EBU

OPERATING EUROVISION AND EURORADIO

TECHNICAL REVIEW

Scene-Based Audio and Higher Order Ambisonics:

A technology overview and application to Next-Generation Audio, VR and 360° Video

NOVEMBER 2019

Ferdinando Olivieri, Nils Peters, Deep Sen,
Qualcomm Technologies Inc., San Diego, California, USA

The main purpose of an EBU Technical Review is to critically examine new technologies or developments in media production or distribution. All Technical Reviews are reviewed by 1 (or more) technical experts at the EBU or externally and by the EBU Technical Editions Manager. Responsibility for the views expressed in this article rests solely with the author(s).

To access the full collection of our Technical Reviews, please see: https://tech.ebu.ch/publications .

If you are interested in submitting a topic for an EBU Technical Review, please contact: tech@ebu.ch

1. Introduction

Scene Based Audio is a set of technologies for 3D audio that is based on Higher Order Ambisonics. HOA is a technology that allows for accurate capturing, efficient delivery, and compelling reproduction of 3D audio sound fields on any device, such as headphones, arbitrary loudspeaker configurations, or soundbars.

We introduce SBA and we describe the workflows for production, transport and reproduction of 3D audio using HOA. The efficient transport of HOA is made possible by state-of-the-art compression technologies contained in the MPEG-H Audio standard. We discuss how SBA and HOA can be used to successfully implement Next Generation Audio systems, and to deliver any combination of TV, VR, and 360° video experiences using a single audio workflow.

1.1 List of abbreviations & acronyms

CBA	Channel-Based Audio
SBA	Scene-Based Audio
HOA	Higher Order Ambisonics
OBA	Object-Based Audio
HMD	Head-Mounted Display
MPEG	Motion Picture Experts Group (also the name of various compression formats)
ITU	International Telecommunications Union
ETSI	European Telecommunications Standards Institute

2. Next-Generation Audio

Next Generation Audio (NGA), defined in DVB ETSI TS 101 154 [1], introduces the following audio features:

- the provision of **immersive audio** (e.g. the inclusion of height elements),
- the possibility for end users to **personalize** the content,
- the introduction of Audio Objects to facilitate both immersive and personalized audio.

According to its definition, NGA represents a significant shift [2] from traditional channel-based audio systems (e.g. stereo or 5.1) and offers new opportunities for content producers, broadcasters, technology providers, and consumers alike [3]. Since the inception of the NGA requirements, new technologies have proliferated [2], [4] and various institutions are providing guidelines [1], [5] to enable a coherent growth of the NGA ecosystem, such as the EBU's open source NGA renderer (EAR) [6] which is based on the Audio Definition Model (ADM) [7].

CBA formats such as stereo, 5.1[1], etc. have so far been the prevailing audio formats used in broadcasting. However, CBA formats would be unsuitable to fulfil all the requirements set by the definition of NGA. For example,

- channel based formats implicitly assume that the sound reproduction system used by the consumer matches the sound system used to produce the content. This assumption generally does not hold true. What typically happens is that content produced for a 5.1 speaker layout is consumed using headphones, for example.
- Given the inhomogeneity of the audio reproduction ecosystem and the plethora of reproduction devices in the market, NGA systems must be designed to optimally render the audio content according to the consumer's reproduction environment. This includes, but is not limited to, headphone reproduction, non-standard loudspeaker layouts, and soundbars. Clearly, channel-based content does not allow for this. In fact, CBA content is merely a set of loudspeaker feeds and not a description of the audio scene as intended by the content creator.
- In applications where customization of the content is needed, e.g. to adjust the relative level of the commentary versus the ambience to enhance intelligibility, channel-based content also fails because the commentary is usually mixed into the loudspeaker feeds.

To address the limitations of using Channel Based Audio (CBA) as the sole format for NGA, two audio formats, Object Based Audio (OBA) and Scene Based Audio (SBA) have been proposed for the implementation of the NGA requirements.

The three audio formats (CBA, OBA, and SBA) are recognized by the International Telecommunication Union [7] and they are simultaneously supported by the MPEG-H Audio standard [5], a comprehensive portfolio of technologies that provides broadcasters with the ability to transmit immersive audio content using any combination of CBA, OBA, or SBA [8].

For example, a content creator may decide to produce the commentaries using OBA (dialogues and audio descriptions, all in multiple languages) whilst SBA can be used to produce the Music and Effects (M&E) component. This scenario is described in more detail below.

This article primarily focuses on the SBA format and the MPEG-H Audio System. More information on CBA and OBA is available in [9] and [5].

[1] Broadcasters don’t habitually make use of the “0.1” LFE channel and so broadcast a 5.0 or 7.0 surround signal.

3. An overview of SBA and HOA technology

The SBA portfolio of technologies for immersive audio aims at accurate capturing of live audio, intuitive tools and workflows with low complexity post processing to produce 3D audio content, efficient compression and transport and accurate reproduction of 3D audio content of arbitrary audio scene complexity.

Under the hood, SBA is based on Higher Order Ambisonics (HOA) [10], a format for the modelling of 3D sound fields defined on the surface of a sphere. SBA and HOA are deeply interrelated, to the extent that the two terms are often used interchangeably. Whilst SBA may be thought of as a portfolio of technologies for 3D audio, HOA can be viewed as the underlying format that enables the technologies which are part of the SBA portfolio.

Ambisonics was described by Michael Gerzon in 1973 [11]. Based on psycho acoustical considerations, Gerzon devised a mathematical framework for the capture and reproduction of immersive sound. Jérôme Daniel later extended and generalized Ambisonics to Higher Order Ambisonics [12]. After Daniel's publications, Gerzon's work was referred to as First Order Ambisonics (FOA).Since then, HOA has triggered an enormous interest in the research community and with audio engineers alike [13], [14].

As previously stated, HOA is a 3D audio format [15] that is based on a mathematical framework for the modelling of 3D sound fields defined on a spherical surface [13]. More specifically, the modelling consists of a projection of the spatial sound field (sampled at several points on a sphere) onto a set of *spherical harmonics*, a set of special functions. The result of this "projection" is, in turn, a set of signals hereafter referred to as HOA signals, which contain the spatial information of the original sound field. The operation of transforming spatial sound fields to HOA coefficients is also known as the HOA transform[2].

In layman's terms, the HOA signals contain a loudspeaker agnostic representation of the original sound field wherein the sound sources are mixed into a fixed number of HOA signal channels. These HOA signals ("mixing coefficients") are calculated based on the spatial location of the sound sources and their audio signals. At the reproduction site, a renderer transforms the HOA signals into loudspeaker feeds according to the actual loudspeaker configuration used for reproduction.

One remarkable property of HOA is that the production and the reproduction stages are completely decoupled from each other [12], [13], meaning that content produced in the HOA format can be reproduced on any loudspeaker layout (including mono, stereo, surround speakers, and irregular layouts) [12], [16] as well as on headphones and soundbars. For this reason, Ambisonics has often been referred to as a future

[2] The "HOA transform" is often referred to as "HOA encoding". Similarly, the operation of "HOA rendering" is often referred to as "HOA decoding". In this article, we use the terms "encoding" and "decoding" from a perceptual audio coding perspective.

proof format for audio productions [17, p.] as it does not “constrain” producers and consumers to, respectively, produce and consume 3D audio content on a reproduction environment.

The performance of HOA systems increase with increasing values of the HOA order N [13]. This has been shown and validated by means of numerous objective and subjective experiments. More specifically, as the HOA order, N increases,

- the accuracy of the HOA representation of the original sound fields increases [22],
- localization accuracy increases [23],
- the size of the “sweet area” at the reproduction side increases [24].

The drawback is that the number of HOA signals necessary to represent a 3D sound field[3] increases with the HOA order, N. This can be solved using modern signal processing techniques for transmission, described below, thus making HOA implementable in practice.

Other concepts that are important for proper functioning in HOA systems are

- the HOA channel ordering, such as Single Index Designation (SID) [18], Furse Malham [19], or Ambisonics Channel Number (ACN) [20], and;
- the HOA gain normalization schemes, such as *SN3D*, *N3D* and *maxN* [21].

In recent years, the combination of ACN order and *SN3D* gain normalization became popular. For the HOA systems to work, it is important to keep consistency between the HOA normalization and channel ordering conventions of choice. It is nevertheless possible to convert between different conventions [21].

3. The three audio formats for NGA production

Although CBA, OBA, and SBA can all be used to produce NGA content there are distinct differences with respective advantages and disadvantages between them. To highlight these differences, let’s consider an immersive audio production of a live event such as a football match, illustrated by Figure 1.

Spot microphones (3, 4, 5, & 6) are installed around the football pitch. In addition to these, two live commentaries, in English (1) and in French (2), are provided. All the audio signals are routed to the production room (e.g. a broadcast truck) where the mixing engineer mixes the signals and produces an audio output.

For simplicity, it is assumed that the mixing console can provide output in all the immersive audio formats, but that the broadcaster can only select one format for the delivery of the content.

[3] The number of HOA signals is $(N+1)^2$ for a 3D sound field or $2N+1$ for a 2D sound field [18].

