atsc.org/wp-content/uploads/2021/03/A342-2021-Part-3-MPEG-H.pdf

[8] ATSC A/331:2019, "ATSC Standard: Signaling, Delivery, Synchronization, and Error Protection," Advanced Television Systems Committee, Washington, DC, 20 June 2019, https://www.atsc.org/wp-content/uploads/2017/12/A331-2017-Signaling-Deivery-Sync-FEC-1.pdf

[9] TTAK-KO-07.0127R1: TTA - Transmission and Reception for Terrestrial UHDTV Broadcasting Service, Revision 1, December 2016.

[10] MPEG-H Audio selected to enhance Brazilian digital television with immersive and personalized sound, https://www.audioblog.iis.fraunhofer.com/mpegh-brazil-isdbtb

[11] ABNT NBR 15602-2:2020, Televisão digital terrestre - Codificação de vídeo, áudio e multiplexação - Parte 2: Codificação de áudio.

[12] ABNT NBR 15603:2020, Televisão digital terrestre - Multiplexação e serviços de informação (SI), https://forumsbtvd.org.br/legislacao-e-normas-tecnicas/normas-tecnicas-da-tv-digital/english/

[13] ABNT NBR 15604:2020, Televisão digital terrestre – Receptores, https://forumsbtvd.org.br/legislacao-e-normas-tecnicas/normas-tecnicas-da-tv-digital/english/

[14] ETSI TS 126 118 V15.0.0 (2018-10), 5G; 3GPP Virtual reality profiles for streaming applications (3GPP TS 26.118 version 15.0.0 Release 15).Available:https://www.etsi.org/deliver/etsi_TS/126100_126199/126118/15.00.00_60/ts_126118v150000p.pdf

[15] TS 101 154 v2.3.1: Digital Video Broadcasting (DVB) – Specification for the use of Video and Audio Coding in Broadcasting Applications based on the MPEG-2 Transport Stream.

[16] ETSI EN 300 468 V1.16.1 (2019-08), Digital Video Broadcasting (DVB); Specification for Service Information (SI) in DVB systems, https://www.etsi.org/deliver/etsi_en/300400_300499/300468/01.16.01_60/en_300468v011601p.pdf

[17] International Telecommunications Union (ITU) Recommendation ITU-R BS.1196-7 (01/2019), Audio coding for digital broadcasting https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1196-7-201901-S!!PDF-E.pdf

[18] South Korea launches UHD TV with MPEG-H Audio, https://www.audioblog.iis.fraunhofer.com/south-korea-uhd-tv-mpeg-h

[19] A. Murtaza and S. Meltzer, "First Experiences with the MPEG-H TV Audio System in Broadcast," SET INTERNATIONAL JOURNAL OF BROADCAST, 2018 Available: https://www.set.org.br/ijbe/ed4/Artigo%206.pdf

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

[20] Shows do Rock in Rio transmitidos com tecnologia desenvolvida pelo Fraunhofer IIS, https://panoramaaudiovisual.com.br/shows-do-rock-in-rio-transmitidos-com-tecnologia-desenvolvida-pelo-fraunhofer-iis/
[21] Centro de treinamento de áudio MPEG-H no Brasil, https://www.brazil.fraunhofer.com/pt/news/noticias-mais-recentes/novo-centro-de-treinamento-de-audio-mpeg-h-abre-suas-portas-em-s.html
[22] Studio Recommendations for 3D Audio productions with MPEG-H Audio, https://www.iis.fraunhofer.de/content/dam/iis/de/doc/ame/wp/FraunhoferIIS_TechnicalPaper_Studio_Recommendations_3DAudio-MPEG-H.pdf
[23] ITU-R BS.2076-2, Recommendation ITU-R BS.2076-2, Audio definition model, Geneva 10/2019
[24] ITU-R BS.2125, Recommendation ITU-R BS.2125, A serial representation of the Audio Definition Model, Geneva 01/2019.
[25] The MPEG-H Authoring Suite, https://www.iis.fraunhofer.de/en/ff/amm/dl/software/mas.html
[26] The Spatial Audio Designer (SAD) Plugin, https://newaudiotechnology.com/products/spatial-audio-designer/
[27] The MPEG-H ADM Profile. Available: https://www.iis.fraunhofer.de/en/ff/amm/dl/whitepapers/adm-profile.html

Adrian Murtaza received his M.Sc. degree in Communication Systems from the École Polytechnique Fédérale de Lausanne, Switzerland in 2012 with a thesis on "Backward Compatible Smart and Interactive Audio Transmission". Upon graduation he joined Fraunhofer IIS, where he works as a Senior Manager, Technology and Standards.

Adrian joined MPEG in 2013 and since then contributed to the development of various audio technical standards in MPEG-D and MPEG-H. He serves as Fraunhofer's Standards Manager in a number of industry standards bodies, including SBTVD, ATSC, CTA, DVB, HbbTV and SCTE, and is the co-author of multiple specifications in those groups.

More recently he focused on specification of Next Generation Audio delivery and transport in ATSC 3.0 systems and MPEG-2 Transport Stream based systems, as well as on enabling of MPEG-H Audio services in different broadcast and streaming ecosystems. With a strong interest in VR/AR media solutions he is actively involved in MPEG-I efforts targeting future immersive applications.

Stefan Meltzer studied electrical engineering at the Friedrich-Alexander University in Erlangen, Germany. In 1990 he joined the Fraunhofer Institute for Integrated Circuits (IIS) in Erlangen, Germany. After working in the field of IC design for several years, Stefan became the project leader for the development of the WorldSpace Satellite Broadcasting system in 1995 and in 1998 of the XM Satellite Radio broadcasting system. His team was responsible for the system design, chip set design, field trials and development of a reference signal generator.

In 2000 he joined Coding Technologies in Nuremberg as Vice President for business development, Germany. His responsibilities included broadcasting and consumer electronics. During his time at Coding Technologies, HE-AAC was accepted in numerous broadcasting standards and applications. After Coding Technologies was acquired by Dolby Labs, Stefan joined Iosono as CTO in April 2008.

From January 2010, Stefan worked as independent technology consultant with the focus on audio and multimedia. In this role he supported Fraunhofer IIS in the business development and marketing activities within the TV broadcast market. In April 2018, Stefan joined Fraunhofer IIS again and is now in charge of business development for TV broadcast applications.

Yannik Grewe received his M.Eng. degree in 'audiovisual media – sound' with a thesis on 'Perception and reproduction of floor level sound in consumer audio playback'. Yannik joined Fraunhofer IIS in 2013 and serves today as senior engineer for audio production technologies, focusing MPEG-H 3D Audio. He is extensively involved as a sound engineer in producing immersive music applications and MPEG-H immersive and interactive audio. His current role includes a close relation to major broadcasters and streaming service providers in Asia, Europe, and South America to enable MPEG-H Audio in their ecosystems.

