SMPTE Meeting Presentation

Scene-based Audio Implemented with Higher Order Ambisonics (HOA)

Nils Peters

Qualcomm Technologies, Inc., San Diego, CA, USA, npeters@qti.qualcomm.com

Deep Sen

Qualcomm Technologies, Inc., San Diego, CA, USA, dsen@qti.qualcomm.com

Moo-Young Kim

Qualcomm Technologies, Inc., San Diego, CA, USA, kimyoung@qti.qualcomm.com

Oliver Wuebbolt

Technicolor, Hannover, Germany, Oliver.Wuebbolt@technicolor.com

S. Merrill Weiss

Merrill Weiss Group LLC, Metuchen, NJ, USA, merrill@mwgrp.com

Written for presentation at the

SMPTE 2015 Annual Technical Conference & Exhibition

Abstract. *Scene-based Audio uses a sound-field technology called "Higher Order Ambisonics" (HOA) to create holistic descriptions of both live-captured and artistically-created sound scenes that are independent of specific loudspeaker layouts. For efficient representation, the audio can be carried as a set of PCM channels that contain predominant sounds and ambience in separate tracks. Standard audio bandwidth-compression techniques then can be applied to the PCM channels. This approach is in contrast to conventional channel-based sound representations, in which one signal is used for each loudspeaker of a target reproduction system, with the implication that upmixing or downmixing is required when loudspeaker configurations other than the intended one are used for actual reproduction. This paper examines how, with Scene-based Audio, there can be satisfactory reproduction of immersive sound at bitrates corresponding to the equivalent of just 6 channels, while an alternative sound-field method that exclusively employs audio objects typically involves much higher bitrates.*

Keywords. Scene-based Audio, HOA, Spatial Audio, Next-Generation Audio, MPEG-H, ATSC 3.0

Introduction

In this paper we show how the practical Scene-based audio concept, specifically Higher Order Ambisonics (HOA), can be used to broadcast immersive audio efficiently.

This paper is organized as follows: First we compare Scene-based Audio with the other two prominent immersive audio concepts – Channel-based Audio and Object-based Audio. Next we provide further details on the principles of HOA. Then the Scene-based compression specified in the recent MPEG-H 3D Audio standard is described, including a review of other approaches. The paper ends with an explanation of how HOA is rendered specifically to a consumer's personal reproduction environment.

The Three Immersive Audio Concepts

There are three distinct immersive audio concepts, recognized and supported by the International Telecommunication Union (ITU) [1] and the European Broadcast Union (EBU) [2]. These concepts are entitled Channel-based Audio, Object-based Audio, and Scene-based Audio. Each of them offers different features and advantages to content creators and audio consumers while fundamentally differing in the ways in which content is broadcast and then is decoded and processed.

Channel-based Audio

Channel-based Audio has been the de facto format in the broadcast industry with the wide deployment of Stereo 2.0 and Surround Sound 5.1 formats. The name originates from the fact that loudspeaker feeds are mixed by content creators for pre-defined (standardized) loudspeaker setups and delivered in pre-defined sequences (see Figure 1). Because content creators ``bake" the content into loudspeaker channels, complex rendering equipment at the decoder is not necessary. To perceive the intended audio image, consumers have to connect their loudspeakers in the correct order and and have all loudspeakers placed as dictated by the channel-based format. This has the drawback that a loudspeaker setup that does not match the intended setup will result in a degraded sound image.

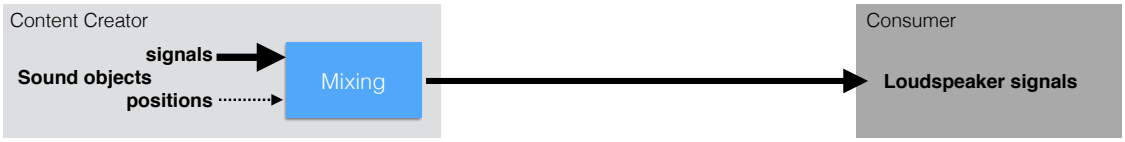

Figure 1. Channel-based Audio concept

Object-based Audio

In contrast to Channel-based Audio, Object-based Audio delivers unmixed audio components in the form of audio objects in conjunction with object metadata. Based on these metadata, the audio objects are rendered to final loudspeaker feeds at the listening location. Rendering at the listening location has the advantage that each consumer's loudspeaker configuration is taken into account in creating the loudspeaker feeds (see Figure 2). Therefore, individual non-

standardized loudspeaker configurations as well as legacy and future immersive loudspeaker configurations are supported. Further, if the system and content allow for it, individual audio objects can be manipulated by users, e.g., by changing the location or the loudness of each object. On one hand, this enables a personalized audio experience; on the other hand, it reduces the content creator's control over the reproduced mix compared to channel-based content.

