Audio Engineering Society

Conference Paper

Presented at the Conference on
Audio for Virtual and Augmented Reality
2016 September 30–October 1, Los Angeles, CA, USA



# Efficient, compelling and immersive VR audio experience using Scene Based Audio / Higher Order Ambisonics

Shankar Shivappa, Martin Morrell, Deep Sen, Nils Peters and S. M. Akramus Salehin

*Qualcomm Technologies Inc. (QTI), San Diego, California, USA*

Correspondence should be addressed to Shankar Shivappa (sshivapp@qti.qualcomm.com)

**ABSTRACT**

For a fully immersive and compelling VR experience, the acoustic-illusion of being 'present' in the virtual world must be created. To achieve this illusion two aspects are compulsory: (1) authentic spatial audio production and (2) the need to track and adapt the audio scene to the listener's head position and orientation. This paper shows how Scene-based audio (SBA), often synonymous with Higher Order Ambisonics (HOA), is ideal for VR because its ease of acoustic capture, offline content creation, post-production, transmission and interactive rendering. Compared to object-based audio, the rendering complexity is much lower for SBA. Also, SBA can offer higher and more coherent spatial fidelity when compared to channel-based audio. One of the advantages of SBA is flexible rendering, which means that the same audio stream can be rendered to various speaker formats including binaural rendering for headphone consumption. The paper discusses the need for efficient SBA compression for VR content delivery, and presents MPEG-H as an efficient and versatile delivery system for SBA. For a personalized VR experience, accurate binaural rendering is essential. SBA can be efficiently binauralized. Its number of convolutions is proportional to the number of HOA coefficients, rather than proportional to the number of virtual loudspeakers. This means that SBA can render to a high number of virtual loudspeakers without impacting the binauralization computation cost. Furthermore, to improve the spatial perception, SBA binauralization can utilize grids of ideally positioned virtual loudspeakers based on platonic solids or otherwise regularly spaced loudspeaker configurations that are impractical in reality and unsupported in channel-based audio formats.

Interactive soundfield rotation in real time is indispensable for creating VR experience. We show how SBA can be rotated and even further enhanced with other user-controlled effects, such as zooming. The paper will discuss use cases to demonstrate the capture, processing, and playback of SBA and will show potential pitfalls and design strategies for an end-to-end spatial audio system for VR. The authors will then conclude that SBA is a

robust and compelling audio format for VR, and that SBA can be easily distributed via broadcast or OTT for real-time end consumer use.

## 1 Introduction

For creating compelling immersive VR experiences three important technical components have to work together: One is the visual quality of the video reproduction, the second is the intuitive user interaction with the VR world, and finally the fidelity and authenticity of the spatial sound reproduction.

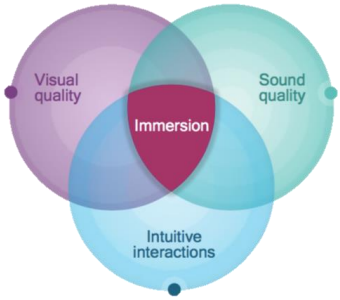


Fig 1. The three pillars of a VR experience.

VR has many use cases and applications. VR finds application in live broadcasting events where audio is captured, transmitted and played back by remote users. In cinematic VR the audio is post produced with stems for later playback. Immersive computer games is another driving factor for VR technology. There are several other use cases such as education and training, marketing, medicine and forensics where VR is being used. In all the VR use cases, having a realistic audio experience, in addition to the visuals, creates and enhances the sense of "presence" or "immersion" in the virtual scene.

Scene Based Audio refers to any format that endeavours to capture and/or define the acoustic pressure field as a function of time (i.e. the acoustic scene). [1] Scene Based Audio (SBA) encompasses audio capture to first order ambisonics, higher order ambisonics or any format that is sensor or loudspeaker location and characteristic agnostic by mapping to coefficients of bases. SBA is particularly advantageous for VR since the audio can be easily rotated to compensate for user head rotations and allows capturing, production, and reproduction equally well in all directions. Channel based audio is hard to adapt to changing viewing directions requiring a large number of HRTFs for all virtual loudspeaker positions. Rotation of channel based audio is sensitive and results in artefacts especially for fast head movements. Objects for VR audio can be complex needing recording and tagging of many objects making transmission complex with a large bandwidth requirement for complex scenes.

On the other hand, SBA based on spherical harmonics allows smooth soundfield rotations and the complexity of the reproduction to headphones or loudspeakers as well as the compression being independent of the scene complexity. Furthermore, SBA can be captured easily with compact microphone arrays and spot microphones as well as being augmented with audio stems.