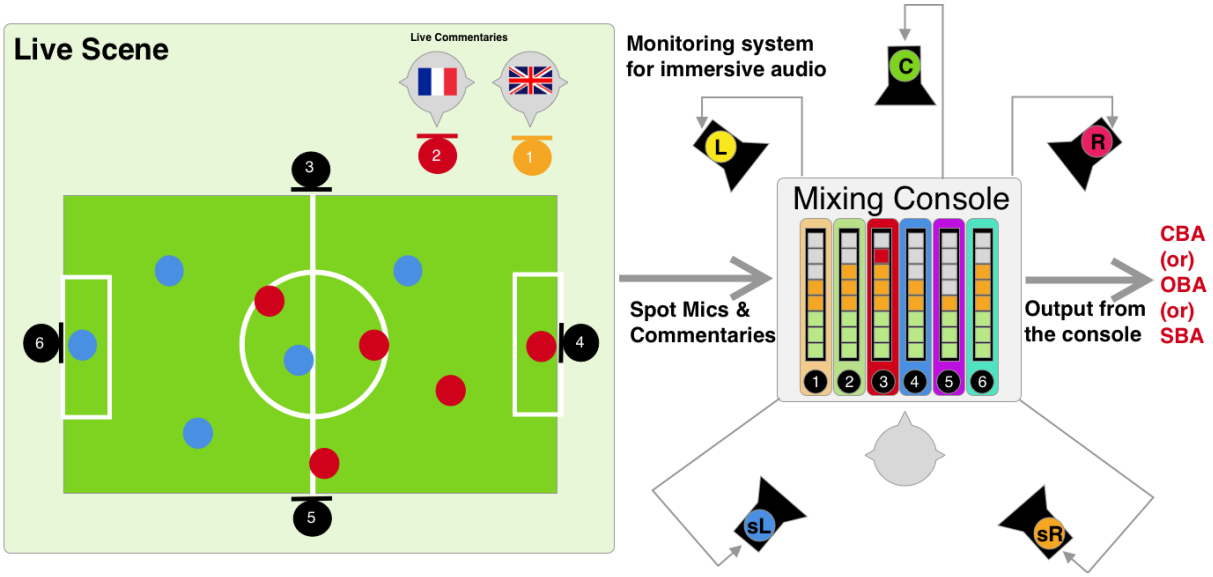

Figure 1: Live event with multiple microphones; the mixing engineer's environment

Two broadcasters, "Broadcaster 1" and "Broadcaster 2", take the output (in one of the three audio formats) from the console and deliver it to their end-users. At the listening side, the listeners receive the audio content, which is reproduced on their respective audio systems.

Let us analyse the implications for the listeners if the broadcasters selected **Channel Based Audio** as the format for their content delivery. In CBA the audio material (the signals from the spot microphones, commentary and/or pre-recorded stems) is mixed for a specific loudspeaker layout (e.g. stereo or 5.1). The end users, in order to perceive the (virtual) audio image intended by the content creator, have to place their loudspeakers as dictated by the channel based format in which the content was created, a requirement which may or may not be easy to fulfil in practical domestic scenarios.

If the production is solely based on the CBA format, the mixing engineers will need to produce different mixes for different target loudspeaker layouts and, in our live football production use case, two languages of commentary. The different mixes (e.g. 2.0 and 5.0) will then be delivered by different broadcasters. Because of bandwidth constraints, each broadcaster may choose to simultaneously deliver the stereo and the 5.0 mixes with a single language "baked in" the audio mix.

With reference to Figure 1, let us assume that Broadcaster 1 delivers both the 5.0 and stereo mixes with English commentary and Broadcaster 2 delivers both the 5.0 and stereo mixes with French commentary.

At Listening Site #1, the listener has a 5.0 loudspeaker system, but for reasons of domestic harmony the loudspeaker positions are not in the ideal 5.0 layout and neither can the front L and R loudspeakers be used to satisfactorily present the

stereo signal. Because the loudspeaker layout at the listening site differs from that intended by the content producer, the listener may have a suboptimal experience.

At Listening Site #2, there is a different issue. This time the English-speaking user has a loudspeaker setup that matches that intended by the content producer for the content delivered in stereo by Broadcaster 2, but the commentary is not in his/her native language, which is again a suboptimal experience. In this case the suboptimal experience is due to the lack of a personalization option (i.e., the possibility of selecting a different language).

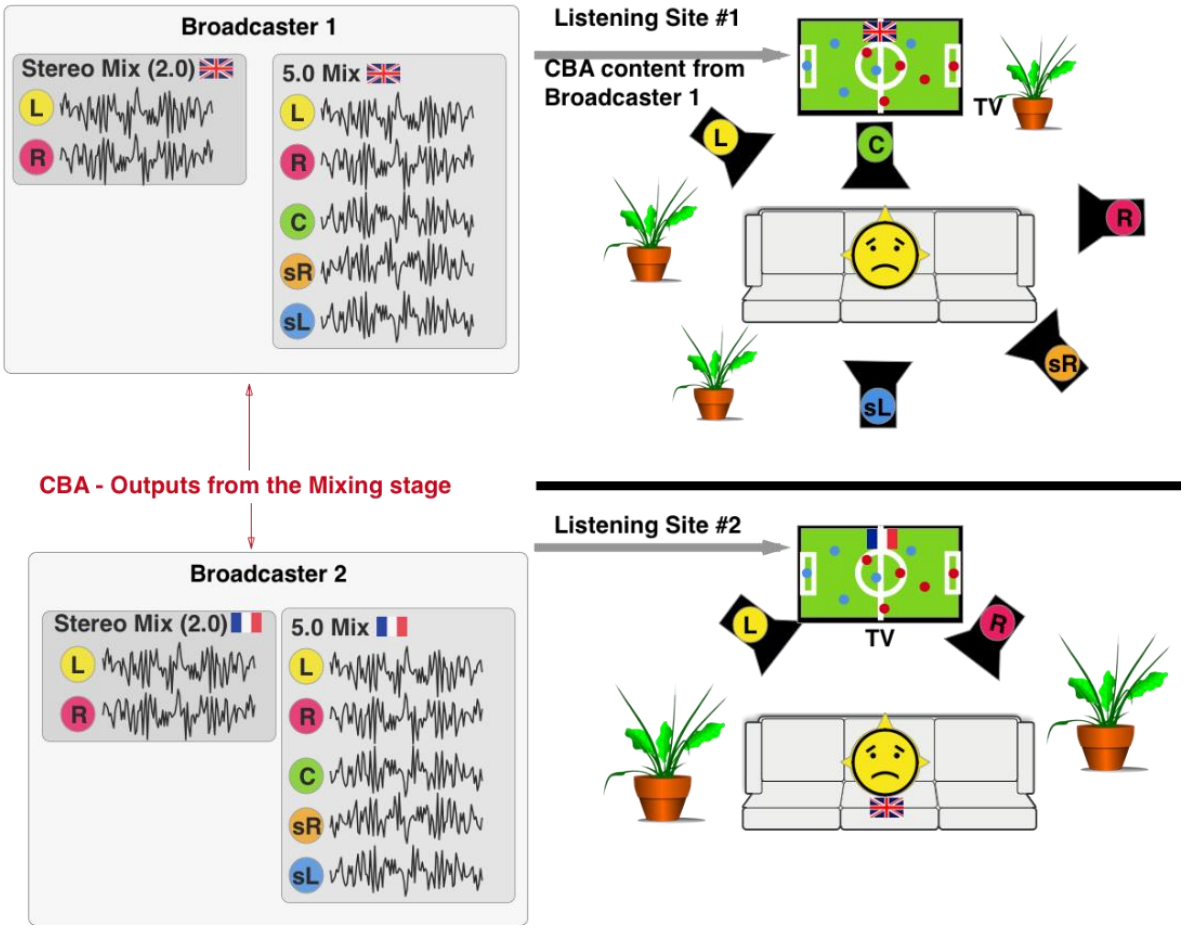

Figure 2: Delivery and reproduction of Channel Based Audio Content

There are numerous reasons why CBA alone may not be a good fit for NGA:

- Multiple languages require multiple mixes. This may be expensive to produce. Furthermore, bandwidth constraints limit the simultaneous broadcast of multiple mixes.
- CBA content cannot be personalized. As an example, the user may wish to customize the relative level of the commentary versus the Music & Effects (M&E) to increase the intelligibility of the commentary. In order to achieve this with CBA content, additional processing at the rendering side is required which, for example, separates the speech from the background

music, an operation which may or may not be successful, depending on the content and on the algorithm employed for the separation.

- The rendering is not flexible, meaning that if the loudspeaker layout at the listening site does not match that used for the mixing, the listening experience may be suboptimal.

In short, with CBA, end users will only receive audio in the format that the content producer and broadcasters make available (e.g. stereo, 5.1, etc.). Furthermore, personalization options are limited.

In contrast to CBA, **Object based Audio (OBA)** comes in the form of unmixed audio components (also referred to as audio objects) with associated object metadata. This object metadata contains information that characterizes the audio object, such as its spatial position in the scene and its audio level, as intended by the content creator.