An aspect that often is overlooked in Object-based Audio is the increased cost of transmitting and rendering of audio objects:

The exclusive use of dynamic Object-based Audio in the broadcast world is challenging because of the transmission bandwidth needed and the processing power required to decode and render many audio objects simultaneously. To circumvent this challenge, audio objects often are advocated for use in combination with Channel-based or Scene-based Audio.

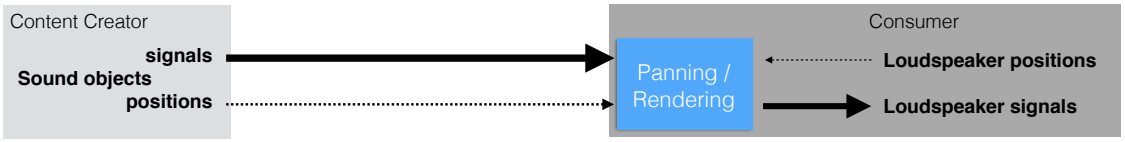

Figure 2. Object-based Audio concept

Scene-based Audio

A concept that is receiving increasing attention is Scene-based Audio. Scene-based Audio is especially interesting for broadcasting because it combines benefits known from both Channel-based and Object-based Audio. Similar to Channel-based Audio, the content creator ``bakes'' the content into a fixed number of audio signals. In contrast to Chanel-based Audio, and similar to Object-based Audio, however, Scene-based Audio signals are agnostic with respect to the loudspeaker reproduction layout. In Scene-based Audio, content is transmitted in the form of so-called spherical harmonic coefficient signals, which describe all sounds and their time-varying spatial properties within a sound scene, independent of any particular loudspeaker setup. Based on the actual loudspeaker configuration at the listening location, these spherical-harmonics signals are rendered into appropriate loudspeaker signals (see Figure 3). Consequently, non-standard, legacy, as well as future immersive loudspeaker configurations can be accomodated – a benefit in common with Object-based Audio. In contrast to Object-based Audio, however, in Scene-based Audio, the number of channels is constant and independent of the number of sound sources. Thus, bandwidth cost and renderer complexity do not scale with scene complexity and are easy to predict for Scene-based Audio systems.

Figure 3. Scene-based Audio concept

With Scene-based Audio, content creation in both live-capture and cinema-based workflow situations is possible. Live capture is especially simplified. Using a microphone array such as the baseball-sized em32 Eigenmike (see Figure 4), along with any number of optional spot

mics, it is possible to capture live immersive sound scenes realistically with little to no human intervention. As with the workflow for Object-based Audio, one also can design scene-based content in a digital audio workstation (DAW) by panning audio-objects to desired positions. Captured live or designed in a DAW, Scene-based Audio also offers a new palette of audio effects to shape immersive content. For instance, it is easy to rotate, mirror, and stretch an entire sound scene (c.f. [3]) or to ``photoshop" sounds originating from a specific direction.

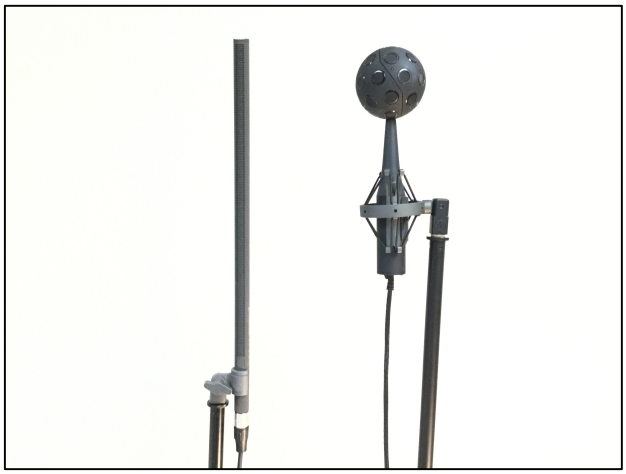

Figure 4. The em32 microphone array (right) and for visual size comparison a Sennheiser MKH 8070 directional microphone (left)

Table 1 provides an overview of the benefits of the three different immersive audio concepts. Noteworthy, all three concepts can be combined simultaneously to create any desired degree of personalization (such as with multiple language objects) and coding efficiency.