Nicolas Faecks received a B.Sc. degree in 'media technologies and a M.A. degree in 'time-based media – Sound – Vision'. Before joining the Fraunhofer Institute for Integrated Circuits IIS in 2014, he was a researcher on lighting technologies at Airbus. At Fraunhofer, he focused on All-IP-Workflows and MPEG-H 3D Audio Systems. As a System Engineer, he was responsible for the MPEG-H 3D Audio broadcast and streaming systems during major events, such as Roland Garros, the European Athletics Championships, the Eurovision Song Contest or the Youth Olympic Games.

Lucas Gregory graduated from Université Polytechnique des Hauts de France (INSA HDF) in 2017 with a Master Degree in Audio and Video System Engineering. He then joined the ATEME Research and Innovation team, where he helped in several French collaborative projects to study and promote innovative technologies such as 360° video, Next Generation Audio codecs and Next-Gen TV (ATSC 3.0). Today, he works on immersive audio technologies as well as the trans-packaging of low-latency content.

Dr. Mickaël RAULET received his Ph.D. degree in 2006. He joined ATEME in 2015, where he is now CTO. He is leading the standardization effort at ATEME and is following the different activities in ATSC, DVB, 3GPP, ISO/IEC, ITU, MPEG, DASH-IF and UHD Forum. He is managing several collaborative R&D projects for ATEME.

Received in 2021-08-25 | Approved in 2021-12-07

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.3. Web Link: http://dx.doi.org/10.18580/setijbe.2021.3

Immersive Audio Application Coding Proposal to the SBTVD TV 3.0 Call for Proposals

Oliver Major
Ziad Shaban
Bernd Czelhan
Adrian Murtaza

CITE THIS ARTICLE
Major, Oliver; Shaban, Ziad; Czelhan, Bernd and Murtaza, Adrian; 2021. Immersive Audio Application Coding Proposal to the SBTVD TV 3.0 Call for Proposals. SET INTERNATIONAL JOURNAL OF BROADCAST ENGINEERING. ISSN Print: 2446-9246 ISSN Online: 2446-9432. doi: 10.18580/setijbe.2021.4. Web Link: http://dx.doi.org/10.18580/setijbe.2021.4

SET INTERNATIONAL JOURNAL OF BROADCAST ENGINNERING - SET IJBE V.7, 2021, Article 4, 9p.
 / ISSN Print: 2446-9246 | ISSN Online: 2446-9432

Immersive Audio Application Coding Proposal to the SBTVD TV 3.0 Call for Proposals

Oliver Major, Ziad Shaban, Bernd Czelhan, and Adrian Murtaza,
Fraunhofer Institute for Integrated Circuits (IIS)

***Abstract*— In July 2020 the Brazilian Terrestrial Television System Forum (SBTVD) has issued a Call for Proposals (CfP) for their next-generation digital TV system called TV 3.0. Fraunhofer IIS and ATEME have proposed the MPEG-H Audio system, based on the open international standard ISO/IEC 23008-3, MPEG-H 3D Audio, as a candidate technology for the Application Coding component of SBTVD TV 3.0. The submitted proposal specifies a new Application Programming Interface (API) enabling applications to make use of the next-generation interactivity features of the MPEG-H Audio system.**

This paper provides a detailed description of the proposed API, as well as the submitted JavaScript implementation, and the architecture of the prototype system. Additionally, the paper outlines the proposed evaluation process demonstrating how the MPEG-H Audio system fulfills the TV 3.0 Application Coding requirements for 3D object-based immersive audio interaction and emergency warning information delivery.

***Index Terms*— 3D and Immersive Audio, Accessibility, Adaptation and customization of content, ATSC 3.0, Application Coding, API, Broadcast, Broadband, Emergency warning system, HTML 5, Hybrid Delivery, Immersive Sound, MPEG-H Audio, Next Generation Audio, Object-based broadcasting, Personalized Sound, SBTVD TV 3.0, Streaming**

I. Introduction

THE Brazilian Digital Terrestrial Television System Forum (SBTVD) has issued in July 2020 a Call for Proposals (CfP) seeking input for Brazil's next-generation Digital TV system under the name “TV 3.0 Project” [1]. The SBTVD Forum has established a detailed set of TV 3.0 requirements and use cases covering six system components (Over-the-air Physical Layer, Transport Layer, Video Coding, Audio Coding, Captions, and Application Coding). The CfP was divided into two phases: Phase 1 required an initial submission from proponents identifying the candidate technology and providing basic information, while in Phase 2 the proponents were expected to submit a full specification of the candidate technology as well as hardware and software solutions for the feature evaluation.

In response to the SBTVD TV 3.0 Call for Proposals, Fraunhofer IIS, ATEME, the Digital Broadcasting Experts Group (DiBEG) and the Advanced Television Systems Committee (ATSC) have jointly proposed the MPEG-H Audio system, based on the open international standard ISO/IEC 23008-3, MPEG-H 3D Audio [2], as the audio component. Additionally, Fraunhofer IIS and ATEME have submitted a proposal for the Application Coding component, specifying a new Application Programming Interface (API) fulfilling the requirements for 3D object-based immersive audio interaction and emergency warning information delivery using an interactive application.

MPEG-H Audio was already adopted in Brazil as part of the TV 2.5 Project to enhance the audio experience over ISDB-Tb with immersive and personalized sound and it is currently fully specified in the ABNT standards [3][4][5][6]. This has enabled broadcasters and content creators in Brazil to gain experience in advanced audio productions with MPEG-H. Having professional broadcast equipment from several major providers available has been essential for using the system in the existing ISDB-T broadcast infrastructure and consequently the MPEG-H Audio system can ensure a smooth transition from TV 2.5 to TV 3.0.

The MPEG-H Audio system was developed to allow highly efficient immersive audio transmission and new capabilities such as advanced accessibility, interaction, personalization and adaptation of audio to different usage scenarios, delivering the best possible experience and taking audio to the next level. A detailed technical description of the MPEG-H Audio system is provided in [7] and lessons learned during live broadcast of major events using MPEG-H Audio are described in [8].

The use of audio objects, usually in combination with channel-based audio, enables the viewers to interact with the content in new ways and create a personalized listening experience. The MPEG-H Audio metadata carries all the information needed to allow viewers to change the properties of audio objects by attenuating or increasing their level, disabling them, or changing their position in three-dimensional space. Additionally, the MPEG-H Audio metadata structures empower broadcasters to enable or disable interactivity options and to strictly set the limits to which extent a user can interact with the content.