At the listening site, an OBA renderer attempts to reproduce the audio objects based on the specific loudspeaker configuration available (which may be non-standard) [4]. In contrast to CBA, OBA enables interactive features by allowing users to customize specific characteristics of audio objects (if allowed by the content creator and supported by the renderer), e.g. altering the spatial position or customizing the levels of a given object.

OBA unlocks interesting possibilities for both content producers and end users alike. For example, in the live sport production scenario considered previously, the mixing engineer may deliver the two commentaries as audio objects (separated from the mix). This would solve the problem of the language selection in Listening Site #2 (shown in the lower half of Figure 2).

In fact, if the content were produced and delivered in OBA, the English-speaking user at Listening Site #2 would be able to select the English commentary, as shown in Figure 3. In addition to this, the user may also customize the relative level of commentary and M&E.

However, there is still the issue of how the M&E is produced (i.e., in which audio format) and how it is delivered, which we shall now discuss.

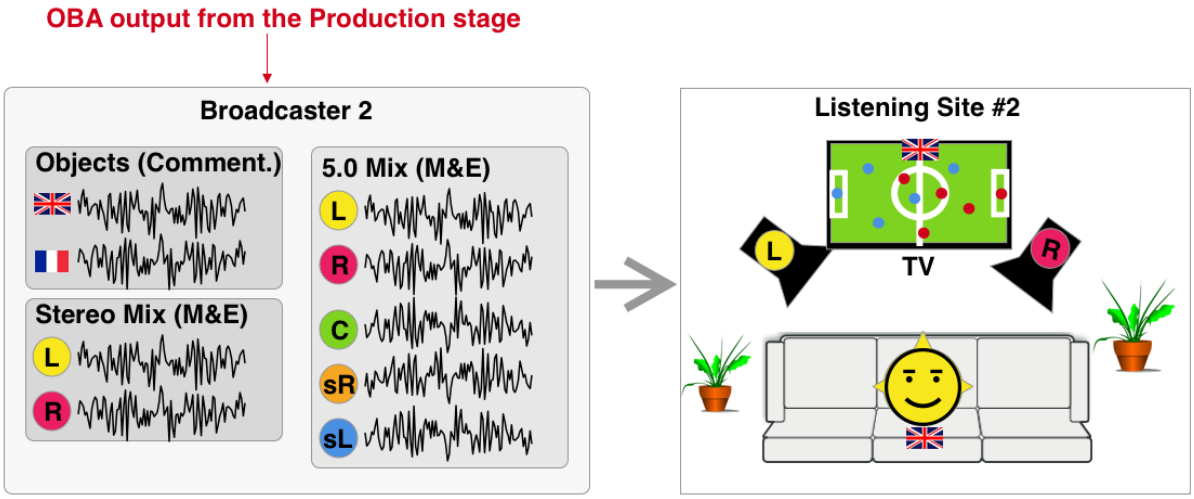

Figure 3: OBA output and the listening environment

From an NGA perspective, OBA solves many of the constraints present in CBA, including the possibility of customizing the relative levels of dialogue and M&E. However, using OBA only workflows would present the following limitations:

- The bandwidth required to transmit individual objects as well as their associated decoding complexity increases with the number of objects. Whilst it may be fair to assume that for most broadcasting applications the number of objects would not be excessive (if the M&E components are transmitted using CBA), there are other use cases (such as episodic content or VR) where the number of objects may be significant.

- In an OBA scenario, the M&E component is usually produced for a given loudspeaker layout (e.g. 7.1+4), hence relying on the CBA format. As previously stated, at the listener site, if the loudspeaker layout does not match that used for the production, additional processing may be required to adapt the content produced for the original layout (e.g. 7.1+4) to the target layout at the listening site (e.g. 5.1). This operation is typically called down mix (or up mix, depending on the relation between original and target layouts, for example to go from 5.1 to 7.1+4) and it relies on processing algorithms available in the reproduction system. The bottom line is that if the M&E is delivered in the CBA format, then the drawbacks of CBA apply.

- Indeed, the M&E can be delivered as audio objects, in which case the OBA flexible renderer would be able to overcome all the problems related to a CBA M&E. However, as previously stated, the bandwidth required to transmit many objects (audio and metadata) may be prohibitive and hence not practical in many applications.

- The complexity of the object renderer increases with the number of objects. Whilst this may not be of direct concern to broadcasters, it would be a concern for technology providers. For content consumed on mobile devices

where the limited availability of power resources calls for limited complexity of rendering and personalization algorithms this may be particularly problematic.

- In VR scenarios, where the sound scene must be adapted depending on the movements of the listener's head, the paradigm objects plus CBA M&E may not provide convincing virtual images for the listener. Whilst objects may be rotated (and the complexity of the rotation operation increases as the number of objects increases) and hence adapted to the movements of the head, the CBA M&E component may not be easily manipulated, which may cause perceptual artefacts that degrade the overall VR experience.

In order to mitigate these problems that may arise with OBA from limited delivery bandwidth, complexity constrains of the consumer device and scene manipulation, immersive audio content can be produced and delivered in the **Scene based Audio (SBA)** format.

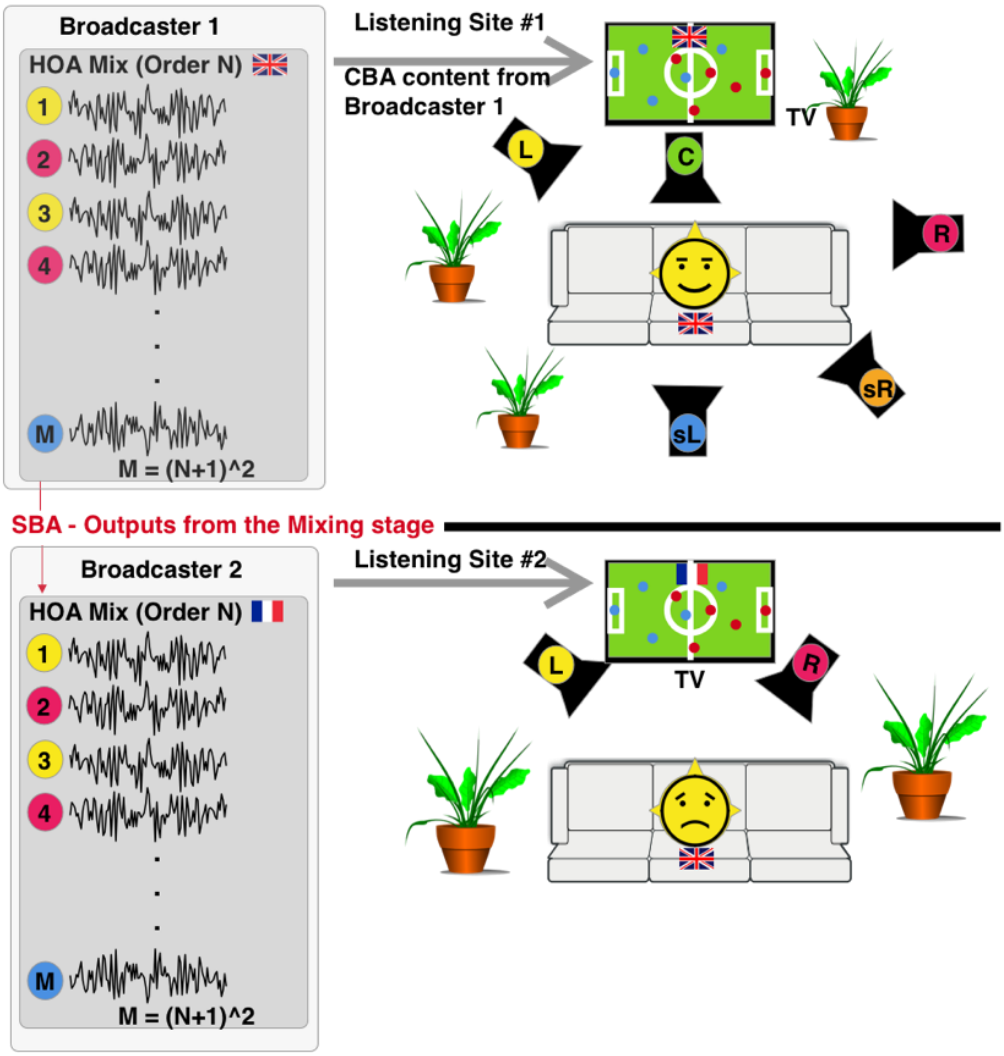

Figure 4: The SBA listening scenario

As previously explained, SBA is based on **Higher Order Ambisonics (HOA)**. Content in the SBA format is encoded into a fixed set of audio signals whose number is independent of the number of audio elements (such as microphone signals or stems) featured in the original content.

As with OBA, the SBA format is agnostic to the loudspeaker layout used for reproduction, thus allowing the rendering of SBA content on arbitrary loudspeaker layouts. Additionally, the SBA format enables users to personalize and interact with the immersive audio content. Returning to the immersive audio production of a live football match, assuming an SBA only production (meaning that the content is mixed and delivered in HOA), the listener at Listener Site #1 with poorly positioned loudspeakers will take advantage of the flexible HOA renderer. However, at Listening Site #2, the listener will not be able to select a different language.