Table 1. Comparison of benefits provided by the three immersive audio concepts

Feature	Channel -based	Object -based	Scene -based
Easy and efficient live capture of immersive sound scenes			x
Easy post-processing of live capture of immersive sound scenes			x
Creation of immersive content within DAWs	x	x	x
Final creator control over mix through ``baking" content into fixed number of audio channels	x		x
Decoder complexity independent from scene complexity	x		x
Transmitted audio signals independent of consumer loudspeaker layout		x	x
Interactive user control of single audio elements		x	
Binaural headphone rendering enabled	x	x	x
Easy deployment in VR applications		x	x

Principles of Higher Order Ambisonics (HOA)

Higher Order Ambisonics [4, 5] represents a three-dimensional dynamic sound scene by using a set of orthogonal basis functions. These basis functions are the so-called Spherical Harmonics, which are hierarchically organized by an increasing order *N*. Figure 5 shows the polar pattern of these harmonics from order *N*=0 (top row) to order *N*=3 (bottom row). With increasing order, these harmonics become more and more spatially directive and, because higher order harmonics build up on lower order harmonics, the spatial resolution of the scene representation increases with each additional order and is (for *N*>3) approximately 180/*N* degrees [6].

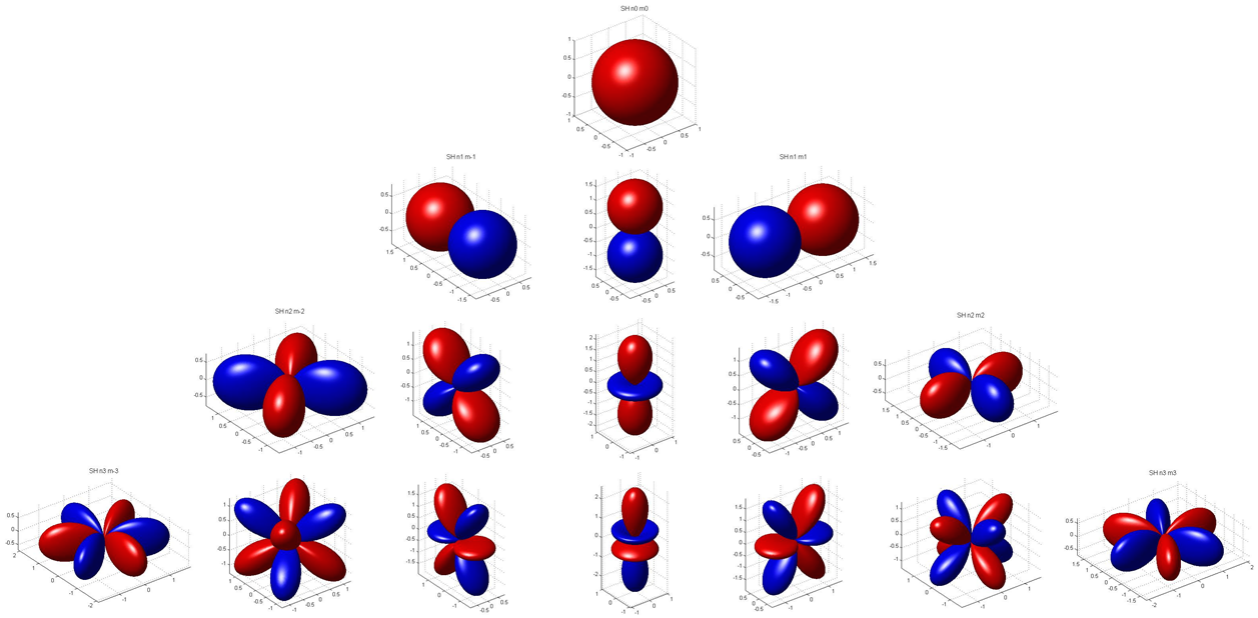

Figure 5. Spherical harmonics from order N=0 (top row) to order N=3 (bottom row)

The order of the spherical harmonics is theoretically unlimited; in practice a sound scene is described with spherical harmonics up to a given order *N*. This order *N* determines the number of audio signals (a.k.a. HOA coefficient signals) required to convey the sound scene, which number is equal to $(N+1)^2$. The HOA coefficient signals describe how much weight is given to the associated Spherical Harmonic basis functions over time.Typically, the order *N* is between 3 to 6, resulting in 16 to 49 HOA coefficient signals.