One of the most important use cases for personalization is dialog enhancement. With MPEG-H Audio the dialog or commentary for a program is sent as an audio object and associated metadata. This allows the viewer at home to adjust the relative volume or “presence” of the dialog relative to the rest of the audio elements in the program. This simple case can be extended to offer two or more dialog objects with different languages or commentaries (e.g., biased

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.4. Web Link: http://dx.doi.org/10.18580/setijbe.2021.4

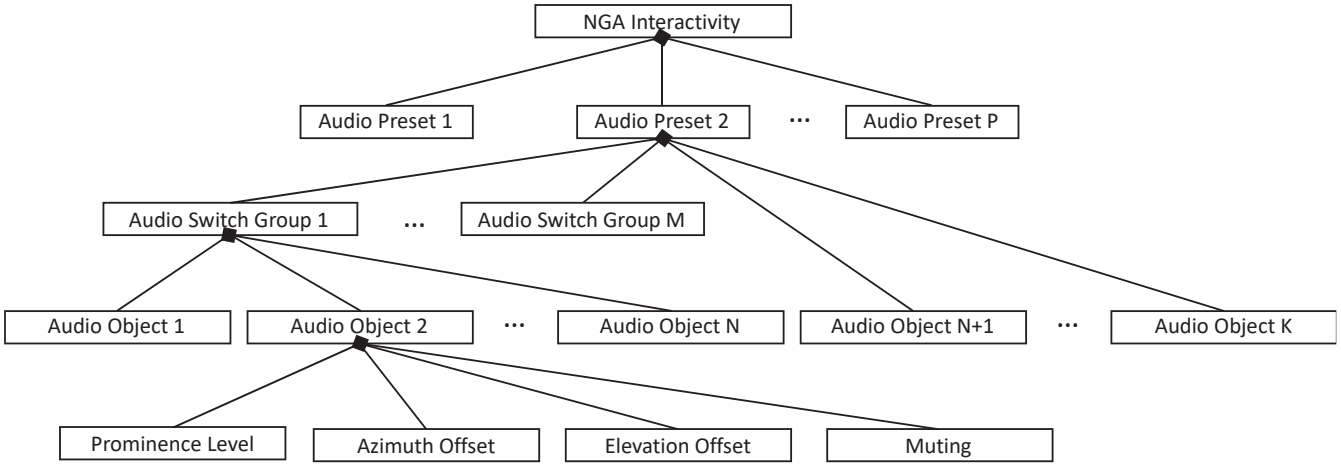


Fig. 1. NGAInteractivity overview.

commentators for each of the teams during a football game).

Moreover, the MPEG-H Audio metadata enables broadcasters to provide several versions of the content, as so-called "presets", which describe how all channels and object signals are mixed together and presented to the viewer. Choosing between different presets is the simplest way to interact with the content. Additionally, advanced interactivity settings can be offered to more experienced users for manipulating objects individually.

The MPEG-H Audio system standardizes a rich metadata set that enables the most advanced and flexible end-user interactivity and personalization experience, while still offering full control over these features to the broadcasters. The presence of metadata in the audio bitstream is extremely important for enabling the Application Coding layer to offer all these options to the user in a controlled and well-defined way.

In order to allow application developers to make use of these advanced audio features in broadcast applications, new APIs for interacting and controlling the rendering of the current audio scene are necessary. Therefore, new NCL (audio) properties have been proposed in [10]. The current document continues the work on new APIs for controlling advanced audio features and will describe the MPEG-H Audio System proposal for the Application Coding part of the SBTVD TV 3.0 Call for Proposal [1].

As part of this submission, a new JavaScript [14] API for controlling advanced audio features was defined. The submission also contains a prototype receiver device and prototype application, which practically show that the proposal fulfills the Application Coding requirements of the SBTVD TV 3.0 Call for Proposal [1].

In this paper, we present the API specification and explain the design principles which have been used. Additionally, the prototype implementation and the test results submitted for evaluation in the TV 3.0 Project will be described. The SBTVD TV 3.0 evaluation process of the proposal is supported by the proposed API and prototype. Lastly, we conclude the paper with a summary of the discussed topics and an outlook to further work.

II. API Description

This section describes the proposed API for the Application Coding part of the SBTVD TV 3.0 Call for Proposal [1]. The proposal extends the Ginga-HTML5 [9] environment by a JavaScript API that allows interaction with the built-in MPEG-H Audio decoder. This allows application developers programmatic control over the MPEG-H user interactivity features.

The basis for the proposed API is an extension of the *HTMLMediaElement* [21] with an *ngaInteractivity* attribute of type *NGAInteractivity*. This serves as an entry point for programmers to take advantage of next-generation audio features on an *HTMLMediaElement* that is playing back an MPEG-H audio stream. The structure of the *NGAInteractivity* interface and its sub-interfaces is described in the following overview:

- The *NGAInteractivity* interface contains a list of audio presets, with exactly one preset active at a time.
- Each *AudioPreset* object is representing an audio preset and contains a list of audio objects and a list of audio switch groups. This implies that different audio presets can have different audio objects and audio switch groups associated with them. It also allows each preset to maintain the state of its audio objects and audio switch groups.
- The state of each *AudioObject* instance consists of its prominence level in dB, its azimuth offset in degrees, its elevation offset in degrees, and whether or not it is muted.
- Each *AudioSwitchGroup* instance contains a set of audio objects with the same parameters as described above. Additionally, an audio switch group contains exactly one active audio object at a time.

The base interfaces used in the API are *AudioElement* and *AudioProperty*. Both of these are observable by extending the *EventTarget* API [19]. This means that the user can add and remove event listeners. The system will dispatch an event of type "change" when any of the associated parameters have changed. In addition, the *onchange* property on these interfaces allows the user to set an event handler for change events.

Fig. 1 summarizes the fundamental structure mentioned above. In the rest of this section, we describe the proposed interfaces in detail and explain how they enable users to use interactivity features.

A. NGAInteractivity API

The NGAInteractivity API, shown in Table I, allows the user to list all available AudioPresets, get and set the active

AudioPreset and get the default AudioPreset. Obtaining the default AudioPreset can be useful in case no active AudioPreset has been set by the user. The interface also enables the user to reset the audio scene to its default state, and set the language used for displaying the User Interface (UI) text and labels.

NGAInteractivity extends the EventTarget API [19]. A "change" event will be dispatched when the AudioPresetList or any of the parameters therein changed.

TABLE I
NGAINTERACTIVITY

Property/Method	Return type	Type
audioPresets	AudioPresetList	Read-only property
defaultPreset	AudioPreset	Read-only property
activePreset	AudioPreset	Read-only property
onchange	EventHandler	Property
setActivePresetById(int presetId)	void	Method
resetToDefault()	void	Method
setDisplayLanguage(string language)	void	Method

B. AudioPresetList API

The AudioPresetList API, shown in Table II, represents an iterable dynamic list of AudioPresets. Exactly one AudioPreset within a list of AudioPresets is enabled at a time.

