Audio Engineering Society

Conference Paper

Presented at the Conference on
Sound Field Control
2016 July 18–20, Guildford, UK

This conference paper was selected based on a submitted abstract and 750-word precis that have been peer reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This conference paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Efficient Compression and Transportation of Scene Based Audio for Television Broadcast

Deep Sen, Nils Peters, Moo Young Kim, and Martin Morrell

Qualcomm Multimedia R&D.

Correspondence should be addressed to Deep Sen (dsen@qti.qualcomm.com)

Scene-based Audio is differentiated from Channel-based and Object-based Audio in that it represents a complete soundfield without requiring loudspeaker feeds or audio-objects (with associated meta-data) to recreate the soundfield during playback. Recent activity at MPEG [1], ATSC and DVB has seen proposals for the use of Higher-Order-Ambisonics (HOA) for Scene-based Audio. The many benefits of Scene-based Audio is countered by the bandwidth requirements as well the ability to transport the multitude of HOA coefficient channels through current day Television plants. In this paper, we report on research and standardization activities directed at solving these issues. These solutions enable the Television broadcast and delivery of both live-captured and artistically-created sound scenes using HOA.

1 Introduction

Scene-based Audio or the representation of Acoustic 'scenes' or soundfields provides many benefits over Channel-based or Object-based Audio. When using HOA, some of these benefits include the ability to capture live and accurate 3D audio while also catering for artistically-created 3D scenes; the ability to render accurate and consistent audio over practically any number of loudspeakers as well as headphones; and the ability to easily adapt the rendered soundfield to a selected Point of View (PoV) as well as adapt to head movements of a headphone wearing listener - allowing for compelling Virtual Reality renditions.

In addition, a palette of new tools become available to the Sound Mixer for both live and offline broadcast. These include the ability to attenuate sound from certain spatial locations; and the ability to augment and amplify sounds from specific directions and microphones. At the playback end, the listener is able to interact with the scene by amplifying certain directions and allow for his/her selected PoV. These are in addition to the interactive features made available when transmitting extra audio elements (objects) in addition to HOA.

The benefits listed above 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 this paper, we show how technology adopted in the recently ratified MPEG-H standard, ISO 23008-3 [1] 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 required by 5.1 Channel-based Audio. A further bottleneck in broadcasting HOA lies in the infrastructure that is used in current Television plants. The workflow in a typical US TV plant is shown in Figure 1.

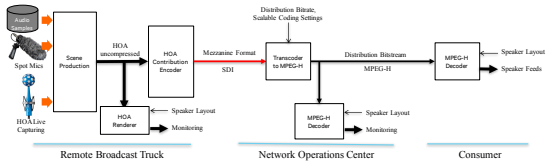

Figure 1. Workflow in a typical TV plant

Most Television plants use SD-SDI or HD-SDI infrastructure to transport their Audio channels through live broadcast trucks, Network-Operating-Centers and their Affiliate stations before emission codecs such as MPEG-H have a chance of compressing the signals into a single stream. This essentially limits the number of discrete PCM signals in the TV plants to 8 or 16 – a far cry for anything greater than a 3rd order HOA signal. An SD-SDI based framework can use the 8 embedded PCM channels to send a 5.1 program along with a stereo program for example. An HD-SDI framework allows 16 embedded PCM channels to be transported through the TV plant. While a second or third order HOA program would fit into the HD-SDI framework, it would significantly limit the TV network from adding other signals that add to the interactive/personalized experience (such as multi-lingual or home/away commentary in a sports broadcast) or signals that are meant to aid the visually impaired. The techniques used in the MPEG-H standard are again applicable to solve this limitation [20]. The MPEG-H standard employs a two-staged approach to compress the HOA signal. First, a spatial compression engine employs deterministic and/or stochastic techniques to decorrelate the signal 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. Applying only the first stage of the MPEG-H encoding process allows the HOA signal to be represented in the dimensionality reduced form, allowing it to be easily transported through the TV plant and even leaving enough signals available for features such as interactivity and personalization.

In this paper, we go through the entire value proposition of Scene-based Audio for Television broadcasting - emphasizing the signal processing techniques that make it possible to broadcast immersive, interactive, personalized and 3D audio to consumers. The next sections of the paper will go through an introduction to HOA, discuss the MPEG-H HOA compression technology and discuss the transport of MPEG-H through television plants.