Because HOA is a hierarchical format, it s possible to remove high-order coefficient signals of an HOA signal without loss of audio elements in the scene but with the penalty of reduced spatial resolution.

Compression of Scene-Based Audio

Review

Evolving from research on first-order Ambisonics, compression of HOA coefficient signals has been investigated in recent years. The authors of [7] and [8] studied whether HOA coefficient signals could be coded with traditional audio coders, such as AAC. It was found that quantization and bit allocation, as a function of the HOA order, shape the perceived reproduction error within the listening area and that the sound quality of the lower orders is more important than the sound quality of higher order HOA coefficients.

Daniel [10] presented an approach to dynamically threshold HOA coefficient signals based on a psycho-acoustic model that accounts for spatial masking and the minimum audible angle (MAA).

An extension to Directional Audio Coding (DirAC) to accomodate HOA was proposed in [10]. In the conventional DirAC [11], an energy-based time-frequency analysis is used to estimate the dominant direction and a diffuseness value per frequency bin from a 1^{st}-order Ambisonics signal. The transmitted payload consists of these frequency-dependent directional and diffuseness parameters and the onmidirectional 0^{th}-order ambisonics coefficient signal (c.f. Figure 5), which can be further compressed with a perceptual audio codec. In the proposed HOA extension [10], an N^{th}-order HOA signal is subdivided into $M \geq N$ even-sized spatial sectors, and each of them is processed with conventional DirAC. The transmitted payload increases to M times the frequency-dependent directional and diffuseness parameters and M audio signals.

Somewhat related to DirAC, the authors of [12] present a method entitled Spatially Squeezed Surround Audio Coding (S^3AC). Here, the sound at the dominant direction of each time-frequency bin is extracted from the three-dimensional soundfield and downmixed into a two-channel stereo signal in which the original sound direction is monotonically squeezed into a direction in the ±30° stereo field. The downmix can be further compressed with a traditional audio coder to create the compressed bitstream.The S^3AC decoder analyzes the stereo signal and remaps the squeezed directions to their original values.

Efficient Compression of Scene-Based Audio for Broadcasting

The fundamental assumption of the scene-based audio compression presented in this paper is that the sound scene consists of a small and time-variant number of predominant sounds (the foreground) and an ambient soundfield (the background). This concept models the way humans decipher complex soundfields: Humans tend to pay attention to a few important sources at a given time [13, 14]. Even if a soundfield consists of many contributing sound sources (e.g, an orchestra or a cocktail-party), sound sources sensed as less important fade away into a diffuse and reverberant background where spatial and timbral properties are less critical.

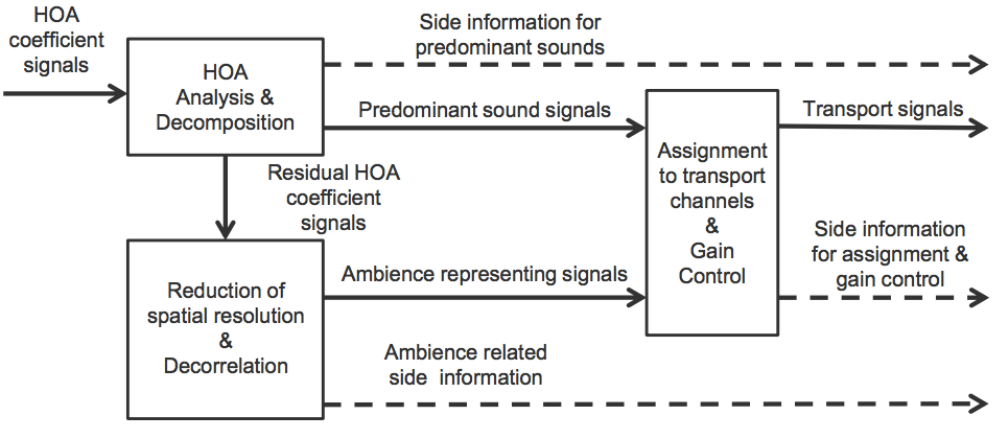

Figure 6. HOA Spatial Encoder, [15]

Figure 6 exemplifies a scene-based encoder that can be used to create MPEG-H 3D Audio [16, 17] bitstream. The *HOA Analysis & Decomposition* module analyzes the input HOA coefficient signals over a short time frame and extracts a small number of predominant components of the sound scene during that given period. The length of the time frame determines the temporal resolution of the soundfield analysis, and the number of preserved predominant sound elements may depend on the content as well as on the desired spatial compression ratio.