For VR games, SBA overlaid with objects provides the most advantages since SBA is highly efficient for creating and capturing ambience as well as size and diffusion properties of sound scenes. The interactive components in VR games are set as audio objects which are overlaid onto HOA scenes.

An example of a complete SBA based system is shown in Fig. 2. The soundfield is captured by microphone arrays and spot microphones and then converted to HOA. These HOA signals are then compressed and transmitted. At the receiving side the HOA stream is decompressed and rendered over headphones for VR play back.

The organization of this paper is as follows: Section 2 presents the requirements for a VR audio format and Section 3 compares SBA against channels and objects for audio in VR. Section 4 describes VR use cases which include cinematic VR, VR games and VR for live events such as sports and concerts. Section 5 describes the available SBA based microphone array devices and the conversion to higher order ambisonics audio format. Post production enhancements of captured SBA and processing of multi-microphone capture using our SBA plug-in suite is described in Section 6. Section 7 describes the process of compressing SBA and Section 8 presents methods to render SBA to loudspeakers and headphones. Subjective testing to evaluate performance of SBA is reviewed in Section 9 and Section 10 provides a summary of the key points presented in this paper.

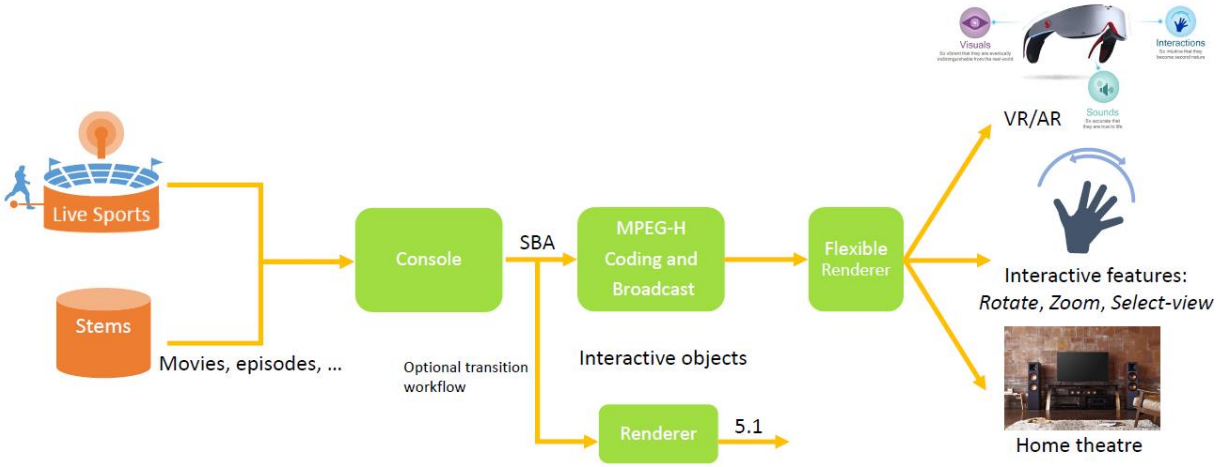


Fig 2. Complete SBA based flow chain starting from audio capture to compression and then to rendering.

## 2 Requirements for VR Audio

In addition to general audio/video quality requirements such as AV sync, Lip-sync, and natural timbral sound quality, for compelling immersive VR experiences using headphones, the audio system has to fulfil the following requirements:

- Accurate localization of sounds in all directions.
- Dynamic binauralization of the soundfield (headtracking).
  - Accurate high resolution soundfield rotation matching human perception up to 1° resolution [2].
  - Unperceivable motion-to-sound latency [3].

One of the essential novelties of VR is that a user is free to change their viewing gaze at will, allowing individual immersive experiences in any viewing direction at any given moment. Consequently, the methods of audio creation, transmission and reproduction applied for VR must be able to accompany the dynamically changing visual perspective.

This means that audio must be reproduced equally well in all directions, allowing the presentation of sounds from below or above the viewer with the same spatial accuracy than sounds from the front.

Further, for a realistic listening experience the audio presentation method must seamlessly adapt the spatial audio processing and recreate the sound scene coherently with respect to the dynamically changing viewer's gaze.

Although VR audio is not limited to headphones, we primarily consider VR headphone reproduction in this paper.

The common binauralization method for VR audio is to use impulse response measurements of loudspeakers at given angles for both the left and right ear. These are commonly known as HRTFs (Head Related Transfer Function), HRIRs (Head Related Impulse Response) or BRIRs (Binaural Room Impulse Response). The first two defines how a person in a non-reverberant environment receives a (possibly from a loudspeaker) sound from a specific direction and distance. In contrast, the BRIRs also captures the acoustic effects of a reverberant room. Therefore, BRIRs are often used to capture and simulate an intended listening room.