AudioPresets within the list can be accessed either by iterating over the list indices, or by providing a *presetId* to the *getAudioPresetById* method. If the specified *presetId* matches the *id* of an AudioPreset within the list, that AudioPreset is returned, otherwise *getAudioPresetById* returns null.

TABLE II
AUDIOPRESETLIST

Property/Method	Return type	Type
length	int	Read-only property
getActivePresetById()	AudioPreset or null	Method

C. AudioPreset API

The AudioPreset API, shown in Table III, represents a singular AudioPreset. The Audio Preset encapsulates a list of AudioObjects and a list of AudioSwitchGroups, accessible to the user via the *audioObjects* and *audioSwitchGroups* properties respectively.

AudioPreset extends the AudioElement API and inherits its *id* and *label* properties. It also transitively inherits the *EventTarget* API, by which it will receive a "change" event when any property within an associated AudioObject or AudioSwitchGroup.

TABLE III
AUDIOPRESET

Property/Method	Return type	Type
id	int	Read-only property
label	string	Read-only property
audioObjects	AudioObjectList	Read-only property
audioSwitchGroups	AudioSwitchGroupList	Read-only property
onchange	EventHandler	Property

D. AudioElement API

The AudioElement API, shown in Table IV, serves as a base class for named observable UI elements. Each AudioElement is identified by a unique *id* and contains a *label* describing the AudioElement in the display language chosen by the user via the *setDisplayLanguage* method in the NGAInteractivity interface.

AudioElement extends the EventTarget API [19]. The system will dispatch a "change" event when any property associated with that AudioElement changed.

TABLE IV
AUDIOELEMENT

Property/Method	Return type	Type
id	int	Read-only property
label	string	Read-only property
onchange	EventHandler	Property

E. AudioSwitchGroupList API

The AudioSwitchGroupList API, shown in Table V, represents an iterable dynamic list of AudioSwitchGroups.

AudioSwitchGroups within the list are accessible either by iterating over the list indices, or via the *getAudioSwitchGroupById* method. If the specified *switchGroupId* matches the *id* of an AudioSwitchGroup within the list, that AudioSwitchGroup is returned, otherwise null is returned.

TABLE V
AUDIOSWITCHGROUPLIST

Property/Method	Return type	Type
length	int	Read-only property
getAudioSwitchGroupById(int switchGroupId)	AudioSwitchGroup or null	Method

F. AudioSwitchGroup API

The AudioSwitchGroup API, shown in Table VI, represents a switch group of AudioObjects. AudioSwitchGroup allows the user to list all associated AudioObjects through the *audioObjects* property. Within an AudioSwitchGroup, exactly one AudioObject is active at a time. The AudioSwitchGroup allows the user to get and set the active AudioObject and get the default AudioObject.

Moreover, AudioSwitchGroup objects can have a *mutingProperty*, which is an optional property of type AudioBooleanProperty. It allows the AudioSwitchGroup to be muted, regardless of which AudioObject is active and which is not. If muting for the AudioSwitchGroup is disallowed, the property will be null.

AudioSwitchGroup extends the AudioElement API and inherits its *id* and *label* properties. It also transitively inherits the EventTarget API, by which it will receive a "change"

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.4. Web Link: http://dx.doi.org/10.18580/setijbe.2021.4

event when the active AudioObject or any property within an associated AudioObject changed.

TABLE VI
AUDIOSWITCHGROUP

Property/Method	Return type	Type
id	int	Read-only property
label	string	Read-only property
mutingProperty	AudioBooleanProperty or null	Read-only property
audioObjects	AudioObjectList	Read-only property
activeAudioObject	AudioObject	Read-only property
defaultAudioObject	AudioObject	Read-only property
onchange	EventHandler	Property
setActiveAudioObjectById(int objectId)	void	Method

G. AudioObjectList API

The AudioObjectList API, shown in Table VII, represents an iterable dynamic list of AudioObjects.

AudioObjects within the list are accessible either by iterating over the list indices, or via the *getAudioObjectById* method. If the specified *objectId* matches the *id* of an AudioObject within the list, that AudioObject is returned, otherwise *getAudioObjectById* returns null.

TABLE VII
AUDIOOBJECTLIST

Property/Method	Return type	Type
length	int	Read-only property
getAudioObjectById(int objectId)	AudioObject or null	Method

H. AudioObject API

The AudioObject API, shown in Table VIII, represents a singular AudioObject. The state of the AudioObject is defined by its prominence level value in dB, its azimuth offset value in degrees, its elevation offset value in degrees, and whether or not it is muted.

Each of these features is accessible through a property on the AudioObject interface. By changing the values of these properties via the AudioNumericProperty and AudioBooleanProperty interfaces, the user is able to control the state of the AudioObject.

All of these properties are optional. If user interaction with one of these properties is disallowed, it will return a null value. In case all properties are null, *isActionAllowed* returns false, indicating that no user interaction is possible on the AudioObject.

The *contentKind* property is a number representing the "mae_contentKind" as defined in [2].

AudioObject extends the AudioElement API and inherits its *id* and *label* properties. It also transitively inherits the EventTarget API, by which it will receive a "change" event when a property of the AudioObject was changed.

TABLE VIII
AUDIOOBJECT

Property/Method	Return type	Type
id	int	Read-only property
label	string	Read-only property
isActionAllowed	boolean	Read-only property
contentKind	int	Read-only property
mutingProperty	AudioBooleanProperty or null	Read-only property
prominenceLevelProperty	AudioNumericProperty or null	Read-only property
azimuthOffsetProperty	AudioNumericProperty or null	Read-only property
elevationOffsetProperty	AudioNumericProperty or null	Read-only property
onchange	EventHandler	Property

I. AudioProperty API

The AudioProperty API, shown in Table IX, serves as a base class for unnamed observable UI elements. Each AudioProperty has interfaces for getting and setting its value depending on its concrete subtype.

AudioElement extends the EventTarget API [19]. The system will dispatch a "change" event when the property's value was changed.

TABLE IX
AUDIOPROPERTY

Property/Method	Return type	Type
onchange	EventHandler	Property

J. AudioBooleanProperty API

The AudioBooleanProperty API, shown in Table X, represents a singular Boolean value. It allows the user to get and set the value through its methods. It also allows the user to get the default value of the property.

AudioBooleanProperty extends the AudioProperty API and transitively the EventTarget API, by which it will receive a "change" event when the property's Boolean value was changed.