2 Higher Order Ambisonics

The philosophy of Scene-based audio is to represent a localized pressure field $p(x,y,z,t)$ as accurately as possible. To do this using Higher Order Ambisonics, the pressure field is represented as a solution to the Wave equation [8] using Spherical Harmonic basis functions:

$$p(r,\theta,\phi,\omega,t) = \left[\sum_{n=0}^{\infty} j_n\left(\frac{\omega r}{c}\right) \sum_{m=-n}^{n} a_n^m(\omega,t) Y_n^m(\theta,\phi)\right] e^{j\omega t} \quad (1)$$

where, c is the speed of sound, $j()$ is the Spherical Bessel function of degree n and $Y_n^m(\theta,\phi)$ are the Spherical Harmonic functions of order n and degree m. This can viewed as a decomposition of the spatial components of the pressure field on the orthonormal basis functions, $Y_n^m(\omega,t)$. The decomposed coefficients $a_n^m(\omega,t)$ completely describe the soundfield and are known as the Spherical Harmonic coefficients, HOA coefficients, HOA signals or just as the 'coefficient signals'. For practical purposes, the infinite sum in Equation 1 is truncated to n=N, resulting in $(N+1)^2$ coefficient signals.

At the outset, a few advantages of representing the soundfield in this manner should be pointed out:

1) To rotate the soundfield, one needs to just multiply the coefficients, $a_n^m(\omega,t)$, by an appropriate rotation matrix. This makes the format highly conducive to applications such as immersive playback over headphones where tracking head-rotations and movements of the listener is either a necessity (for Virtual Reality type experiences) or at the least makes the

binaural listening experience more compelling [18,19].

2) It is easy to record a live soundfield [2] as these coefficients using a number of off-the-shelf microphones including the Soundfield microphone and the well-known Eigenmike™ from mhAcoustics. Once in this format, a plethora of new tools become available to a mixer. These include the ability to spatially attenuate sounds emanating from specific regions of the 3D space.

3) It is also possible to support the offline cinematic or TV episodic workflow where a soundfield is created from audio 'stems'. This is done by modelling the stems as plane or spherical waves emanating from a certain position in 3D space. They can be further 'mixed' to add effects such as 'width' and 'diffusion'.

4) Since the $a_n^m(\omega, t)$ representation is oblivious to loudspeaker positions, a renderer [2,9,16,17] is required to convert the coefficients into loudspeaker feeds. Renderers of this type usually take into account the number and positions of loudspeakers that are available and produce an optimum rendering for that environment. Advanced renderers can also account for local acoustical conditions such as room reverberation. This method of not 'tying' the audio to loudspeaker positions, allows the format to be adaptable to practically any loudspeaker layout. It also maintains the spatial resolution required to acoustically focus into spatial regions – allowing the consumer to interact with the soundfield – in a way that is not possible with traditional channels-based audio. This approach cannot be compared to 'up-mixing' or 'downmixing', with channels-based audio, since those processes are essentially trying to maintain the same audio experience with fewer or larger number of speakers. For example, an up-mixed 7.1.4 signal cannot be expected to provide additional 'acoustic' information when played back over 22.1 loudspeaker.

5) A layered approach to delivering audio is also enabled by Scene-based audio. A base layer containing low spatial resolution can be complemented by one or more enhancement layers which add to the spatial resolution. These can be streamed independently and while consumers would always need the base layer, the enhancement layers are completely optional. These allow interesting use-cases for both a broadcaster and a consumer - that is not possible with either object- or channel-based audio.

3 Audio Compression of HOA signals

The advantages listed above make Scene-based audio ideal for transmission. However, what lies in the way, is the sheer bandwidth required. A 4th order HOA signal requiring 25 coefficient channels and a 6th order HOA signal requiring 49 coefficient channels would each need a nominal 40 Mbps and 80 Mbps respectively.

A quick review of past attempts at compressing HOA signals reveal [10, 11] the investigation of audio codecs such as AAC for compressing HOA coefficient signals. Daniel [12] 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). In a technique known as Directional Audio Coding (or DirAC) [14], an energy-based time-frequency analysis is used to estimate the dominant direction and a diffuseness value as a function of frequency bin applied to a 1st-order Ambisonics signal. The transmitted payload consists of these frequency-dependent directional and diffuseness parameters and the onmidirectional 0th-order ambisonics coefficient signal, which can be further compressed with a perceptual audio codec. An extension to DirAC to accommodate HOA was proposed in [13]. In the proposed extension, an N^{th}-order HOA signal is subdivided into $M \geq N$ even-sized spatial sectors, and each of them 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 [13] present a method 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.