For creating a two-channel binaural headphone version of an immersive audio representation, sets of HRTFs/BRIRs are used for filtering loudspeaker signals with the appropriate impulse response. Consequently, this requires that the HRTF/BRIR for the specific loudspeaker location exist. The use of the BRIR of the local room (where the headphone audio is being consumed) also helps to improve externalization, immersiveness and localization of sounds. For VR experiences, where the intention is to exclude all local cues and only allow cues that are relevant to the virtual world, this use of local room's response may not be required.

## 3 Comparison with Other Formats – Channel, Object and FOA

In this section, we compare and contrast scene-based audio with other spatial audio formats. Pros and cons of each format are discussed in the context of VR.

### 3.1 Channels

Loudspeaker-based audio reproduction such as stereo 2.0 or surround 5.1 has been the de-facto standard for production and audio delivery to consumers for decades. To ensure the intended sound reproduction, channel-based audio requires the same standardized loudspeaker placement at the production facility and the listener's reproduction location. Standardized loudspeaker configurations include simple mono and stereo, horizontal-only (5.1) to immersive 7.1+4 and 22.2 [22][23]. Most of these configurations are primarily tailored toward cinematic AV experiences using a non-uniform and non-isotropic loudspeaker placement which prioritizes the frontal sound stage area for accompanying the events on the video screen. Other areas (such as the floor area) are either not covered or only sparsely covered by loudspeakers, which decreases the ability to accurately reproduce sounds from these directions.

Further, channel-based audio is hard to adapt to a changing viewing direction: for a faithful reproduction the entire immersive virtual loudspeaker configuration must be virtually displaced by updating the HRTFs associated to the new direction of each speaker. This requires an accurate set of HRTFs available for all possible virtual loudspeaker positions as well as careful signal processing when updating the HRTFs in real time. This update process is sensitive to artefacts such as sound coloration especially for fast head movements [20] and if not done correctly will destroy the illusion completely. It was shown that immersive channel-based content can be faithfully reproduced over headphones using dynamic head-tracking with the same (relative) immersive quality compared to speaker reproduction [Ref: FhG-DAGA2016]. However, there is no standardized speaker format that delivers auditory cues in all directions around the listener. Therefore, channel-based audio is not the ideal immersive audio format for the VR use case.

### 3.2 Objects

The concept of audio objects, possibly first introduced with MPEG-4 AudioBiff [24], gained commercial attention for cinematic content with Dolby Atmos. The basic concept is that an audio scene is assembled at the consumer side from individually transmitted sound sources (objects) and their metadata describing the intended object position and other spatial properties. Based on these metadata, the audio scene is constructed using an audio object rendering algorithm.

The capturing of audio objects for VR can be a challenging task because all sound sources must be individually captured and tagged with the correct location metadata. Leakage across the captured audio objects (e.g., due to proximity or reverberation) needs to be avoided because it can affect the sound localization and timbre of the rendered audio objects.

A pure object-based representation requires the use of individual audio tracks. This means that the bandwidth necessary for transmitting a sound scene depends on the number of simultaneous objects present at any point in time – or the scene complexity. Typical cinematic content requires the use of hundreds of simultaneous objects and their metadata - meaning that the bandwidth requirements of object-based audio is prohibitively high for streaming or broadcast. Solutions involving coalescing multiple objects and or using a channel bed reduces the original attraction of object based audio – that of high spatial resolution and flexible rendering.

Objects can be either individually binauralized using one discrete HRTF convolution process per object, or rendered (e.g., using VBAP [4]) to a set of virtual loudspeakers which is then binauralized using one HRTF convolution process per virtual loudspeaker.

### 3.3 Scene-based Audio

Using the concept of spherical harmonic basis functions, scene-based audio describes how the sound pressure in a scene changes as a function of time and direction. In contrast to channel-based audio, scene-based audio is rendered to loudspeakers at the consumer side and thus can cater for all standardized as well as non-standardized reproduction setups.

First Order Ambisonics (FOA) (also commonly referred to as B format [28]) is a basic form of scene-based audio in which the soundfield is described by only the lowest four spherical harmonic coefficients. Higher order ambisonics (HOA) provides a concise and accurate representation of the entire sound field and overcomes the spatial resolution constraints of FOA. Also, the listening area in which the reproduced soundfield can best be perceived is larger for higher orders than FOA [25]. SBA provides an efficient and accurate representation of the sound field with a limited number of coefficients, dependent only on the order of the representation that is chosen. Moreover, the SBA coefficients can be compressed to a fixed bandwidth irrespective of the complexity of the scene [26].