TABLE X
AUDIOBOOLEANPROPERTY

Property/Method	Return type	Type
defaultValue	boolean	Read-only property
onchange	EventHandler	Property
getValue()	boolean	Method
setValue(boolean value)	void	Method

K. AudioNumericProperty API

The AudioNumericProperty API, shown in Table XI, represents a singular numeric value. It allows the user to get and set the value through its methods. It also allows the user to get the default value of the property.

The value of the property can be restricted to a specific range of values that the user can access through its *minValue*

and *maxValue* properties. It is worth mentioning that the unit of the value depends on the property it represents. For instance, an AudioNumericProperty representing azimuth offset will have numeric values in degrees, while an AudioNumericProperty representing prominence gain will have numeric values in dB.

AudioBooleanProperty extends the AudioProperty API and transitively the EventTarget API, by which it will receive a "change" event when the property's numeric value was changed.

TABLE XI
AUDIONUMERICPROPERTY

Property/Method	Return type	Type
minValue	int	Read-only property
maxValue	int	Read-only property
defaultValue	int	Read-only property
onchange	EventHandler	Property
getValue()	int	Method
setValue(int value)	void	Method

III. PROTOTYPE IMPLEMENTATION

To prove the feasibility of the proposed API, we have implemented a prototype system that a) contains JavaScript bindings of the API itself as long as a native version in the browser is not available, b) shows the usage of the API to render functional user interfaces for audio interactivity in two use-case applications, and c) sets up a TV environment and a broadcaster environment to embed the use-cases in a context similar to production systems. We explain these three parts of the prototype system in this section.

A. JavaScript Implementation

The API proposal was written with a browser implementation and the possibility of a later standardization as a web standard in mind. As such, the specification resembles the standard documents published by the W3C in its form and content. In many cases, we resolved discussions in the development of the API in favor of consistency with other web standards, in order to give web developers, who are familiar with other browser APIs, the best possible development experience. This means that the translation of the specification into a JavaScript [14] interface follows the strict rules of the WebIDL [13] language used for the proposal. There is little room for interpretation regarding the interfaces.

To show the adequacy of the APIs for the purposes of enabling user interfaces capable of triggering audio interactivity in an MPEG-H Audio [2] playback, we implemented them in the *MpeghUiLib* JavaScript library. For the implementation, the TypeScript language[1] was chosen for three reasons. First, it is strictly typed, allowing us to declare the interface types and having them checked automatically. Through this, we enforced that our implementation's interfaces match the API proposal with the help of static checking tools. Secondly, TypeScript generates pure JavaScript code which can run on any modern and unmodified browser. Lastly, we also provide type declarations with the library to aid the user in developing applications for the proposed API. TypeScript's wide adoption in the web developer community lowers the barrier of entry to inspect and modify the prototype's code.

The interfaces specified in the proposal are fully implemented in the MpeghUiLib. The structure of NGAInteractivity objects and its sub-objects is fully available as described in the specification. That should allow carefully developed applications dependent on the MpeghUiLib to work with the browser implementation once it is available.

Contrary to the interfaces, the full implementation of the specified behavior is in a prototype state. The functionality that is needed for the use-case apps to work correctly is fully implemented as specified and was the main focus of the submission. In its current state, all parsing of MPEG-H audio scene data and the full translation to JavaScript objects is done correctly, as well as the dispatching of decoder events in the case of user interaction.

The main difference between the specification and the implementation in the MpeghUiLib is the entry point, which is supposed to be an extension of the HTMLMediaElement as defined in [21]. In the current phase of the proposal, we connect to an external decoder and cannot take advantage of HTMLMediaElements natively decoding MPEG-H in the browser. Hence, it makes no sense to connect the NGAInteractivity object to an HTMLMediaElement directly. Instead, our implementation offers the MpeghUiLib constructor to the user, so they can explicitly instantiate an MPEG-H user interface for their decoder. The additional *parseAudioScene* interface function and the user-defined *onUiAction* callback are available to feed audio scenes from the decoder to the library and user interactions from the library to the decoder respectively. These two additional functions will not be available in the native implementation of the API.

Furthermore, since multi-client interactivity was not needed for the use-case applications in the submission, the EventTarget dispatching is not fully implemented as specified: in the current state, only the respective element that a change event was dispatched on will trigger its *onchange* callbacks. According to the proposed API, also the parent elements of changed properties should be notified about a change, similar to bubbling events in the DOM specification [19]. This offers users to implement more granular handling of changes in their applications and is mostly useful if we assume that property changes can also be initiated in places other than the implemented user interface, which is not applicable to the provided use-case apps. The only case that needs to be considered in the current state is when the decoder itself triggers a scene change. This proves not to be a problem though, because the user can redraw the whole user interface in that case, which does not require bubbling events.

B. Use-case Applications

To show the usage of the API from an application developer's point of view, we have included two use-case applications in our submission. The difference between the two applications is primarily the content being played back to showcase the different application scenarios, specifically audio interactivity and the emergency warning system. Beyond that, there is no difference between the applications.

[1] https://www.typescriptlang.org/

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.4. Web Link: http://dx.doi.org/10.18580/setijbe.2021.4

© 2021 SET / ISSN Print: 2446-9246 | ISSN Online: 2446-9432

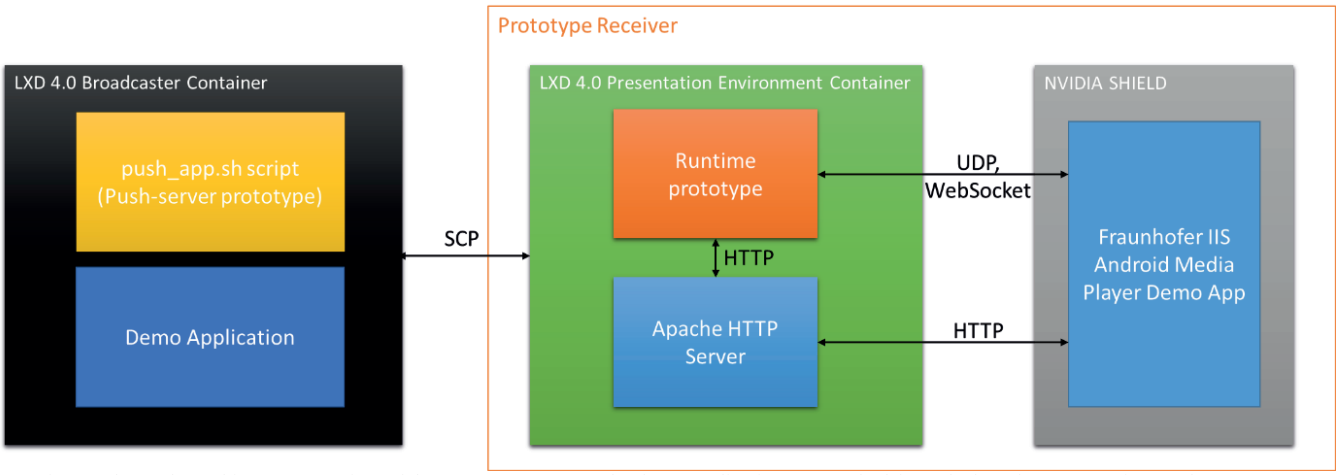


Fig. 2. Shows the architecture overview of the prototype system. The demo applications are pushed from the broadcaster container to the prototype receiver, where the runtime prototype containing the use-case apps is executed and controls the decoding on the NVIDIA Shield.