Each audio format has its own advantages and disadvantages when it comes to implementing the NGA paradigm. It would be best to use a combination of formats for NGA purposes, more specifically, the best combination of formats that solves all the problems listed above is OBA with SBA. This option is fully supported by the MPEG-H Audio standard.

4. SBA in practice, from content producers to consumers via broadcasters

The challenges of deploying an audio workflow based on HOA for broadcasting were originally discussed by Gerzon [25], [11], [26] between 1973 and 1985. Since then, technological advances in audio processing techniques based on HOA have made the SBA workflow implementable in practice [10], [27], [28]. Broadly speaking, using MPEG-H Audio, the SBA workflow consists of three stages:

- **Production stage**: the mixing engineer creates the 3D audio content in HOA by mixing various audio sources (feeds from spot microphones, stems, Ambisonics microphones, etc.) by using appropriate tools to perform the HOA transform.
- **Transport stage**: the set of HOA signals are compressed and sent to the end user via an MPEG-H Audio bitstream.
- **Reproduction stage**: the MPEG-H Audio decoder at the user's end receives and decodes the MPEG-H Audio bitstream to retrieve the HOA signals. The HOA signals can then be further manipulated and customized (e.g. rotation of the sound field in VR applications or audio "zoom" in a desired direction). Finally, the HOA renderer creates the appropriate feeds for the reproduction device.

  Dialogues, commentaries, or audio descriptions can be sent as separate audio objects, as required.

4.1 Production stage

HOA provides mixing engineers with an unprecedented flexibility for 3D audio content creation. Any audio source can be transformed to HOA [29]. This includes any combination of live feeds from spot microphones, pre-recorded audio stems, as well as signals captured live or pre-recorded with Ambisonics microphones (of any order), to name but a few. In addition to this, previously produced HOA content (e.g. HOA content stored in the ADM format [7]) can be added to a HOA mix.

The signal flow in a typical HOA production is shown in Figure 5. Even though the elements in Figure 5 are based on the Qualcomm® 3D Audio tools[4] HOA plugin suite (available for download at [30]), similar principles apply to HOA workflows implemented in software (e.g. DAW plugins) or hardware (e.g. mixing consoles).

Referring to Figure 5, mono or stereo audio feeds (either previously recorded or captured live) are processed with the so called HOA panner, a tool that allows mixing engineers to produce a “virtual” 3D sound scene by positioning the audio material (e.g. stems and microphone feeds) in the 3D space.

The HOA panner creates the HOA signals based on:

- the audio inputs (e.g. the audio objects and spot microphones) and,
- the properties of the sound source (such as the position of the sound source in the 3D space and its width).

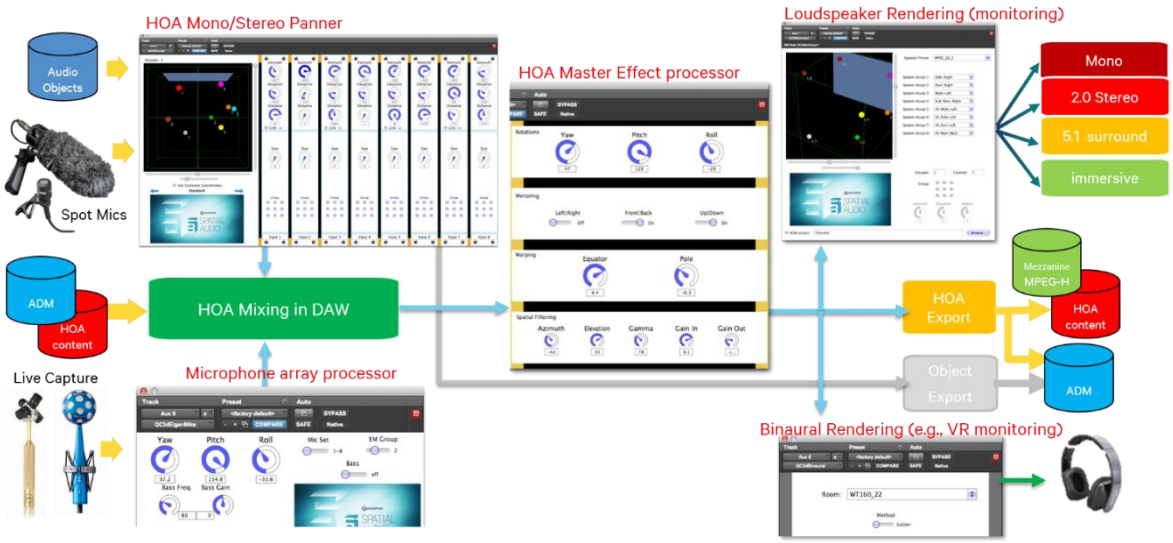

Figure 5: An example of a 3D audio production workflow based on HOA and the Qualcomm® 3D Audio tools suite [30]

In addition to specifying the position of the sources in the HOA scene, the mixing engineer can apply spatial effects such as spatial filters to enhance (or attenuate) sounds from certain regions of the space (e.g. attenuating the sound coming from a

[4] Qualcomm 3D Audio tools are a product of Qualcomm Technologies, Inc. and/or its subsidiaries.

PA system or a noise source). If any of the mono or stereo sources are marked as “objects” they are sent directly to the “Object export” bus and are not mixed in the HOA signals.

The HOA workflow for 3D audio is compatible with existing audio workflows. For example, engineers can apply sound effects (such as equalization, compression, etc) to a given channel strip before injecting the audio in the HOA panner module.

In addition to the items processed by the HOA panners, a 3D sound field can be enhanced with signals captured by microphone arrays, an option which may be particularly useful in live scenarios (e.g. a sport event, a concert, etc.). The feeds from the microphone arrays are transformed to HOA via appropriate processing such as the “Microphone Array processor” module in Figure 5 [30]. The HOA signals generated by each instance of the HOA panner and/or Microphone Array processor are summed in the HOA mixer[5].

At this point, the HOA signals contain the HOA representation of the whole 3D sound field designed by the mixing engineer.

If desired, the HOA signals can be further processed by the “HOA Master Effect Processor” for further spatial effects, such as rotation of the entire sound scene, mirroring, warping, or spatial filtering (e.g. audio zooming to a specific direction). For example, the rotation of the sound field is particularly useful to align the sound field with the camera view.

The interested reader may find more information on HOA effects in [14].

Sound engineers can take advantage of the Loudspeaker and Binaural Rendering modules for monitoring purposes.

By virtue of the HOA flexible rendering, a sound field represented in the HOA format can be reproduced on any loudspeaker layout (including standard and non-standard layouts). A viable choice for HOA rendering is the open source EBU NGA renderer [6].

HOA can be reproduced through headphones [31], [32] via the HOA to Binaural Rendering module and paired to a Head Mounted Display (HMD) to allow real time adaptation of the sound field to head rotations in VR productions [33].

[5] In a Digital Audio Workstation, the “HOA mixer” is a multichannel signal bus containing $(N+1)^2$ signals where the output from each HOA panner/transform module are summed together.

Figure 6: An example of a monitoring system for the simultaneous production of SBA for VR and linear TV content. The workstation features a 7.1.4 loudspeaker system, an HMD for VR monitoring, and a DAW.

The HOA production workflow can successfully be applied both to live and episodic productions. Whilst the concepts described above hold for both production scenarios, the tools that the mixing engineer uses may be different. For example, in live productions, engineers may exploit HOA flexible rendering to easily switch between different loudspeaker layouts for monitoring, whereas in an episodic content scenario, mixing engineers would use a DAW and may find the spatial effects processor useful to easily align the sound field to the camera perspectives.

4.2 Transport stage: Contribution and Emission encoders

HOA is an ideal format for productions that involve large numbers of audio sources, such as episodic content, which typically involves many stems. If these stems were to be transmitted as individual OBA objects (i.e., audio information and related metadata), the bandwidth required to transmit the audio scene may be prohibitive. SBA is advantageous in this case as it limits the number of PCM channels transmitted to the end user to $(N+1)^2$ HOA signals.

However, increasing values of N (and, hence, increased performance of the HOA system), also lead to an increase of the bandwidth required to transmit the HOA PCM channels. This is particularly true for live broadcasting from production infrastructures based on Serial Digital Interfaces (SDI), which typically support up to 8 or 16 PCM channels (e.g. 8 PCM channels for SD SDI and 16 PCM channels for HD SDI).

To allow the transmission of HOA signals over interfaces of limited bandwidth (e.g. SD SDI), the HOA signals can be “spatially compressed” by means of appropriate

and advanced signal processing techniques [27] aimed at reducing the required bandwidth for the transmission while also maintaining high levels of spatial accuracy (if compared to the “uncompressed” HOA signals). Such signal processing techniques are part of MPEG-H Audio [5] and they are also supported by ETSI TS 103 589 [34].

Hereafter, we provide a brief description of how the HOA spatial compression technology works [5], [34].