The spherical harmonic underpinnings of SBA allows for efficient and smooth rotation of the sound field. There are algorithms (described in section 8) for efficient binauralization of HOA coefficients that are also independent of the complexity of the scene

and the number of virtual speakers used in rendering. These computational advantages are invaluable in enabling head-tracked binauralization for VR on consumer devices. By selecting a high enough order and a large number of virtual speakers, SBA can provide very high spatial resolution and fidelity of the sound scene without incurring the complexity and bandwidth limitations of objects and channels. Yet another advantage of SBA is that there is a practical way to capture the sound field using compact microphone arrays and also the flexibility to author and augment the scene from individual stems and spot microphones, as presented in Sections 5 and 6 respectively.
The advantages of SBA in the context of VR are made evident by Google and Facebook announcing support for first order ambisonics and they might support higher orders in the future.

## 4 SBA Relevance in Different VR Use Cases

VR encompasses a wide range of experiences and a few representative ones are described below along with a discussion of why SBA is extremely relevant in these use cases.

### 4.1 Infinite Seat

Sports, concerts and other live events can be captured, streamed and virtually experienced by remote users, leading to this concept of the "infinite seat". The live capture capability of SBA using compact microphone arrays makes it a suitable candidate for this use case. The movement of the audio sources in the scene, the reverberation effects and the ambience such as applause from the audience are all conveniently captured by a relatively small microphone array. There is no need to separately track the audio sources, which can be impractical in a sports scene, for example. In addition, the sound field captured from one location in the live event can be augmented in real-time with additional audio features such as mixing audio from additional spot-mics, commentary tracks and warping the sound field to create artistic effects, without increasing the bandwidth requirements from that of the baseline sound field.
The compact microphone arrays that can capture the sound fields in a live scene are especially conducive in enabling related use cases such as news reports from disaster areas, where a realistic capture of the audio scene is invaluable. Imagine a weather reporter with a small spherical microphone array, standing near a shoreline while a storm is approaching – such a setup, using SBA, can create an experience of viewer being immersed in the scene, with the sound of the waves crashing and the reporter's commentary all coming together to create a compelling VR experience.

### 4.2 Cinematic VR

Cinematic VR refers to a broad class of VR experiences where the underlying assumption is that of post-produced episodic content. The ability of the artist to create a compelling audio experience by mixing pre-recorded stems is vital. Composing a scene using audio objects could result in a large bandwidth requirement for transmission and storage due to the number of objects in a realistic scene. In some cases such as documentaries that have real world scenes in them, it could be difficult to capture isolated audio objects and their location information. SBA provides a practical fixed-bandwidth solution to content creators by being able to represent very complex audio scenes in a concise manner, irrespective of the scene complexity. The ability to capture SBA using compact microphone arrays makes it possible to easily deploy this technology in the field. The SBA production tools described in the following sections will provide the content creators with a suite of tools that are comparable to tools available in other formats. In addition, the creation of a SBA mix will allow the VR experience to be consistent over a wide range of rendering options from binaural to immersive surround sound systems.

### 4.3 VR Games

Object-based audio has traditionally been the format of choice in gaming engines. Many of the sounds in games are generated interactively, in response to user actions and audio objects are a convenient way to represent them. However, to create realistic sounding experiences, a gaming engine has to not only recreate the location of the audio object but also other attributes of the object such as its size and diffusion properties. This leads to higher computational demands on the consumer-end devices. SBA can offer an alternative approach to creating realistic audio scenes within a game by allowing the game designer to compose complex audio scenes with fixed complexity and bandwidth requirements. A few objects can be overlaid on top of the HOA scene if necessary. SBA can also be an alternative rendering option for a scene composed of many objects.
In the above illustrative use-cases, one observes that SBA has advantages at several stages of the VR workflow. In the next few sections, the application of SBA to different stages of the VR ecosystem from production to transmission and consumption are discussed.

## 5 SBA Acquisition

One of the key advantages of SBA is that compact microphone arrays can be used to capture the sound field. In VR, audio localization needs to match up with the visual cues to provide a complete immersive experience. There has been significant development in video capture for VR with two fisheye lens capture devices such as the theta360 [9] as well as more expensive 360 degree video capture consisting of camera arrays [10]. The video capture needs to be complemented with a 3D audio capture system. There are several consumer and professional grade audio capture devices that are geared toward FOA and HOA on the market, propelled by the need for scene based audio in VR.