Hence, the rest of this section is applicable to both unless stated otherwise.

The apps are divided into two main parts. The first part is a renderer library that renders a view of the NGAInteractivity data on an HTML page. Its main task is to receive an NGAInteractivity object and render a DOM subtree that maps one-to-one onto the object's structure. As this is a library that we intend to reuse, we chose to implement it in TypeScript for the reasons mentioned above. In practice this is not strictly required; a plain JavaScript implementation or any other compatible language could also be used.

The second part is the entry point for the browser and consists of a simple HTML page with space for the UI on the bottom. The accompanying JavaScript file loads the MpeghUiLib and renderer libraries and initializes them. It also sets up the connection to the decoder which, as this is not a browser-internal implementation, takes some additional effort. In the prototype, a WebSocket [20] is used for communicating the interactivity information with the external decoder. This connection is also set up by the JavaScript file.

The final result is packed into one easily deployable file by Webpack[2]. An example of a user interface as rendered by the test application can be seen in Fig. 3. The item provides four different presets of which the "Padrão" preset is active. Inside the active preset, the user has access to an audio switch group "Língua" with selectable audio objects "Inglês" and "Francês" the former of which is active at the time.

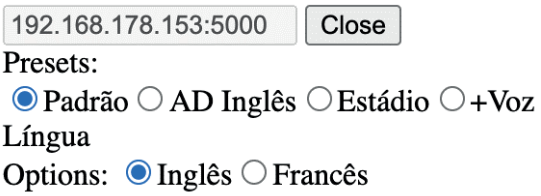


Fig. 3. Illustrates the user interface of the use-case apps.

C. *Prototype System*

The use-case apps generally work fine on their own, but require some additional components to be set up to reflect a real broadcasting scenario. These are a) an external decoder to do the actual MPEG-H Audio decoding and interpretation of the interactivity commands, and b) respective presentation and broadcaster environments that transmit and launch the apps from a broadcasting facility to a consumer's simulated TV set. An architectural overview of the demonstration setup can be seen in Fig. 2.

For the external decoder, we use an Android-based NVIDIA Shield[3] running the Fraunhofer IIS Android Media Player App. The App is capable of decoding MPEG-H Audio from local MP4 [16] files, as well as network streams containing MPEG-H and using transport formats like MPEG DASH [15], Transport Stream [17] or plain fragmented MP4 files with progressive HTTP download. We chose MPEG DASH and progressive MP4 playback over the network as content delivery options for the use-case test applications. This is done to avoid pre-deploying content to the decoding device and instead have the content come from the presentation environment for a more realistic delivery scenario.

In addition to the content delivery via HTTP, the player app listens for JSON-formatted playback commands on a UDP port. These commands also contain the type and URL of the resource to be played back, so the use-case apps have control over the playback, even though they are running on another device.

Furthermore, once the playback is started, the player will listen to WebSocket connections on another TCP port. This second, bidirectional connection is used for the transmission of user interactivity data. The player uses an established connection to send audio scene information including available presets, audio objects, switch groups, labels, and interactive properties with their ranges to the presentation environment. The user interface uses the same connection to send user interactivity commands to the player. All information is sent in an XML [18] format and is a detail of the player implementation.

The decoding is done on an external device in order to demonstrate the capability to deploy the system as a distributed architecture across multiple devices. At the same time, the already available features in the Fraunhofer IIS Android Media Player Application could be used during the evaluation. In practice, the player and the user interface can be encapsulated in a single system; a player can run on the same device as the user interface, they can draw to the same screen or buffer, and the connection does not need to happen

[2] https://webpack.js.org/

[3] https://www.nvidia.com/en-us/shield/

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.4. Web Link: http://dx.doi.org/10.18580/setijbe.2021.4

TABLE XII
APPLICATION CODING REQUIREMENTS (EXCERPT FROM [12])

Use case		Minimum technical specification		Over the air delivery	Internet delivery
AP13	Enable emergency warning information delivery using an interactive application	AP13.1	Emergency warning information interactive application	desirable	desirable
AP14	Support for immersive TV	AP14.4	3D object-based immersive audio interaction	required	required
		AP14.5	3D media positioning and interaction	required	required

over a WebSocket, but can also be achieved through other protocols or natively with a suitable decoder API.

Finally, the presentation environment and the broadcaster environment were submitted as LXD[4] containers due to the easy handling and deployment in a test environment. The broadcaster container is mainly responsible to push the use-case apps and content from the broadcaster to the presentation environment container running on a consumer's simulated TV set. The presentation environment container is then responsible for delivering the content to the decoder (using the Apache 2 HTTP server[5] in the demonstration) and running the use-case apps to render the HTML user interfaces on a Chromium Browser[6].

IV. EVALUATION

The "SBTVD Forum – TV 3.0 – CfP Phase 2 / Testing and evaluation" document [12] defines 17 use-cases for the Application Coding component. Two of these use cases are addressing next-generation audio and all corresponding requirements are fulfilled by our Application Coding proposal for MPEG-H Audio. Table XII lists these fulfilled use-cases and requirements.

For evaluation purpose, we have defined and conducted five specific test procedures, which allow an easy evaluation of the proposed system as required in [12]. Each of these test procedures covers a unique feature of MPEG-H Audio [2] that can be mapped to a concrete requirement specified in [12] as shown in Table XIII. It is important to note, that the more advanced tests (4 and 5) are fulfilling multiple requirements. For example, test procedure 5 is fulfilling AP 13.1 and AP 14.4 from [12].

TABLE XIII
TEST PROCEDURES

Test procedure	MPEG-H Audio feature	SBTVD requirement
1	Preset interactivity	AP 14.4 - 3D object-based immersive audio interaction
2	Switch group interactivity	AP 14.4 - 3D object-based immersive audio interaction
3	Gain interactivity	AP 14.4 - 3D object-based immersive audio interaction
4	Position interactivity	AP 14.5 –3D media positioning and interaction
5	Emergency warning	AP 13.1 – emergency warning information interactive application

As an example, for one of these test procedures, Fig. 4 illustrates the MPEG-H Audio UI of test procedure 3. The gain of the "Língua" switch group can be decreased or increased by moving the corresponding "ProminenceLevel" slider to the left or right respectively. Since MPEG-H Audio [2] includes advanced Loudness Normalization functionality [7], the overall loudness of the audio output will stay the same, and only the balance between the "Língua" object and the rest of the audio scene will change.