The recently ratified MPEG-H standard [1], ratified in 2015, is the result of a competitive process (that saw multiple submissions) that required high quality immersive (or 3D audio perception) audio at bitrates as low as 48 kbps for HOA coefficient signals upto 6th order. As such, the technology provides state-of-

the-art compression for HOA signals allowing broadcast quality (80 MUSHRA points) transmission of full resolution HOA signals at bitrates as low as 300 kbps and transparent quality transmission at 500 kbps (90 MUSHRA points) for all HOA orders (see Figure 6). The layered approach allows a low resolution base-layer as low as 38 Kbps with enhancement layers gradually adding to the spatial resolution. The following paragraphs describe how the MPEG-H technology is able to achieve this amount of compression.

The first stage of MPEG-H encoding for HOA coefficient signals attempts to decompose the signals into 'predominant' (or 'foreground') components and 'ambient' (or 'background') components. This is shown in Figure 3. The decomposition can be interpreted as a way to de-correlate the HOA signal. The 'predominant' or 'foreground' components are signals that are perceptually distinguishable from the ambience and are therefore decorrelated from the rest of the soundfield. The residual soundfield (left-over after the extraction of the predominant sounds) constitute the ambient component. Each pre-dominant component has associated with it a set of data (side-information) that define its spatial characteristics – such as location and width.

One way to conceptualize the decomposition of the predominant signals is using the following equation:

$$H = s_i d_i^T, \qquad (2)$$

where, s_i and d_i are column vectors depicting the i^{th} predominant component and its associated directional characteristics respectively. The d_i vectors carry the directional characteristics decoupled from the temporal characteristics of the predominant signal. The resulting H matrix is the spatio-temporal representation of the predominant signal. The decomposition can be carried out in a frame-by-frame manner at a resolution of approximately 20 ms per frame. Examples of the d_i vector plotted in 3D space is shown in Figure 2.

Various techniques [3] can be used to achieve this decomposition such that maximal decorrelation and energy compaction is achieved. The predominant components, using its associated set of information, is used to recombine with the ambient components to recreate the HOA signal at the decoder.

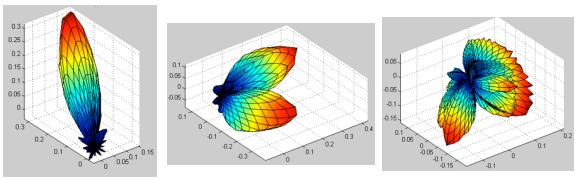

Figure 2 Spatial plots of various d_i vectors.

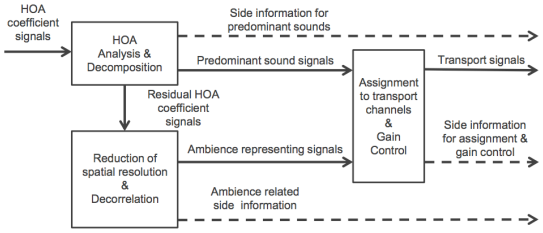

Figure 3. HOA Encoding in MPEG-H

The number of predominant signals ($P(t)$) and ambient signals ($A(t)$) add up to the total number of 'Transport' ($T = P(t)+A(t)$) signals. The residual soundfield may go through a further decorrelation process to represent the ambient part of the soundfield using $A(t)$ number of signals. The numbers $P(t)$ and $A(t)$ can change over time while the total number of transport signals, T, is usually kept constant. This means that channel-assignment information needs to be sent to the decoder. The channel-assignment data, side-information data for the predominant components along with any further information about the ambient components constitute the total side-information – requiring approximately 10 kb/s. T is usually a function of a target bit-rate and is set between 6 and 10. For a sixth-order HOA signal, this represents a dimensionality reduction factor of approximately 8 and 5. For a fourth-order signal the reductions are approximately 4 and 2.5.

The T transport signals are subsequently analyzed for redundancy and irrelevancy according to traditional psychoacoustic source coding concepts [4] to achieve further coding gain. The d_i vectors are quantized using scalar or vector quantization methods – both being allowed in MPEG-H.

At the decoder, the T transport channels are reconstituted from the psychoacoustic decoder before the ambient and pre-dominant components are reconstructed and recombined to reproduce the HOA coefficients. This is shown in Figure 4. The HOA signal can be subjected to loudness and gain control using Dynamic Range Control (DRC) parameters that are either sent with the MPEG-H

bitstream or are known for a particular playback device. MPEG-H provides an interface to the decoder to allow the local loudspeaker layout to be communicated to the HOA renderer which then produces optimal feeds for that specific layout.