Sennheiser created a four microphone array, AMBEO, placed at the vertices of a tetrahedron specifically for recording audio content for VR [11]. AMBEO is a first order ambisonic microphone outputting A format audio. The ambisonics A format audio is the raw microphone capsule signals for a first order ambisonic microphone array. The A format is converted to B format which is the first order ambisonic audio consisting of W, X, Y and Z coefficients. The W channels contains audio from all directions, X, Y and Z channels correspond to sound directions along x, y and z axis, respectively. The B format audio can be easily rotated for VR and is integrated into YouTube and Oculus cinema for playback. The Core Sound TetraMic [12] and Soundfield MKV are other first order ambisonic microphone arrays based on the tetrahedral arrangement. Horizontal only B format comprising of only W, X and Y channels without any height Z channel is supported by Google's JUMP system for VR. This uses the portable Zoom H2N audio recorder which has five microphones on the horizontal plane.

Higher order ambisonic recording devices overcome the low spatial resolution of the first order systems mentioned previously. The MH Acoustics Eigenmike [13] is a spherical microphone array on a rigid sphere of radius 4.2 cm. It comes with an Eigenstudio application for recording raw signals as well as Eigenunits VST plug-ins for conversion to third order ambisonics. The larger number of microphones on the Eigenmike means that the raw signals can be converted to fourth order ambisonics. VisiSonics has their own higher order audio capture device. VisiSonics has an integrated higher order ambisonic microphone array and a camera array on a 20 cm rigid sphere [14]. Their system includes 64 microphones and 5 video cameras. The larger number of microphones means that the captured soundfield can be decomposed to 7th order ambisonics providing a really high spatial resolution of up to two degrees. Figure 1 shows a few of the ambisonic audio devices available on the market.

Fig. 2 Various ambisonic audio capture devices available on the market. (a) Core Sound TetraMic, (b) Zoom H2n and (c) MH Acoustics Eigenmike.

The conversion from raw microphone signals to higher order ambisonics involves weighted summing of the raw signals together with convolutions. The advantage of doing this conversion is that the HOA signal does not need to contain any information of the locations or characteristics (e.g. directionality or frequency response) of the microphones to decipher the directions of audio. Furthermore, recordings from single microphones can be added to these HOA signals provided we know which directions these recorded signals should impinge from in the final mix or they can be added based on the locations of the microphones. Audio capture to HOA is not limited to spherical microphone arrays but can be done using several single microphone recordings. These can be mixed to a single HOA audio signal based on the microphone locations or offline.

## 6 SBA Production

Production for SBA can take one of two approaches; a spherical microphone array or multi-microphone capture like traditional channel-based capture. The previous section discussed in-depth capture and processing from a spherical microphone array. This section will discuss processing of multi-microphone capture and post-processing of an SBA signal for creative enhancement.

### 6.1 Plug-In Suite

Our plug-in suite allows for the creation and monitoring of SBA format content within a digital audio workstation (DAW). The suite is able to create $6^{th}$ order content and output to any number of loudspeaker configurations by employing an internal SBA buss structure between input, fx and output type plug-ins.

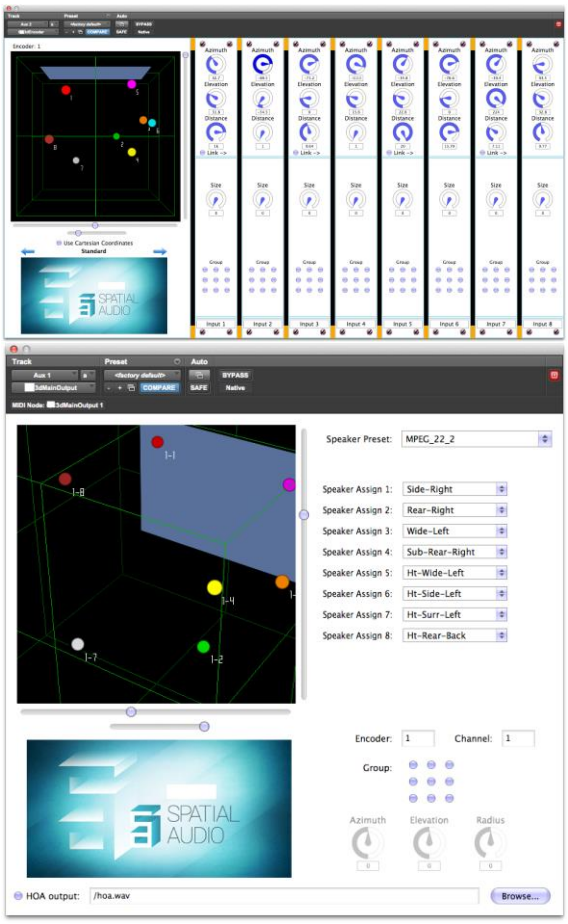


Fig. 3. Screenshots of our SBA plug-ins.