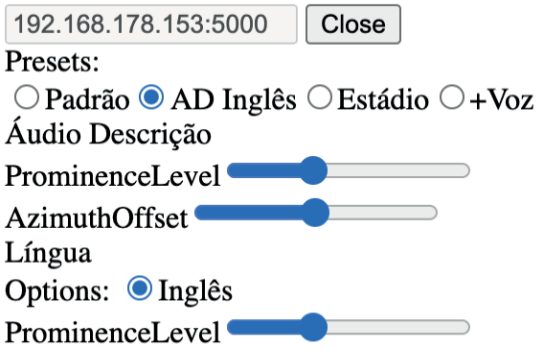


Fig. 4. Illustrates the prototype UI while decoding an audio item with 4 presets. The "ProminenceLevel" slider can be used to control the balance between the corresponding audio switch group or audio object in relation to the overall audio scene.

V. OUTLOOK

The previous sections describe the design, implementation and internal evaluation of our approach to enable next-generation 3D interactive audio for the Application Coding component of the SBTVD TV 3.0 Project using the MPEG-H Audio system. Our proposal is a proof-of-concept that demonstrates the required features according to the TV 3.0 CfP [1]. After the TV 3.0 Phase 2 evaluation process will be finalized, it is foreseen that all accepted proposals will be adapted and further refined for standardizing the best solutions for the TV 3.0 Application Coding layer. As an active member of the SBTVD Forum, Fraunhofer IIS will continue to support the standardization effort and is committed to work closely with the SBTVD Forum experts for integrating a 3D object-based audio API into the TV 3.0 suite of standards according to the Forum decisions.

It should be noted that the user interaction in the illustrated scenario benefits from bidirectional communication between the decoder and the user interface component. The decoder needs to be notified about user interactions and has to adapt the audio scene rendering accordingly. On the other hand, the rendered user interface depends on the metadata present in the audio stream and has to react to changes therein. For ensuring a high quality of experience, it is desired to have low latency between an interaction and the expected effect, which is achieved with a close integration of the two components.

[4] https://linuxcontainers.org/lxd/introduction/
[5] https://httpd.apache.org/
[6] https://www.chromium.org/Home

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.4. Web Link: http://dx.doi.org/10.18580/setijbe.2021.4

The event-based design of our proposal enables this close integration and could optimally be achieved by a native implementation in Ginga-HTML5. As a less strict option, bidirectional communication via WebSockets can be used to enable this form of close integration of the components, as shown in the submitted prototype.

It would also be possible to implement the desired behavior as a library on top of a REST API in Ginga-CC-WebServices similar to the existing SBTVD TV 2.5 specification [10]. However, as this is a fundamentally unidirectional mode of communication, active polling of the REST API or a different method of notifying the user interface about bitstream-triggered UI changes has to be implemented in order to ensure a good quality of the experience.

Similarly, the submitted setup for TV 3.0 is using an MPEG-H decoder running on an external device with stereo or binaural headphone output for the prototype. However, the described API design is not restricted to this setup and can also be used with other audio outputs like immersive AVRs, soundbars or integrated TV speakers.

VI. Conclusion

With its standardized metadata and decoder interfaces, the MPEG-H Audio system [2] offers a stable foundation for extending HTML5 web APIs for controlling advanced audio features in a device-independent and interoperable fashion. It was proposed as a candidate technology for the Application Coding component of the SBTVD TV 3.0 Project Call for Proposals [1]. This API extends the Ginga-HTML5 standard [9] and enables applications to offer user interaction with MPEG-H interactivity features, like switching presets and audio switch groups, as well as adjusting the prominence level and position of audio objects.

A prototype JavaScript implementation of the API was submitted to the SBTVD Forum for testing and evaluation. Accompanying runtime prototypes for broadcast and receiver environments show the feasibility of the proposal. The prototype environments employ a distributed architecture including an external decoder, illustrating the flexibility of the MPEG-H ecosystem.

Furthermore, following the evaluation test procedures described in this paper, the proposed API fulfills the TV 3.0 Application Coding requirements for “3D object-based immersive audio interaction”, “3D media positioning and interaction” and support for “emergency warning information interactive applications”.

Acknowledgment

The authors would like to thank the SBTVD Forum for the efforts to organize the TV 3.0 Project and support us during the development of the proposal, especially Professor Marcelo F. Moreno, who is chairing the Application Coding working group in the SBTVD Technical Module and has led the technical foundation of the proposal together with Rafael Diniz in the paper on immersive audio properties for NCL media elements [11].

References

[1] Brazilian Digital Terrestrial Television System Forum. (2020, July 17). “Call for Proposals: TV 3.0 Project”. [Online]. Available: https://forumsbtvd.org.br/wp-content/uploads/2020/07/SBTVDTV-3-0-CfP.pdf

[2] “Information technology - High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D audio,” International Organization for Standardization (ISO), Geneva, Standard ISO/IEC 23008-3:2019, 2019.

[3] Fraunhofer Audio Blog. (2019, Aug. 27). “MPEG-H Audio selected to enhance Brazilian digital television with immersive and personalized sound”, [Online]. Available: https://www.audioblog.iis.fraunhofer.com/mpegh-brazil-isdbtb

[4] “Televisão digital terrestre - Codificação de vídeo, áudio e multiplexação - Parte 2: Codificação de áudio,” ABNT NBR 15602-2:2020. [Online]. Available: https://forumsbtvd.org.br/legislacao-e-normas-tecnicas/normas-tecnicas-da-tv-digital/english/

[5] “Televisão digital terrestre - Multiplexação e serviços de informação (SI),” ABNT NBR 15603:2020. [Online]. Available: https://forumsbtvd.org.br/legislacao-e-normas-tecnicas/normas-tecnicas-da-tv-digital/english/

[6] “Televisão digital terrestre – Receptores,” ABNT NBR 15604:2020. [Online]. Available: https://forumsbtvd.org.br/legislacao-e-normas-tecnicas/normas-tecnicas-da-tv-digital/english/

[7] R. L. Bleidt et al., “Development of the MPEG-H TV Audio System for ATSC 3.0,” in *IEEE Transactions on Broadcasting*, vol. 63, no. 1, pp. 202-236, March 2017, doi: 10.1109/TBC.2017.2661258.