Each predominant sound element is represented by monaural PCM audio data plus parametric side information describing its spatial properties (i.e., spatial distribution, source width). The residual HOA coefficient signal describes the ambient background of the sound scene with all remaining (subsidiary) sounds in the HOA domain. Because the spatial resolution is less critical for this ambient soundfield, its HOA order can be reduced without significant perceptual impact on the overall sound scene, resulting in fewer coefficients required to describe the ambience. To prevent audible artifacts due to spatial unmasking of quantization noise, the low-order HOA ambience coefficients can go through a decorrelation process before further bandwidth-compression.

In summary, due to decomposition of HOA input content into predominant sound elements and order-reduced ambient background elements, HOA content can be reduced to a lower number of PCM transport signals plus associated side information. Satisfactory results have been achieved with as few as six transport signals. Notably, the number of transport signals is independent of the HOA order of the original content but, rather, depends on the desired spatial compression ratio. The compression ratio is approximately $(N+1)^2/T$, with N being the order of the HOA input content and T being the number of resulting PCM transport signals.

Subsequently the resulting PCM transport signals are coded using the MPEG-H core coder and are multiplexed with the HOA side information to create the final bitstream (see Figure 7). To

prepare the PCM signals for the *Core Encoder*, an automatic gain control ensures that all PCM signals have an optimal amplitude (see Figure 6).

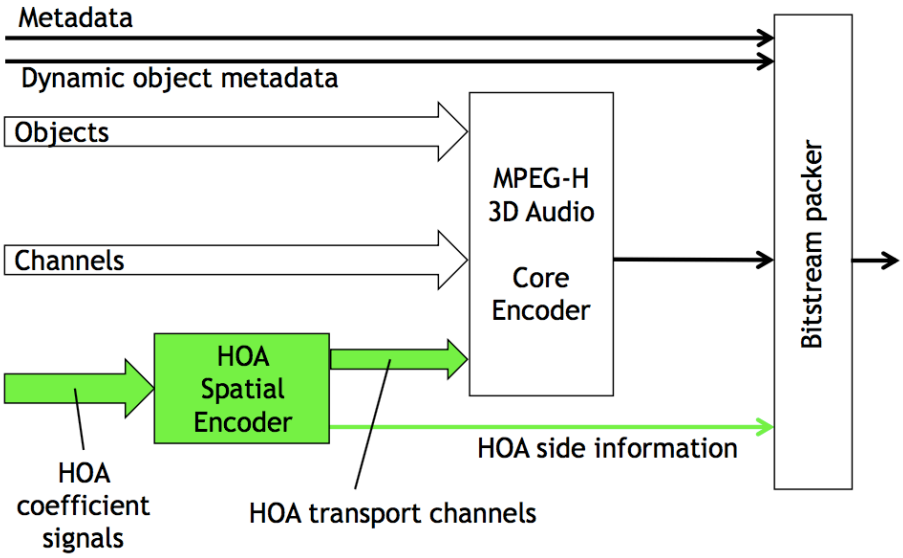

Figure 7. MPEG-H Reference Encoder, [15]

In combination with the MPEG-H Core Encoder, the HOA content can be coded with a rate-distortion curve as illustrated in Figure 8. Excellent broadcast quality (≥80 MUSHRA points) is currently achieved at about 300 kbps, independent from the HOA order of the input content. Because this coding technology is at the beginning of its optimization phase, the encoder efficiency likely will improve further over time.