Input plug-ins convert the input signals into SBA format. One of the input side plug-in can take up to 8 mono microphones which can be spatially placed, have a source trail effect applied to moving sources, grouping across plug-ins, or use a spatial multi-tap delay effect. Any number of these plug-ins can be used within a DAW. Another plug-in takes a 32-capsule spherical microphone array signal across 4 instances on 7.1 tracks and applies the processing described in the previous section, as well as giving rotational control to the user. Another plug-in takes in classic 1st order ambsionics from a tetrahedral microphone array. One plug-in allows for already created SBA content to be inputted into the system.

A single Spatial Sound Fx plug-in works on the SBA signal as a whole. It can apply an overall gain to the SBA signal, rotate the sound field, and mirror the sound field about any axis, warping about the equator or poles and spatial filtering from a user defined direction. This plug-in allows for effects that are not possible with channel-based content or are too complex for object-based content.

On the output side there is the main output plug-in that sets up the internal SBA buss structure between the plug-ins, it has the choice of SBA renderer to use, up to 8 choices for speaker output, has the group controls across all input plug-ins, shows a visualization of all sound sources and can save SBA to a file. Secondary output plug-ins extends the main output plug-in by allowing for up to 8 more speaker outputs per instance, this allows to have more speaker outputs than a DAW's built in buss limit. Finally, there is a plug-in has the ability to directly output the SBA coefficients for real-time streaming.

## 7 Bit Compliant Compression – MPEG-H

The benefits of Scene-Based-Audio are countered by the sheer bandwidth required to transmit uncompressed HOA coefficient signals especially in comparison with bandwidth compressed channel-based Audio. For example, a 4th order HOA signal containing 25 coefficient signals, if uncompressed, would require approximately 40 mbits/s (with a sampling rate of 48000 Hz and 32 bits/sample). In comparison, current bitrates, used in the Television industry, for 5.1 channel-based delivery require approximately 400 kbits/s. In our previous paper [19], we showed how technology adopted in the recently ratified MPEG-H standard, ISO 23008-3 is able to bring the required HOA bitrate down by almost three orders of magnitude (for the lowest rates). This allows the broadcasting of HOA signals at roughly the same amount of bandwidth currently required by 5.1 channel-based audio deployments.

A two-staged approach is used to compress the HOA signal, and produce packets that are bit compliant for the MPEG-H standard. First, a spatial compression engine employs deterministic and/or stochastic techniques to decompose the signal into sets of linearly uncorrelated and energy compacted components, and thereby reduce the dimensionality of the incoming HOA signal. The decorrelated signal and the residual are subsequently coded using psychoacoustic techniques to achieve further reductions in required bandwidth.

The compression engine MPEG-H is also able to offer layered-coding for HOA coefficient signals. In a layered configuration, the base layer carries a spatially-lower-resolution version of the soundfield (along with ancillary information such as how many enhancement layers are available) while the enhancement layers supplement this with higher resolution soundfield information. An example of such a layered configuration could be to send the 0th order HOA signal in the base layer (at bitrates as low as 40 Kbps) while the enhancement layers contain the higher orders.

An MPEG-H bit compliant decoder would be able to obtain the layers in the bitstream based on the indication of layers specified in the bitstream.

This kind of layered approach is difficult to achieve with either channel- or object-based audio. For those formats, it is possible to have multiple streams coded at different bitrates, but that topology suffers from listeners being subjected to coding distortions for the low-bit rate stream – rather than a graceful degradation to a lower spatial resolution listening and, the added redundancy due to the fact that each stream is not adding supplementary information on the lower layers – but is required independent of the other layers. For channel-based audio, it is possible to have multiple streams such as one for mono, one for stereo and so on – but again those are independent streams and don't carry supplementary information costing both storage and bandwidth.

MPEG-H provides a practical and efficient compression solution to enable VR broadcast and streaming services while retaining the high fidelity of the SBA audio experience.

## 8 Rendering SBA

Since scene-based audio is agnostic to the loudspeaker layout when capturing or creating the SBA content, one needs to render the SBA on the consumer device. This rendering can be performed to produce real loudspeaker signals, or otherwise virtual loudspeaker signals that can be convolved with real loudspeaker impulse responses that correspond to the same angular coordinates. For simplicity, we take the case of a regular set of loudspeaker positions [5], those can be of the Platonic Solids, t-designs or Fliege points [6]. The vector of loudspeaker signals $S = [S_1 \ldots S_n]^T$ can be created by

$$S = D.B \qquad (1)$$

Where B is the vector of SBA signals $[B_{(0,0)} \ldots B_{(n,m)}]^T$ and D is the rendering matrix. In this simple case the rendering matrix D can be calculated from

$$D = C^{\dagger} \qquad (2)$$

Where C is a matrix of the spherical harmonics at the given loudspeaker angular coordinates. The simple case is ideal for VR since the renderer is ideal and all directions around the user are equally covered by loudspeakers.