By analysing the spatial characteristics of the HOA content, the HOA spatial compression algorithm reduces the number of PCM channels required for the HOA representation from $(N+1)^2$ to a set, M, PCM channels, called the HOA Transport Format [27]. This includes T audio channels called the “HOA transport channels” and one channel reserved for the HOA side info (metadata) [5], [34], so that $M=T+1$.

The compression algorithm is optimized for the given number of selected transport channels, M, so that the perceptual difference between the original (uncompressed) and the spatially compressed HOA is minimized.

The number, M, (transport channels + metadata) is typically much lower than $(N+1)^2$ and it is set by the content producer/broadcaster to suit the available bandwidth. The M PCM channels in output from the HOA spatial compression algorithm are next compressed using the MPEG-H Audio Core Codec to a chosen target bitrate (e.g. 384 kbit/s).

The performance of the spatial compression technology, which depends on the relation between the HOA order N and the chosen value for M, has been thoroughly assessed by means of listening tests [35]. A score of 80 MUSHRA points (broadcast quality) has been reported for bitrates as low as 300 kbit/s [27], [35].

Here are some examples in the choice of the value for M:

- For SD SDI interfaces (8 channel limit)
    - HOA content only: for example, an Nth order HOA content can be reduced to $M=8$ PCM channels thus enabling the transmission over an SD SDI bus.
    - HOA + one audio object: by selecting $M=7$ as the number of spatially compressed HOA signals, the total number of PCM channels over the SDI interface is $M+1=8$.
- For HD SDI interfaces (16 channel limit),
    - HOA content only: An N^{th} order HOA content (with $N>3$) can be reduced to $M=16$ PCM signals.
    - HOA content and 4 audio objects: the total number of PCM channels transmitted over HD SDI is $M+4+1=16$, where $M=11$ is the number of PCM

channels for the “HOA bed” (with no limitation on the HOA order), 4 is the number of PCM channels for the audio objects, and 1 channel (the control track) is reserved for metadata. However, lower values of M can also be selected if it is necessary to send further PCM channels (e.g. deliver a stereo mix in addition to the audio objects and the HOA bed).

Interfaces of limited bandwidth are not the only use case for HOA spatial compression. It can also be beneficial in applications where the HOA content is decoded and rendered on a mobile device. In this case, the HOA decoder on the mobile device would have to decode and render M signals only, as opposed to the $(N+1)^2$ PCM signals of the uncompressed HOA content.

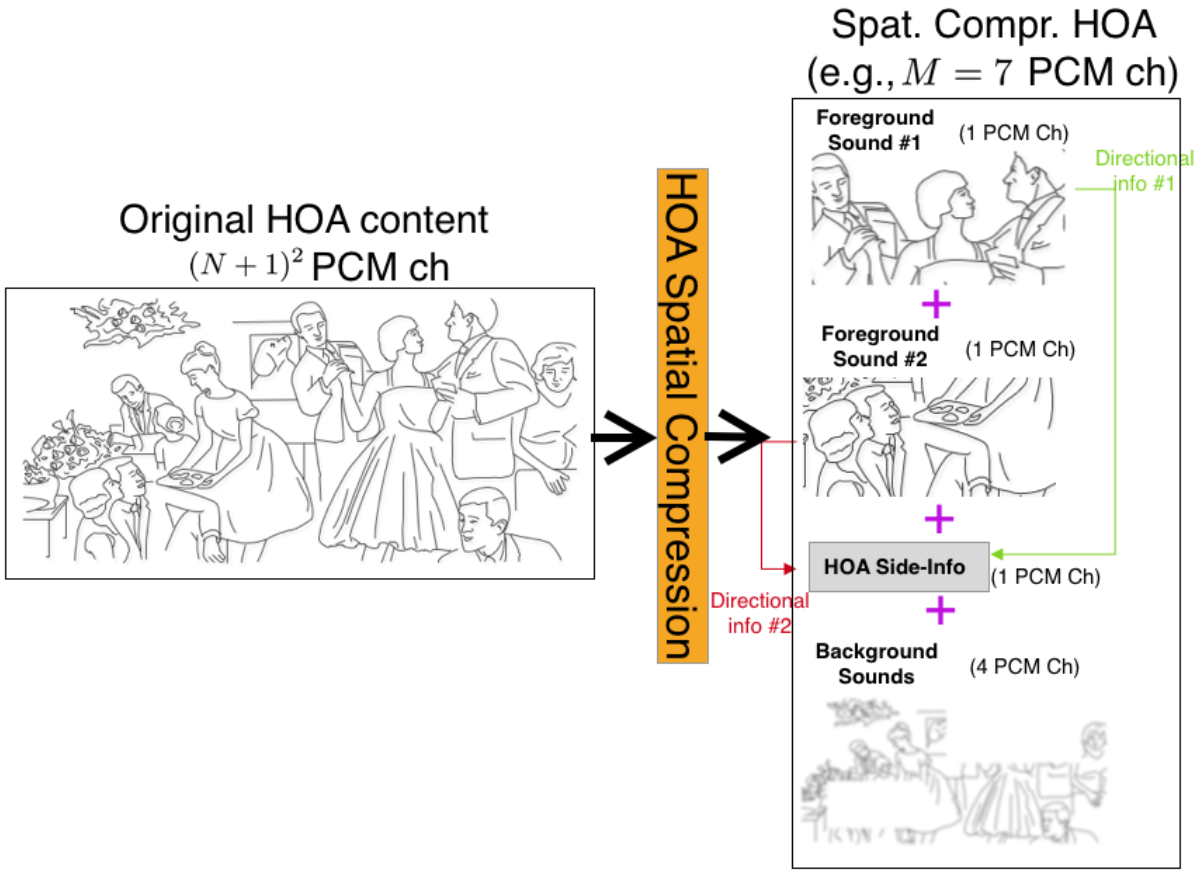

Figure 7: Graphical representation of the HOA spatial compression. The HOA sound scene is analysed and decomposed in foreground and background components and HOA side info.

From a technical standpoint, the HOA spatial compression algorithm uses mathematical and psychoacoustical principles to analyse and decompose the original HOA content in a set of T “foreground” and “background” components.

Figure 7 illustrates an example of a sound scene (represented in the HOA format) that consists of a party where people are having multiple conversations in an environment where ambient sounds are present (background music, indistinct chatter, etc.). The HOA spatial compression algorithm analyses the scene, and it

extracts the foreground components and their directional information, in this case the sounds (and the directions of the sound) produced by two groups of people having conversations. At the same time, it will extract the background component which contains the background music, indistinct conversations, etc.

Block diagrams of an HOA Spatial Encoder and Decoder are depicted in Figure 8

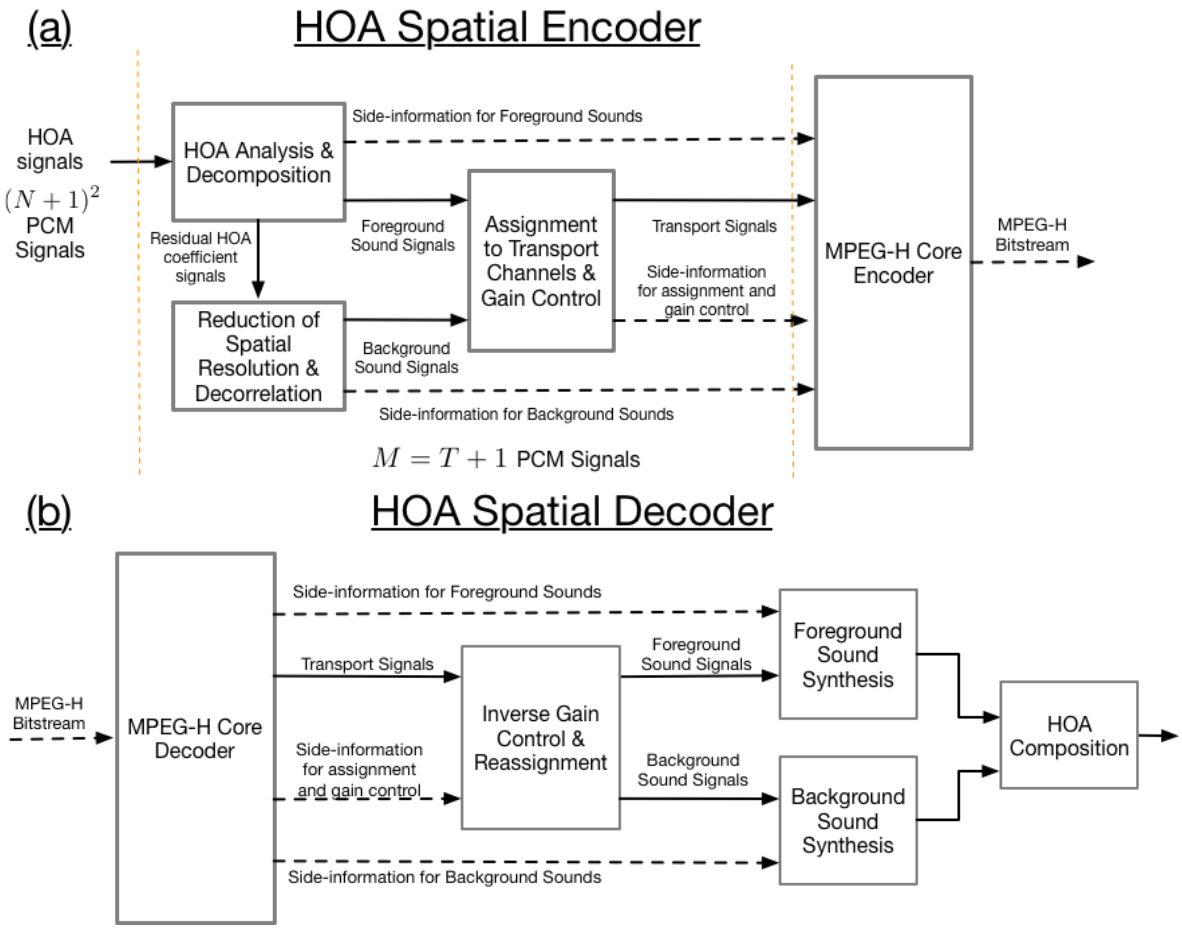

Figure 8: Block diagrams of a possible HOA Spatial Encoder (a) and the MPEG-H HOA Spatial Decoder (b) [5].