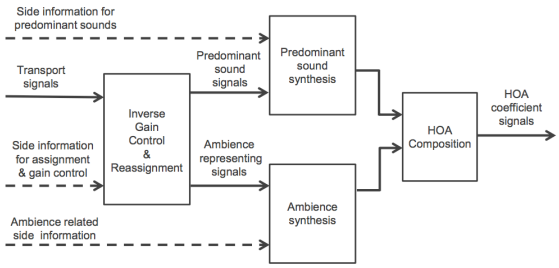

Figure 4. HOA Decoding in MPEG-H

MPEG-H is also able to offer layered-coding for HOA coefficient signals (see Figure 5). 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.

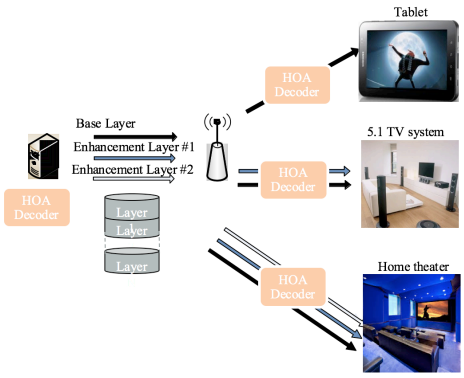

Figure 5. Layered coding using HOA and MPEG-H.

This kind of layered approach is difficult to achieve with either Channels- or Objects-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 Channels-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.

An inherent problem with deploying Scene-Based-Audio in traditional Television plants lies in being able to transport the HOA signals from one part of the plant to the other. Typically, the constraint is due to the use of SD-SDI and HD-SDI routers which limits the number of PCM signals transportable through the TV plant to 8 and 16 respectively. This in turn would limit the HOA order to 1st or 3rd order respectively. The 3rd order HOA signal would take up the entire embedded 16 HD-SDI channel, meaning that other services such as those for the visually impaired or the ability to have a second mix (perhaps stereo) would not be possible. In order to accommodate the ability to send HOA through both SD- and HD-SDI signals, we suggest decoupling the first stage of the MPEG-H encoding (as described above) from the second stage. This allows the representation of any orders of HOA using 7-11 PCM channels (6-10 of PCM channels plus one more for the side-information). This means that an HOA signal can be transported through an SD-SDI framework and still provide one PCM channel for other services. This allows the delivery of complete immersive or 3D audio (played over a multitude of loudspeakers) using one more PCM channel than what it would take to deliver 5.1.

4 Results

The quality of the encoding as a function of bit-rate is shown in Figure 6. The figure represents MUSHRA [21] testing using 20 listeners. There were a total of 12 test items - two 6th order, one 3rd order and nine 4th order items. The hidden reference are the uncompressed items rendered to 22.2 loudspeakers (adhering to the NHK 22.2 layout). The tests were conducted in a BS.1116 compliant room comprising of 28 Genelec 8240A monitor speakers and two 7060B subwoofers.

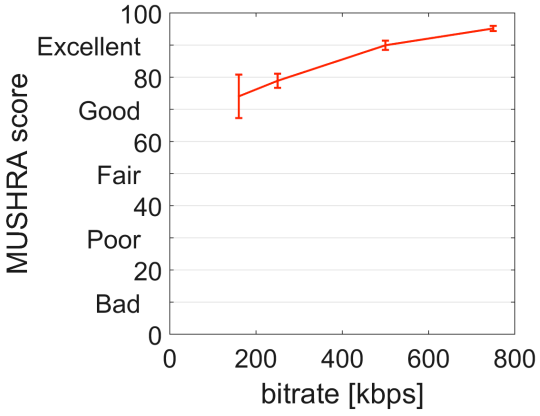

Figure 6. MUSHRA scores for HOA items coded between 96 kbps and 768 kbps.

We have used T=6 for 256 kbps and lower, T=8 for 512 kbps and T=10 for 768 kbps.

5 Discussion

The results from the previous section shows that it is possible to transmit Scene Based Audio signals at 300 kbps for broadcast quality transmission (assuming that a MUSHRA score of 80 meets broadcast requirements). In addition, it should be noted that the results are HOA order independent.

An interesting comparison of the coding efficiency can be made with Channels based audio. Current broadcast requires 384 kbps to transmit 5.1 channels using AC-3 [22]. We can reasonably assume that new generation audio codecs such as AC4 [23] can double the efficiencies previous codecs, such that 11.1 (or 7.1.4) channels would require the same amount of bitrate or 384 kbps. Considering that the Scene Based Audio signals can be rendered to a far higher number of channels (for example the NHK 22.2 layout and even higher) it is seen that a transmission using HOA is far more efficient than transmission using Channels-Based audio especially as the number of intended loudspeakers go up.