### 8.1 SBA Direct Binauralization

The general way to binauralize a loudspeaker signal, $S_{BIN}$, was described above and is written as

$$S_{BIN} = S*IR \qquad (3)$$

Where * is convolution of time-domain signals and IR is a matrix of left and right impulse responses of the loudspeaker positions. Combining the rendering (1) and binauralization (3) stages results in

$$S_{BIN} = (D.B)*IR \qquad (4)$$

However since for an ideal SBA rendering there should be more loudspeakers than there are SBA coefficients, a more efficient way to do the binauralization is to create an SBA IR matrix

$$D_{IR} = D*IR \qquad (5)$$

Which can be done ahead of real-time processing, and then one can use the calculation below

$$S_{BIN} = B*D\_IR \qquad (6)$$

For the combined rendering and binauralization, called SBA to binauralization.

### 8.2 Interactive Features

One of the defining features for SBA being the format for VR audio production is the ability to efficiently apply user-side interactive features and/or effects without degradation to the signal. Rotation of the sound-field related to head-movement of the user is key to provide a realistic VR auditory experience. An effect such as rotation can be applied by applying a $(N+1)^2$ by a $(N+1)^2$ matrix F to the SBA signals

$$B' = F.B \qquad (7)$$

The matrix F is produced by multiplying a renderer matrix for an equally spaced set of loudspeakers points from (2) with a matrix, L, of spherical harmonics of the loudspeaker points rotated by the same amount of head movement, thus

$$F = D.L \qquad (8)$$

Once again, if desired, efficiency can be gained by applying the effects matric F to (5) giving

$$D\_IR = (D.F)*IR \qquad (9)$$

Other effects such as gaze emphasis can be applied in the same manner. Other SBA type effects are described by Kronlachner [7].

## 9 Subjective Testing

Even though SBA has benefits to object and channel-based audio for VR in terms of ease of acquisition, production effects and user-side interactive features, SBA also performs as well, if not better than traditional channel-based and object-based audio formats.

In [17] Palacino et. al. found that a first order ambisonics recording binaurally rendered was almost equivalent to that of a dummy head recording. They found that with increased order that the ITD cue accuracy increased. Since SBA can be reproduced using HRTFs/BRIRs that are better suited to an individual, it can be argued that for some listeners this would be better than the dummy-head where there is only a single choice of impulse response.

Rumsey [18] presents work by Powers et. al. which shows that higher-order ambisonics can perform vastly superior to first order ambisonics dependent on the source content. It also shows that in the content favorable to higher-orders that first order performs worse than 5.1 channel-based content and for the unfavorable content third, first and 5.1 representations produce no statistically significant better results. Ergo, higher order ambisonics is the better choice since it never performs worse than first order or channel-based content, but it can perform drastically better in the right conditions.

Braun and Frank [16] published a comparison between first and fourth order ambisonics using both recorded and synthetically constructed scenes. In both cases, the higher order representation was better than first order. There was a larger difference in recorded content which lends itself to consumer captured VR content and real-space recordings of VR for the "infinite seat" scenario.

## 10 Conclusions

The ease of rotating the sound scene when the underlying format is HOA has been known for a long time. Previous to the recent spate of interest in VR, this property of HOA was used to provide better binauralization experiences – by allowing the playback system to adapt to the slightest of head movement – alleviating the problem of front/back confusions that is prevalent in headphone renderings that do not adapt [8]. Using HOA for VR goes further by adapting not just to the slightest of head movements but complete rotations. It is for this reason and the various disadvantages of the other audio formats, described in this paper, that HOA is a perfect match for VR audio experiences. Not only is HOA suitable for VR, it is also suitable for 360 degree video. Interactive effects such as zooming into every spatial sub-section of a live scene is only possible with HOA.

This paper has discussed VR audio content production using SBA. It has detailed the capture, processing and reproduction steps involved in producing an immersive auditory scene. Through the discussion of use cases and comparisons with channel and object-based audio the authors have shown compelling reasons to use SBA for VR. Moreover SBA for live capture is truly the most effective and computationally efficient method currently available. Although raw SBA signals can have a large bitrate it has been shown that through compression schemes such as MPEG-H that the bandwidth requirements can be dramatically reduced to comparable bitrates that are used today for 5.1. The authors have demonstrated this technology at tradeshows such NAB, conferences and company-based demonstration days with huge amounts of positive feedback. Through the development of such a system it was determined that low latency and exact sync between auditory and visual cues are of the upmost importance to deliver a non-nauseating, realistic and all-round immersive virtual reality experience to an end user.