The input to the spatial encoder is the set of $(N+1)^2$ PCM signals containing the HOA coefficients obtained, for example, with the production workflow previously described. The HOA signals are analysed by the HOA spatial compression algorithms and subsequently decomposed in a set of predominant and ambient components (i.e. the HOA Transport Channels). These T transport channels, and the Side information is then compressed in the MPEG-H Audio Core Encoder at the selected target bitrate.

At the decoder side, the M signals are decoded in the MPEG-H Audio Core Decoder and the HOA coefficient signals are subsequently reconstructed using the transport channels and metadata. The $(N+1)^2$ HOA PCM signals are recovered at the decoder stage after the “HOA Composition” block.

In the context of a broadcasting workflow, the HOA Transport Format can be used for the transmission of HOA through a Contribution Encoder first and then through the Emission Encoder. With reference to Figure 9, the workflow is as follows:

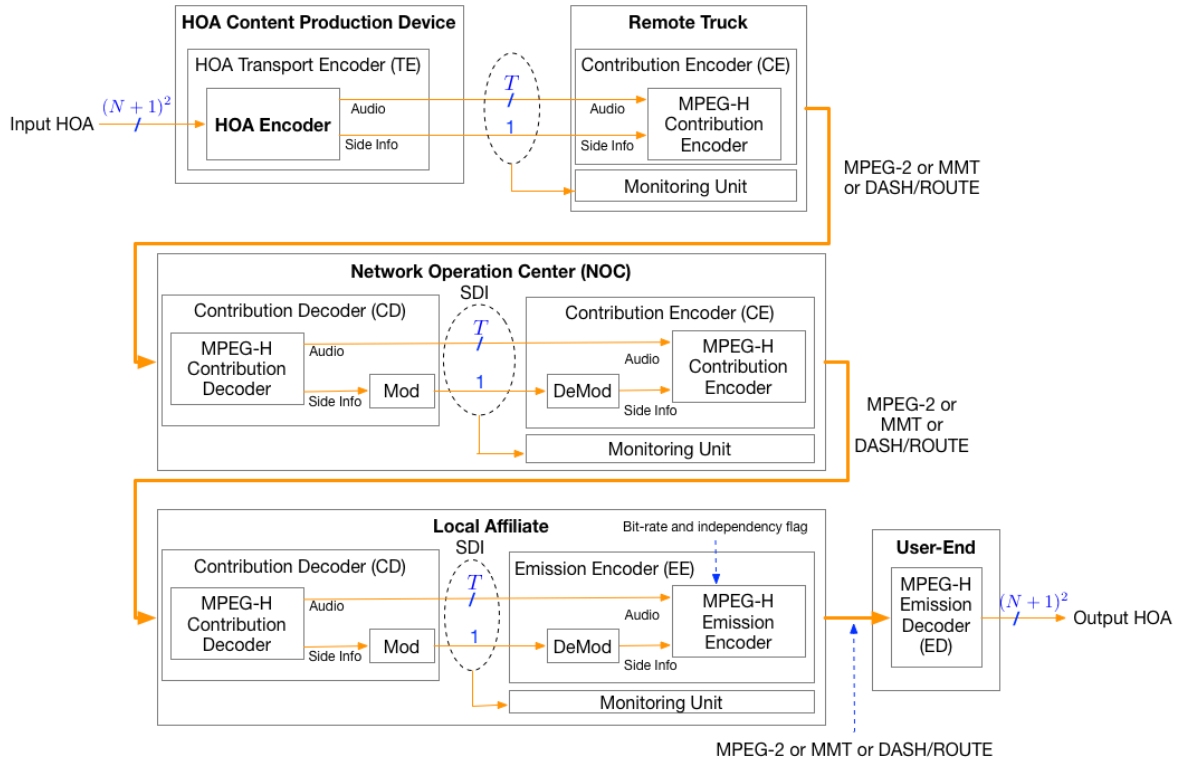

Figure 9: Workflow for the transport of HOA within existing broadcast facilities

The $(N+1)^2$ HOA PCM signals are firstly processed by the HOA Transport Encoder and compressed to M PCM signals (where $T \ll (N+1)^2$ and $T < 8$ if SDI is used, or $T < 16$ if HD SDI is used).

The set of T signals can now be transmitted to the Remote Truck, using MADI, for example. In the Remote Track, the HOA stream can be monitored and further compressed by the MPEG-H Contribution Encoder (CE).

Applying a transmission protocol, such as MPEG-2, MPEG Media Transport (MMT) or DASH/ROUTE, the audio stream is sent to a Network Operation Centre (NOC) and subsequently to a Local Affiliate where a combination of Contribution and Emission encoders/decoders are employed and a Monitoring Unit can inspect the HOA content before the actual transmission to the end user. In this final, transmission stage, the MPEG-H Audio Emission Encoder produces an MPEG-H Audio bitstream at a target bitrate (e.g. 384 kbit/s). The bitstream is transmitted to the user's end, where the MPEG-H Audio decoder reconstructs the $(N+1)^2$ HOA signals from the T transport signals.

The workflow in Figure 9 can easily be extended to accommodate audio objects alongside the HOA content.

4.3 Reproduction stage and interactive features

Once the HOA signals have been reconstructed by the MPEG-H Audio decoder, they can be rendered and further manipulated.

Some of the possible sound field manipulations that have already been mentioned include rotation and spatial filtering. The rotatability of HOA coefficients is particularly useful in the VR use case whilst spatial filtering can be employed to allow users to select parts of the audio scene and enhance sounds coming from their chosen directions.

As previously stated, HOA is a “loudspeaker agnostic” format, whereby the production and the reproduction stages are decoupled. Modern signal processing technologies extend the “loudspeaker agnostic” capabilities of HOA to “device agnostic”, meaning that the HOA signals can be rendered to any reproduction device including headphones, soundbars and mobile devices.

This is incredibly advantageous for content producers because they need only produce and deliver a single format (the HOA signals, with the addition of a few audio objects, if needed) that can be reproduced on any device.

5. A proposed workflow for NGA: a combination of SBA and OBA

Previously we have highlighted the advantages and limitations of using CBA, OBA, and SBA for NGA. In general,

- CBA is “fixed” – no flexible rendering and personalization
- OBA provides high personalization but increased complexity
- SBA provides flexible rendering but limited personalization

For a successful implementation of NGA, we propose a combination of SBA for the production and delivery of the immersive programme component (e.g. the M&E) alongside a few, selected audio objects (OBA) for the delivery of commentaries in different languages, dialogue and audio descriptions. In this way simultaneous flexibility and interaction is achieved to deliver compelling immersive and personalized NGA experiences to end users while limiting the complexity of the whole audio chain. This holds true in terms of production, bandwidth required for the delivery of the content and reduced complexity of the CE devices.

- **Production**: the production is performed in a single format (HOA) and it is no longer necessary to produce separate mixes for each target loudspeaker layout (e.g. 2.0, 5.1). This reduces production costs and efforts, also considering that the HOA production workflow is almost identical to that used for CBA. Sound engineers can monitor the HOA content over any loudspeaker layout while also being confident that their artistic intent is going to be preserved even if the end user also uses an irregular or non-standard loudspeaker layout. [22]

- The **rendering is flexible**, and it is based on the actual configuration used for reproduction. It is loudspeaker layout and device agnostic and it offers an open door to future, yet unknown reproduction environments – benefits in common with OBA. Even if the end user modifies their loudspeaker configuration (e.g. from 5.1 to 7.1.4), a new HOA rendering matrix is designed to faithfully reproduce the 3D audio content in HOA. All this is in contrast with CBA, where content is designed for a specific loudspeaker layout.