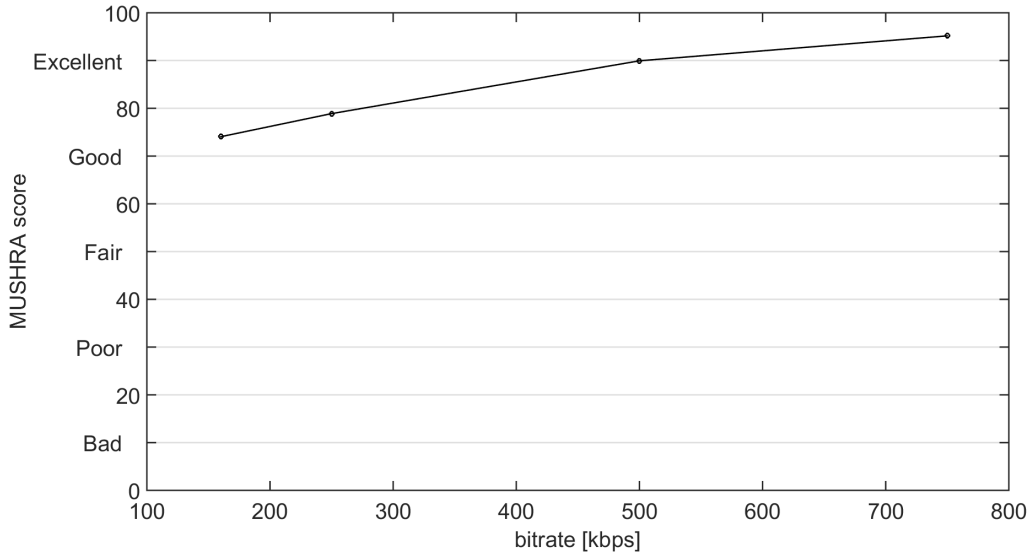

Figure 8. Rate-distortion curve for HOA coding with MPEG-H

Layered Coding with HOA

A coding strategy especially known from the video broadcasting field is to divide media content into hierarchical coding layers. In a layered (a.k.a. scalable) coding method, a base layer delivers media content in a basic audio quality that can be stepwise raised by having access to hierarchical enhancement layer(s). This layering makes it possible, for example, to apply a priority-based error protection scheme to the different bitstream layers or to broadcast the same content to devices with different decoding capabilities.

Because HOA is a hierarchical format (c.f. Section 3), it is perfectly suited for layered coding.The layered coding method in HOA is different from other multichannel layered coding approaches in which, for instance, the frequency spectrum is divided (e.g., no high frequencies in the base layer) or in which scene elements are prioritized, resulting in a base layer with just a subset of all audio elements.

In HOA, a base layer could be created by encoding just the few low-order HOA coefficient signals of the HOA input content. From this base layer, it still is possible to reproduce the entire immersive sound scene but with a reduced spatial accuracy. When the remaining higher-order HOA coefficient signals are compressed into enhancement layer(s), decoders can stepwise regain the spatial resolution of the original HOA content.

Decoding of Scene-based Audio

An overview of the MPEG-H 3D Audio Decoder is shown in Figure 9.

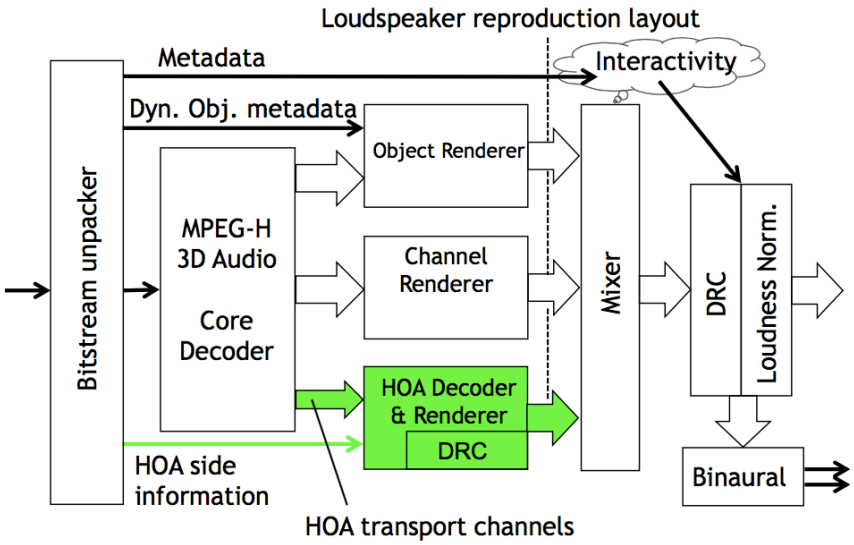

Figure 9. MPEG-H Reference Decoder

The *Bitstream unpacker* module takes the received bitstream and extracts the HOA side information together with other metadata. The *Core Decoder* module decodes the remaining portion of the bitstream into PCM signals. These PCM signals can be loudspeaker feeds (channel-based), audio objects (object-based), and/or HOA transport signals. In the case of the latter, the HOA decoder first restores the amplitude of the transport signals and sorts the transport channels into predominant sounds elements and ambient background elements (see Figure 10). By utilizing the side information, the predominant sound elements are spatially restored into the predominant soundfield. By multiplexing this predominant soundfield with the ambient background soundfield, the final HOA coefficient signals are reconstructed in the *HOA Composition* module.