## References

[1] URL: https://www.qualcomm.com/scene-based-audio

[2] Wang D. L., and Brown G. J., "*Computational auditory scene analysis: Principles, algorithms, and applications*", Wiley-IEEE Press, 2006

[3] Lindau A., "*The perception of system latency in dynamic binaural synthesis*", Proc. of 35th DAGA, pp 1063-1066, 2009

[4] Pulkki, V., "*Virtual Sound Source Positioning Using Vector Base Amplitude Panning*", J. Audio Eng. Soc, Vol. 45, No. 6, pp 456-466, 1997

[5] Daniel J., "Spatial Sound Encoding Including Near Field Effect: Introducing Distance Coding Filters and a Viable, New Ambisonic Format" *23rd International Conference: Signal Processing in Audio Recording and Reproduction*, Audio Engineering Society, 2003

[6] Fliege J. and Maier U., "*The distribution of points on the sphere and corresponding cubature formulae*", IMA Journal of

Numerical Analysis, Volume 19, Issue 2, Pp. 317-334, 1999

[7] Kronlachner M., "*Spatial Transformations for the Alteration of Ambisonic Recordings*", Master Thesis, University of Music and Performing Arts, Graz, June 2014

[8] Wallach H. "*The role of head movements and vestibular and visual cues in sound localization*", J. Exp. Psychol. 27, pp 339–368 10.1037/h0054629, 1940

[9] URL: https://theta360.com/en/

[10] URL: https://gopro.com/odyssey

[11] URL: http://en-us.sennheiser.com/shape-the-future-of-audio-ambeo

[12] URL: http://www.core-sound.com/TetraMic/1.php

[13] URL: http://www.mhacoustics.com/products

[14] URL: http://visisonics.com/products-2/

[15] URL:http://www.matthiaskronlachner.com/?p=2015

[16] Braun S. and Frank M., "*Localization of 3D Ambisonic recordings and ambisonic virtual sources*", Proc. Int. Conf. Spatial Audio, pp. 21-26.

[17] Palacino J., Nicol R., Emerit M. and Gros L., "*Perceptual assessment of binaural decoding of first-order ambisonics*", Société Française d'Acoustique. Acoustics 2012

[18] Rumsey F., "*Immersive Audio, Objects, and Coding*", JAES Volume 63 Issue 5 pp. 394-398, May 2015

[19] Sen D., Peters N., Kim M. and Morrell M., "*Efficient Compression and Transportation of Scene Based Audio for Television Broadcast*", 2016 AES International Conference on Sound Field Control, July 2016

[20] Lindau A., Maempel H. and Weinzierl S., "*Minimum BRIR grid resolution for dynamic binaural synthesis*", Journal of the Acoustical Society of America, vol. 123, no. 5, pp. 3851--3856, 2008

[21] Hanschke J., Fleischmann F., Plogsties J. and Fug S., "*Dynamische binaurale Raumsynthese in der 3D-Tonwiedergabe - eine Untersuchung zur Qualitat verschiedener virtueller Lautsprecherkonfigurationen*", Proc. of 42nd DAGA, 2016

[22] SMPTE ST 2036-2-2008, "*Ultra High Definition Television Audio Characteristics and Audio Channel Mapping for Program Production*", 2008

[23] Tsingos N., Chabanne C., Robinson C. and McCallus M., "*Surround Sound with Height in Games Using Dolby Pro Logic IIz*", 41st International AES Conference - Audio for Games, 2011

[24] Vaananen R. and Huopaniemi J., "*Advanced AudioBIFS: virtual acoustics modeling in MPEG-4 scene description*", IEEE Transactions on Multimedia, vol. 6, no. 5, pp. 661-675, 2004

[25] Bertet S., Daniel J., Parizet E. and Warusfel O., "*Influence of microphone and loudspeaker setup on perceived higher order ambisonics reproduced sound field*", Proceedings of ambisonics symposium, 2009

[26] ISO/IEC 23008-3:2015, "*Information technology --- High efficiency coding and media delivery in heterogeneous environments --- Part 3: 3D audio*", 2015

[27] Noisternig M., Musil T., Sontacchi A. and Holdrich R., "*3d binaural sound reproduction using a virtual ambisonic approach*", IEEE International Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems, VECIMS'03, 2003

[28] Gerzon M. A., "*Practical Periphony: The Reproduction of Full-Sphere Sound*", 65th AES convention, 1980