- **Rendering complexity** in SBA is independent of the scene complexity. In other words, the complexity of the SBA renderer does not increase as the number of objects increases (even if the audio scene is created from thousands of stems/mic signals). This is because the source signals are converted to a fixed number of HOA signals, uniquely dependent on the HOA order, and not on the number of objects present in the scene. This is in contrast with OBA, where rendering complexity increases as the number of objects increases.

- **Bandwidth costs** do not scale with scene complexity and they are easy to predict in SBA systems. The reason for this is the same as for the point above.

- **Interactive features** in SBA allow users to manipulate and customize the overall sound scene based on their preferences. This feature is also provided by OBA if the elements of the scene are provided as objects (and not as a channel bed, for instance). Instead, in SBA, users can select virtually any point in the sound scene and decide whether to e.g. amplify/silence sounds coming from any given direction.

- **Low-complexity manipulations** (e.g. rotation) of the sound scene in the SBA format is a direct consequence of the properties of the HOA format. This is particularly advantageous for VR applications where sensors on a Head Mounted Display (HMD) are used to track the user's head rotation and match the user's direction of view at any given time. It also facilitates 360 degrees video applications where the sound field rotation can typically be controlled by moving a finger on the touch screen of a mobile device.

- **Scalability**: the hierarchical structure of HOA (due to the orthogonality of each order) allows a trade-off between the sound field accuracy and computational complexity (e.g. to save battery power on a mobile device). For example, even if a high HOA order is used for the production, the reproduction device can render the audio content based only on the information contained in the lower orders.

6. Other use cases for SBA and HOA

The rise of VR and 360 degrees video content has seen HOA selected, de facto, as the industry standard in applications of this kind, with HOA being supported by numerous platforms for content creation and sharing[6].

In VR and 360 video applications the audio scene must be manipulated in real time (e.g. rotation) depending on the data captured by sensors which, in turn, capture users' input through devices such as an HMD or a touch screen. It is important to equip devices with low latency rendering processing to minimize the motion to sound delay and to hence provide users with a natural experience without discomfort. HOA provides a convenient signal processing framework to perform such operations in a low complexity and low latency fashion.

6.1 Virtual, Augmented, and Mixed Realities

Several broadcasters have started producing content for Virtual (VR), Augmented (AR), and Mixed Realities (MR) [36]. We'll loosely and succinctly refer to VR, AR, and MR as "new realities". In applications of this kind, the content is typically produced as if the listener were at the "centre" of the scene. The user is free to explore the scene (with different degrees of freedom) [33]. This listener centric paradigm is the same as that developed in the SBA framework. In fact, from a sound field point of view, we can imagine that the listener's head is immersed in an imaginary sphere on which the sound field is represented. This coincides with the underlying assumption of HOA sound field representation. In addition to that, where appropriate, audio objects can be added to a scene designed for these "new realities" and, hence, the SBA+OBA audio workflows described previously apply in these use cases too.

In these applications, listeners will typically use headphones for the reproduction of the 3D audio scene [33]. HOA content can efficiently be binauralized and this, together with the sound field manipulation capabilities, makes HOA the format of choice for applications of this kind [37]. The low computational complexity signal processing enables low latency head tracked binauralization for VR on consumer devices.

To summarize, HOA is conducive for "new realities" applications for various reasons:

- It provides a framework for the representation of sound fields over a sphere,
- It provides low complexity and low latency sound field manipulations (e.g. rotation) efficient enough to be implemented on battery powered, mobile devices.

[6] 1) Facebook 360 Spatial Workstation https://facebookincubator.github.io/facebook 360 spatial workstation/KB.html
2) Youtube Spatial Audio https://support.google.com/youtube/answer/6395969
3)Google Resonance https://developers.google.com/resonance audio/

In new realities applications, the user may be allowed to freely “move” within a virtual scene, a use case which is commonly referred to as “six degrees of freedom” (6DoF). While research in this field is currently ongoing [39], SBA and OBA can be used to address 6DoF scenarios to provide both smooth rotation and realistic translation of virtual sound fields.

6.2 360° videos

A report from the EBU [36] suggests that the 360° video format is being experimented with by many broadcasters [38]. In 360° video applications, the user can “swipe” on the control surface where the video is playing (e.g. a tablet or a phone) and explore the scene. This implies that the sound field needs to be continuously “adapted”, based on interaction with the user. If the audio content for the 360° video is produced with a combination of HOA and objects, the adaptation of the sound field can be performed in real time with relatively low computational complexity, thus enabling users to consume the 360° video content on battery powered devices.

6.3 Simultaneous NGA TV, VR, and 360 videos production and delivery

A use case that is of interest to broadcasters is the production of 3D audio content for the simultaneous delivery of NGA content over TV, VR, and 360° video. In this use case, the video would be captured using a linear TV camera and a VR (or 360°) camera. The audio can be produced in HOA and audio objects using a single audio workflow to produce the 3D audio scene.

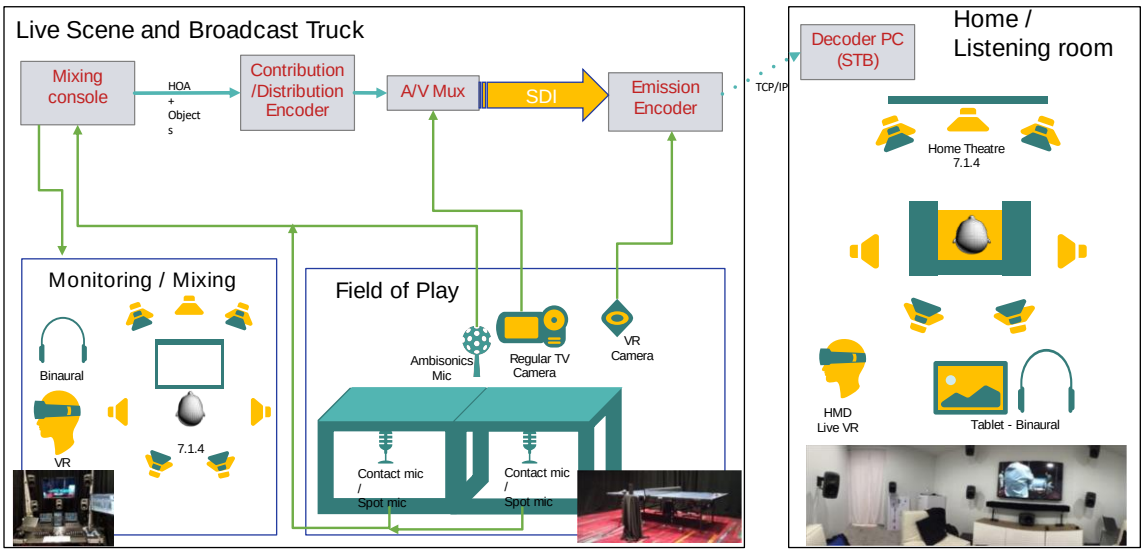

Figure 10: Audio Workflow for simultaneous NGA, VR, and 360 video production

At the user’s end, appropriate HOA and OBA renderers in each device allow the reproduction of the 3D audio scene on HMDs (via binaural rendering on headphones), loudspeaker systems, soundbars, tablets, and mobile phones. The single audio workflow helps broadcasters to control the cost of the production while

ensuring that consumers using different devices will enjoy a consistent and accurate 3D audio experience independent of the device used for the reproduction. An example implementation of such a scenario for a table tennis event is shown in Figure 10.

7. Conclusions

Based on the Higher Order Ambisonics format, Scene Based Audio is a portfolio of technologies for the production, delivery, and reproduction of 3D audio. The HOA format allows the representation of arbitrarily complex sound scenes using a limited number of PCM audio channels (i.e. the HOA signals), which can be further processed for reproduction on any device (e.g. loudspeakers, soundbars, headphones, etc.).

The HOA signals, whose number depends on the order of the HOA representation, can be further reduced to a lower number (e.g. 8 PCM signals). This compression technique allows HOA to be transported within existing broadcast infrastructures and subsequently it allows the efficient delivery of arbitrarily complex 3D audio scenes and their high-fidelity reproduction to many consumer devices. In this article we have discussed how the cost effective and low complexity HOA signal processing scheme can enable broadcasters to successfully produce and deliver immersive NGA content using a combination of audio objects and HOA for linear TV, VR, and 360° video.

8. References

[1] Digital Video Broadcasting, “Specification for the use of Video and Audio Coding in Broadcasting Applications based on the MPEG-2 Transport Stream (ETSI TS 101 154, v2.3.1),” v2.3.1, Feb. 2017.

[2] Roger Miles, “Part 1: Vertically integrated audio platforms and the need for a standard audio renderer,” *EBU Tech-i “Meeting audience expectations,”* no. 29, p. p9, Sep-2016.

[3] David Wood, “Taking A Sounding About UHDTV,” *EBU Tech-i “The Challenges of Next Generation Audio,”* no. 30, p. 1, Dec-2016.

[4] R. L. Bleidt *et al.*, “Development of the MPEG-H TV Audio System for ATSC 3.0,” *IEEE Transactions on Broadcasting*, vol. 63, no. 1, pp. 202–236, 2017.

[5] ISO/IEC, “Information Technology - High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D audio,” 2015.