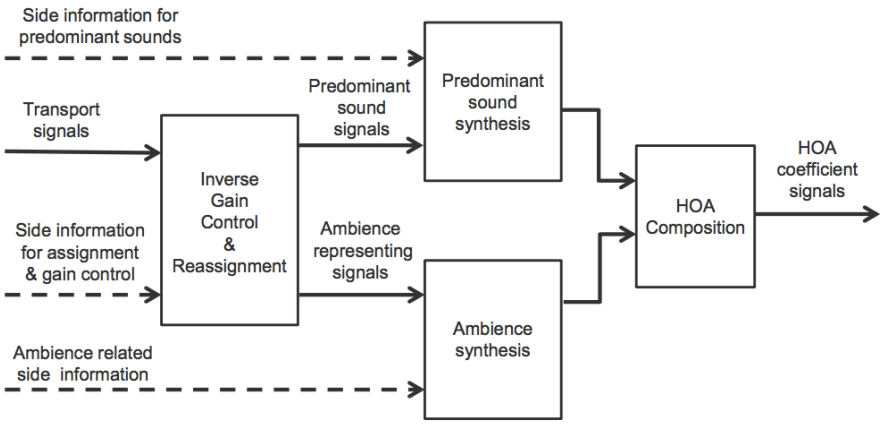

Figure 10. HOA Spatial Decoder

Listener-centric Post-processing

For generating the final audio output, the decoded HOA content is further processed, taking into account the actual reproduction environment. Device-specific or listener-defined settings for Dynamic Range Control (DRC) can be applied, too. Loudness-normalization can be applied to conform to the CALM act or to EBU R128.

Dynamic Range Control (DRC)

Dynamic range control allows for dynamic loudness adaptation to the user's listening situation. For instance, a mobile home cinema is capable of delivering a high dynamic range, while the loudspeakers on a tablet are not designed for that. Similarly, based on DRC gain sequences in the bitstream and a user-specified DRC preset, the dynamic range of the HOA content can be modified, e.g., to reduce the dynamic range for late-night TV watching. Algorithms for DRC modify the dynamic range of loudspeaker feeds. To apply those DRC algorithms in the HOA context, the HOA content is rendered to a pre-defined virtual loudspeaker setup with $(N+1)^2$ nearly equidistant loudspeakers. For these $(N+1)^2$ pre-defined virtual loudspeakers, the (multiband) DRC gain sequences are stored as gain parameters in the bitstream and the desired DRC effect is applied accordingly, similar to the processing for channel-based content. Finally, all $(N+1)^2$ DRC-processed virtual loudspeaker signals are converted back into the HOA domain.

Rendering

To listen to the scene-based audio content over loudspeakers, scene-based audio content is rendered into loudspeaker feeds. This process is efficiently carried out by multiplying the decoded HOA content with a rendering matrix. This matrix can be generated based on the number of decoded HOA coefficients and the locations of the loudspeakers in the reproduction environment. Several approaches for generating an optimal rendering matrix can be found, for example, in [16, 18, 19].

In contrast to the situation with channel-based audio, the infamous workaround of upmixing and then downmixing to accommodate loudspeaker configurations that differ from the configuration for which content was intended is not necessary in scene-based audio.

The HOA order N (and therefore the number of $[N+1]^2$ HOA coefficient signals) does not dictate a certain number of loudspeakers or a certain loudspeaker arrangement. Even HOA content with a small order (e.g., $N=2$, $[N+1]^2 = 9$ reconstructed HOA coefficient signals) can be rendered to a large loudspeaker array; and reconstructed HOA content with a higher order (e.g., $N=6$, 49 HOA coefficient signals) can be rendered for stereo. This all just depends on the rendering matrix, which adapts to the conditions in the listening environment.

Headphone Rendering – Binauralization

An optional processing step for headphone listening is binauralization of the audio content.

The aim is to provide the headphone listener with the sensation of listening to the rendered loudspeaker signals in a desired room. The auditory character of this room is captured by a set of so-called Binaural Room Impulse Responses (BRIRs).

For creating binaural headphone signals, each rendered loudspeaker signal is filtered with its associated BRIR and summed together to create a left ear and a right ear headphone signal.