[8] A. Murtaza and S. Meltzer, “First Experiences with the MPEG-H TV Audio System in Broadcast,” *SET INTERNATIONAL JOURNAL OF BROADCAST*, ISSN Print: 2446-9246 ISSN [Online]. 2446-9432. doi: 10.18580/setijbe.2018.6. Available: https://www.set.org.br/ijbe/ed4/Artigo%206.pdf

[9] “Televisão digital terrestre - Codificação de dados e especificações de transmissão para radiodifusão digital - Parte 10: Ginga-HTML5 - Especificação do perfil HTML5 no Ginga”, ABNT NBR 15606-10:2021. [Online]. Available: https://forumsbtvd.org.br/legislacao-e-normas-tecnicas/normas-tecnicas-da-tv-digital/english/

[10] “Televisão digital terrestre - Codificação de dados e especificações de transmissão para radiodifusão digital - Parte 11: Ginga CC WebServices - Especificação de WebServices do Ginga Common Core”, ABNT NBR 15606-11:2021. [Online]. Available: https://forumsbtvd.org.br/legislacao-e-normas-tecnicas/normas-tecnicas-da-tv-digital/english/

[11] R. Diniz and M. Moreno. “Immersive audio properties for NCL media elements”, in *Anais Estendidos do XXV Simpósio Brasileiro de Sistemas Multimídia e Web, Florianópolis*, 2019, pp. 195-197, doi: https://doi.org/10.5753/webmedia_estendido.2019.8164. W.-K. Chen, *Linear Networks and Systems*. Belmont, CA: Wadsworth, 1993, pp. 123–135.

[12] Brazilian Digital Terrestrial Television System Forum. (2021, Mar. 15). “CfP Phase 2/Testing and Evaluation: TV 3.0 Project”. [Online]. Available: https://forumsbtvd.org.br/wp-content/uploads/2021/03/SBTVD-TV_3_0-P2_TE_2021-03-15.pdf

[13] C. McCormack. (2016, Dec. 15). WebIDL Level 1. *W3C Recommendation* [Online]. Available: https://www.w3.org/TR/WebIDL-1/

[14] ECMA International. “ECMAScript® 2021 language specification,” ECMA-262, 12th edition, June 2021. [Online]. Available: http://www.ecma-international.org/publications/standards/Ecma-262.htm

[15] “Information Technology - Dynamic Adaptive Streaming Over HTTP (DASH) -- Part 1: Media Presentation Description and Segment Formats,” International Organization for Standardization (ISO), Geneva, Standard ISO/IEC 23009-1:2019, 3th edition, 2019.

[16] “Information Technology - Coding of Audio-Visual Objects -- Part 12: ISO Base Media File Format,” International Organization for Standardization (ISO), Geneva, Standard ISO/IEC 14496-12:2020, 6th edition, 2020.

[17] “Information Technology - Generic coding of moving pictures and associated audio information -- Part 1: Systems,” International Organization for Standardization (ISO), Geneva, Standard ISO/IEC 13818-1:2018, 6th edition, 2018.

[18] T. Bray, J. Paoli, M. Sperberg-McQueen. (2008, Nov.). Extensible Markup Language (XML) 1.0. *W3C Recommendation.* [Online]. Available: https://www.w3.org/TR/xml/

[19] Y. Zhu et al. (2021, Jun.). DOM Living Standard. *W3C Recommendation.* [Online]. Available: https://www.w3.org/TR/dom/

[20] I. Fette, A. Melnikov. “The WebSocket Protocol” *IETF, Request for Comments (RFC 6455)*, Dec. 2011. [Online]. Available: https://rfc-editor.org/rfc/rfc6455.txt

[21] S. Faulkner, A. Eicholz, T. Leithead, A. Danilo, S. Moon. “HTML 5.2.” *W3C Recommendation, Jan. 2021.* [Online]. Available: https://www.w3.org/TR/html52/

Oliver Major received both his B.Sc. and his M.Sc. degrees in computer science at the RWTH Aachen University in Aachen, Germany in 2014 and 2016 respectively.

Between 2014 and 2015, he gained some experience with network communication and integrated platforms as a Student Assistant at devolo AG in Aachen, Germany. After his studies, he joined the Mobile Audio Rendering group at the Fraunhofer Institute for Integrated Circuits (IIS) in Erlangen, Germany as a Research Engineer in 2017, where his main topics were binaural rendering, platforms and web technologies. After joining the Web Media Technologies group in 2020, his focus shifted more towards system architecture and web technologies.

Ziad Shaban received his B.Sc. degree in electrical engineering - communications and electronics from the Jordan University of Science and Technology, Irbid, Jordan in 2013 and his M.Sc. degree in communication and multimedia engineering from the Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany, in 2015.

Between 2013 and 2016, he was active as a Research Assistant in the fields of software-defined radio and wireless communication at the Fraunhofer Institute for Integrated Circuits (IIS) in Erlangen, Germany, where he also wrote his thesis titled “A Study of a Testbed for Multi-channel Simulators”. In 2016, he joined the Multimedia Transport group of Fraunhofer IIS as a Research Engineer with a focus on the research and development of streaming technologies. As of 2020, he has been a member of the Web Media Technologies group.

Bernd Czelhan received the B. Sc. in 2011 and M. Sc. in 2012 degrees in Computer Science from the Technische Hochschule Nürnberg Georg Simon Ohm, in Nuremberg, Germany.

In 2012, he joined the Fraunhofer Institute for Integrated Circuits (IIS) as a research engineer, where his main working topic is the next-generation audio codec MPEG-H and Web development. Since 2020, he is heading the Web Media Technologies group. In addition, he is supporting the practical implementation of MPEG-H. He is especially interested in Web technologies and modern transport mechanism and system aspects of today’s audio codecs, such as MMT, DASH/Route, and hybrid delivery.

Adrian Murtaza received his M.Sc. degree in Communication Systems from the École Polytechnique Fédérale de Lausanne, Switzerland in 2012 with a thesis on “Backward Compatible Smart and Interactive Audio Transmission”. Upon graduation he joined Fraunhofer IIS, where he works as a Senior Manager, Technology and Standards.

Adrian joined MPEG in 2013 and since then contributed to development of various audio technical standards in MPEG-D and MPEG-H. He serves as Fraunhofer's Standards Manager in a number of industry standards bodies, including SBTVD, ATSC, CTA, DVB, HbbTV and SCTE, and is the co-author of multiple specifications in those groups.

More recently he focused on specification of Next-Generation Audio delivery and transport in ATSC 3.0 systems and MPEG-2 Transport Stream based systems, as well as on enabling of MPEG-H Audio services in different broadcast and streaming ecosystems. With a strong interest in VR/AR media solutions he is actively involved in MPEG-I efforts targeting future immersive applications.

Received in 2021-08-16 | Approved in 2021-12-07

https://set.org.br/ijbe/ doi: 10.18580/setijbe.2021.4. Web Link: http://dx.doi.org/10.18580/setijbe.2021.4