[6] EBU, “ADM Renderer for use in Next Generation Audio Broadcasting,” EBU, Geneva, Switzerland, TECH 3388, Mar. 2018.

[7] ITU, “Recommendation ITU-R BS.2076-1 - Audio Definition Model,” International Telecommunication Union, Geneva, Switzerland, 2017.

[8] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, “MPEG-H 3D Audio—The New Standard for Coding of Immersive Spatial Audio,” *IEEE J. Sel. Top. Signal Process.*, vol. 9, no. 5, pp. 770–779, Aug. 2015.

[9] Robert Bleidt, Arne Borsum, Harald Fuchs, and S. Merrill Weiss, “Object-Based Audio: Opportunities for Improved Listening Experience and Increased Listener Involvement,” *SMPTE Motion Imaging Journal*, vol. 124, no. 5, p. 13, 2015.

[10] Nils Peters, Deep Sen, Moo-Young Kim, Oliver Wuebbolt, and S. Merrill Weiss, “Scene-based Audio Implemented with Higher Order Ambisonics (HOA),” presented at the SMPTE 2015 Annual Technical Conference & Exhibition, 2015.

[11] M. A. Gerzon, “Periphony: with-height sound reproduction,” *Journal of the Audio Engineering Society*, vol. 21, no. 1, 1973.

[12] Jérôme Daniel, “Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia,” 2001.

[13] M. Poletti, “Three-dimensional surround sound systems based on spherical harmonics,” *Journal of the Audio Engineering Society*, vol. 53, no. 11, 2005.

[14] M. Kronlachner and F. Zotter, “Spatial transformations for the enhancement of Ambisonic recordings,” p. 5.

[15] D. G. Malham, “Higher Order Ambisonic systems for the spatialisation of sound,” presented at the 1999 International Computer Music Conference (ICMC), Beijing, China, 1999.

[16] H. Pomberger and F. Zotter, “An Ambisonics Format for Flexible Playback Layouts,” Graz, Austria, 2009.

[17] Jörn Nettingsmeier, “Higher-order Ambisonics - A future-proof 3D audio technique,” presented at the Verband Deutscher Tonmeister International Convention, 2010.

[18] Jérôme Daniel, “Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia,” Université Paris 6, 2001.

[19] D. Malham, “Higher order Ambisonic systems,” pp. 1–12, Dec. 2004.

[20] Christian Nachbar, Franz Zotter, Etienne Deleflie, and Alois Sontacchi, “Ambix-a suggested ambisonics format,” in *Ambisonics Symposium 2011*, Lexington, KY, USA, 2011.

[21] T. Carpentier, “Normalization schemes in Ambisonic: does it matter?” presented at the Audio Engineering Society 142nd Convention, Berlin, Germany, 2017.

[22] Sébastien Moreau, Jérôme Daniel, and Stéphanie Bertet, “3D sound field recording with Higher Order Ambisonics - Objective measurements and validation of spherical microphone,” in *Audio Engineering Society Convention 120*, Paris, France, 2006.

[23] Stéphanie Bertet, Jérôme Daniel, Etienne Parizet, and Olivier Warusfel, “Investigation on Localisation Accuracy for First and Higher Order Ambisonics Reproduced Sound Sources,” *Acta Acustica united with Acustica*, vol. 99, no. 4, pp. 642–657, Jul. 2013.

[24] M. Frank and F. Zotter, “Exploring the Perceptual Sweet Area in Ambisonics,” May 2017.

[25] Michael A. Gerzon, “Ambisonics in Multichannel Broadcasting and Video,” *Journal of the Audio Engineering Society*, vol. 33, no. 11, 1985.

[26] M. A. Gerzon and G. J. Barton, “Ambisonic Surround-sound mixing for multitrack studios,” presented at the Audio Engineering Society 2nd International Conference, Anaheim, CA, USA, 1984.

[27] Deep Sen, Nils Peters, Moo-Young Kim, and Martin Morrell, “Efficient Compression and Transportation of Scene Based Audio for Television Broadcast,” presented at the Audio Engineering Society International Conference on Sound Field Control, Guildford, UK, 2016.

[28] N. Peters, M. Morrell, and D. Sen, “Sports video production with scene-based audio for MPEG-H,” New York City, USA, 14-Mar-2016.

[29] M. Frank, F. Zotter, and A. Sontacchi, “Producing 3D Audio in Ambisonics,” presented at the Audio Engineering Society 57th International Convention, Hollywood, CA, USA, 2015.

[30] Qualcomm Technologies, Inc., “3D Audio Tools.” [Online]. Available: https://developer.qualcomm.com/software/3d-audio-tools.

[31] M. Noisternig, T. Musil, A. Sontacchi, and R. Holdrich, “3D binaural sound reproduction using a virtual ambisonic approach,” 2003.

[32] D. Menzies, “Nearfield Synthesis of Complex Sources with High-Order Ambisonics, and Binaural Rendering,” in *th International Conference on Auditory Display*, 2007.

[33] Shankar Shivappa, Martin Morrell, Deep Sen, Nils Peters, and S. M. Akramus Salehin, “Efficient, compelling and immersive VR audio experience using Scene Based Audio / Higher Order Ambisonics,” presented at the Audio Engineering Society Conference on Audio for Virtual and Augmented Reality, Los Angeles, CA, USA, 2016.

[34] ETSI, “Higher Order Ambisonics (HOA) Transport Format,” Sophia Antipolis Cedex, France, ETSI TS 103 589, Mar. 2018.

[35] International Organisation for Standardisation, “MPEG-H 3D Audio Verification Test Report (w16584) - The Moving Picture Experts Group - Audio Subgroup.” [Online]. Available: https://mpeg.chiariglione.org/standards/mpeg h/3d audio/mpeg h 3d audio verification test report h.

[36] Paola Sunna and Graham Thomas, “How are EBU PSM members using 360/VR?,” Dec-2017.

[37] Markus Noisternig, Alois Sontacchi, Thomas Musil, and Robert Höldrich, “A 3D Ambisonic Based Binaural Sound Reproduction System,” presented at the Audio Engineering Society 24th International Conference on Multichannel Audio, Alberta, Canada, 2003.

[38] EBU, “Opportunities and challenges for Public Service Media in VR, AR and MR,” European Broadcasting Unit (EBU), Geneva, Switzerland, TR 039, Apr. 2017.

[39] The Moving Picture Expert Group, MPEG-I Part 4: Coded Representation of Immersive Audio, https://mpeg.chiariglione.org/standards/mpeg-i, Last visited: July 2019.

9. Author(s) biographies

	Ferdinando Olivieri received his MSc in Telecommunications Engineering from the University of Florence (Italy), his MSc in Acoustics from the Institute of Sound and Vibration Research (ISVR), University of Southampton (UK), and his PhD in Audio Signal Processing from ISVR. He is currently working as an Audio Research Engineer in Multimedia Advanced Tech R&D, Qualcomm Technologies, Inc, San Diego, CA. His research interests include array signal processing, inverse problems in acoustics, beamforming, and sound field capture and reproduction. Dr Olivieri is a member of the Audio Engineering Society and of the IEEE Signal Processing Society.
	Nils Peters received his MSc. in electrical and audio engineering from the University of Technology, Graz, Austria, and his PhD. in music technology from McGill University, Montreal. He was a Post-Doctoral Fellow with the International Computer Science Institute, Centre for New Music and Audio Technologies, and the Parallel Computing Laboratory, University of California, Berkeley. He is currently a Senior Staff Research Engineer and a Manager in the Multimedia Research and Development Laboratory, Qualcomm Technologies, Inc. He has researched in the field of spatial audio for over a decade and currently on technical and perceptual aspects of spatial audio, including soundfield analysis and compression, acoustic sensing, spatial sound reproduction, auditory perception, and sound quality evaluation. Dr Peters participates in audio standard developments at MPEG, 3GPP, DVB, and ITU-R and serves as the Co-Chair of the Technical Committee for spatial audio at the Audio Engineering Society.
	Deep Sen received his BE. and PhD. degrees from the School of EE&T at the University of New South Wales, Sydney, Australia, in 1990 and 1994, respectively. From 1994 to 2003, he was with the Speech and Audio Laboratories, AT&T Bell Laboratories, NJ, USA, and from 2003 to 2011, he was a Faculty Member with the School of Electrical Engineering, University of New South Wales, Australia. From 2011 to 2018, Deep was a Senior Director with the Multimedia R&D group at Qualcomm Technologies, Inc., San Diego, CA, USA, when he co-authored this article. Since 2018, he has been a Distinguished Scientist/Architect with Apple, Inc., Cupertino, CA, USA. He has multiple peer-reviewed publications and granted patents. His experience includes speech and audio coding, perception, bio-mechanical modelling of the cochlea, blind source separation, beamforming, and hearing prosthetics. Dr Sen is a member of the Acoustical Society of America, Audio Engineering Society, and has served as an elected member of the IEEE Speech and Language Technical Committee.

Published by the European Broadcasting Union, Geneva, Switzerland

ISSN: 1609-1469

Editor-in-Chief: Patrick Wauthier

E-mail: wauthier@ebu.ch