An advantage of Scene-based audio is that it can be binauralized, including dynamic head movements. Dynamic head movements are important for creating a convincing illusion, including out-of-head sensation and accurate localization without front-back confusion [20]. This makes scene-based audio a viable option for emerging virtual (VR) reality or augmented reality (AR) applications. For this purpose, the HOA content is inversely rotated to compensate for head movements sensed by a head tracker. The time-varying soundfield rotation is performed by multiplying the $(N+1)^2$ HOA coefficient signals with a $(N+1)^2$ x $(N+1)^2$ rotation matrix before binauralizing [21].

Conclusion

In this paper, we discussed why Scene-based audio, in particular Higher Order Ambisonics (HOA), is a practical, elegant solution for creating and transmitting immersive content for next generation audio services.

We outlined the advantages of HOA and compared it to other immersive-audio concepts. We further described how, in MPEG-H, Scene-based Audio can be coded and transmitted with high efficiency and rendered specifically to a consumer's personal reproduction environment.

More information on Scene-based Audio can be found in a white paper [22].

References

[1] ITU-R BS.2076-0 Audio Definition Model. Geneva, Switzerland: International Telecommunication Union, June 2015.

[2] EBU Tech 3364: Audio Definition Model. Geneva, Switzerland: European Broadcasting Union, 2014.

[3] M. Frank, F. Zotter, and A. Sontacchi "Producing 3D audio in ambisonics," in Proc. of the 57th International AES Conference: *The Future of Audio Entertainment Technology - Cinema, Television and the Internet*.,2015.

[4] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of the Audio Engineering Society*, vol. 53, no. 11, pp. 1004 - 1025, 2005.

[5] J. Daniel, "Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia," Ph.D. dissertation, University of Paris VI, France, 2000.

[6] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *Journal of the Acoustical Society of America*,vol. 116, no. 4, pp. 2149 - 2157, 2004.

[7] E. Hellerud, A. Solvang, and U. P. Svensson, “Spatial redundancy in higher order ambisonics and its use for lowdelay lossless compression, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 269 - 272.

[8] I. Burnett et al., “Encoding higher order ambisonics with AAC," in *Proc of the 124th AES Convention*, 2008.

[9] A. Daniel, R. Nicol, and S. McAdams, “Multichannel audio coding based on minimum audible angles," in *Proc. of the 40th International Conference: Spatial Audio: Sense the Sound of Space.*, 2010.

[10] V. Pulkki et al., “Parametric spatial audio reproduction with higher-order B-format microphone input," in *Proc of the 134th AES Convention*, 2013.

[11] V. Pulkki, “Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503 - 516, 2007.

[12] B. Cheng et al., “A general compression approach to multi-channel three-dimensional audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1676 - 1688, 2013.

[13] M. Kashino and T. Hirahara, “How many concurrent talkers can we hear out," in *Proc. of the Acoustical Society of Japan*, 1995, pp. 467 - 468.

[14] D. Huron, Voice denumerability in polyphonic music of homogeneous timbres," *Music Perception*, vol. 6, no. 4, pp. 361{382, 1989.

[15] A. Krueger, Higher Order Ambisonics relevant architecture," in ISO/MPEG-H Audio - The New Standard for Universal Spatial / 3D Audio Coding, 138th AES Convention, Warsaw, PL, 2015.

[16] ISO/IEC 23008-3:2015, Information technology - High effciency coding and media delivery in heterogeneous environments - Part 3: 3D audio, 2015.

[17] J. Herre et al., “MPEG-H 3D Audio – the new standard for coding of immersive spatial audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770 - 779, Aug 2015.

[18] F. Zotter, M. Frank, and H. Pomberger, “Comparison of energy-preserving and all-round ambisonic decoders," in *Proc. of the German Annual Conference on Acoustics* (DAGA), 2013.

[19] A. J. Heller, E. Benjamin, and R. Lee, “A toolkit for the design of ambisonic decoders," in *Proc. of the Linux Audio Conference*, 2012.

[20] D. R. Begault, E. M. Wenzel, and M. R. Anderson, “Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *Journal of the Audio Engineering Society*, vol. 49, no. 10, pp. 904 - 916, 2001.

[21] M. Noisternig et al., “3D binaural sound reproduction using a virtual ambisonic approach," *IEEE International Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems (VECIMS'03)*, 2003, pp. 174 - 178.

[22] Qualcomm Technologies, Inc. (2015, April). [Online]. Available: https://www.qualcomm.com/documents/whitepaper-scene-based-audio-mpeg-h