Comparisons can also be made with Objects-based audio. A typical cinematic 3D audio production involves many simultaneous audio elements along with the metadata [5,6] for each element (such as the position of the element). Figure 7 shows the number of simultaneous audio elements in an approximately two-minute-long cinematic content. The maximum number reaches 60, while the average number is 33. The raw, uncompressed bitrate would average 38 mbps (at a sampling rate of 48 kHz and bit depth of 24 bits/sample). This kind of bitrates are only possible over physical media such as Blu-ray discs where some amount of lossless coding could bring this rate down by a factor of 2 or 3. Significant bandwidth savings could be had if the number of audio elements were to be grouped (in a regional proximity manner, for example). It's not unreasonable to assume that such a grouping could result in an average of 16 audio elements being present simultaneously. A lossy compression of these elements would result in an average bitrate of 800 kbps (assuming 50 kbps/element) without even considering the bitrate for the metadata that needs to be transmitted along with the elements. Such a comparison shows the benefits of using Scene Based Audio for transmitting 3D immersive audio. Considerations of complexity of rendering each audio element according to its metadata (diffusion) on end-user devices such as smart-phones provide further compelling reasons on why the only practical method of transmitting Object-based audio is through a Channel-bed (such as 7.1) and a few audio elements which add to the immersive effect by adding height perception.

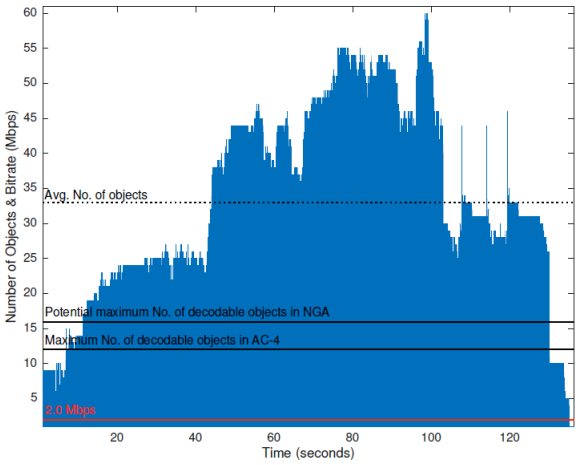

Figure 7. Number of simultaneous objects present as a function of time for some cinematic content.

An additional benefit of Scene-based audio is the ease in which the soundfield can be rotated given the fact that representation is on Spherical Harmonic basis functions. This ease of rotation, the ability to create a binaural feed (with the aid of HRTFs and rendering to multiple virtual speakers), the inherently high spatial resolution of the representation (roughly equivalent to greater than $(N+1)^2$ uniformly spaced loudspeakers on a sphere, where N is the order of the HOA signal) makes the format ideal for Virtual Reality type experience – where the soundfield has to be rotated to adapt to the head rotations of the listener.

6 Summary

We have described the MPEG-H encoding technology for Scene Based Audio (HOA) signals. The standard provides the state-of-the-art solution for 3d-audio via traditional broadcast network infrastructure using HOA. Further, a spatial encoding technology provides the means to transport the HOA signals through television plants for both SD-SDI and HD-SDI based frameworks.

7 References

[1] ISO/IEC: "Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D audio," Doc. 23008-3:2015, International Standards Organization / International Electrotechnical Commission, Geneva Switzerland.

[2] Poletti, M.A., "Three-Dimensional Surround Sound Systems Based on Spherical Harmonics," JAES, 53 (11), pp 1004-1025, Nov 2005.

[3] Goyal, V.K., "Theoretical foundations of transform coding," IEEE Signal Processing Magazine, vol. 18, no. 5, pp. 9-21, Sept. 2001.

[4] Painter, T. and Spanias, A., "Perceptual coding of digital audio," Proc. IEEE, 80(4): pp 451–513, Apr. 2000.

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

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

[7] Frank, M., Zotter, F., and Sontacchi, A., "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.

[8] Daniel, J., "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.

[9] Rafaely, B., "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *Journal of the Acoustical Society of America*, 116 (4), pp. 2149 - 2157, 2004.

[10] Hellerud, E., Solvang, A., and Svensson, U.P, "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.

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

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

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

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

[15] 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.

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

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

[18] Begault, D.R., Wenzel, E.M., and Anderson, M.R., "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.

[19] Noisternig, M., 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.

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

[21] ITU-R Recommendation BS.1534-2, "MUlti-Stimulus test with Hidden Reference and

Anchor (MUSHRA)".

[22] ATSC Standard: A/52:2010: Digital Audio Compression (AC-3) (E-AC-3) Standard.

[23] ETSI: "Digital Audio Compression (AC-4) Standard; Part 1: Channel based coding," Doc. TS 103 190-1 V1.2.1 (2015-06), ETSI, Sophia Antipolis Cedex, France.