Audio Engineering Society

Convention Paper 7366

Presented at the 124th Convention
2008 May 17–20 Amsterdam, The Netherlands



# Encoding Higher Order Ambisonics with AAC

Erik Hellerud[1], Ian Burnett[2], Audun Solvang[1] and U. Peter Svensson[1]

[1] *Centre for Quantifiable Quality of Service in Communication Systems*, Norwegian University of Science and Technology, Trondheim, Norway*

[2] *University of Wollongong, Australia*

Correspondence should be addressed to Erik Hellerud (`erih@q2s.ntnu.no`)

**ABSTRACT**

In this work we explore a simple method for reducing the bit rate needed for transmitting and storing Higher Order Ambisonics (HOA). The HOA B-format signals are simply encoded using Advanced Audio Coding (AAC) as if they were individual mono signals. Wave field simulations show that by allocating more bits to the lower order signals than the higher the resulting error is very low in the sweet spot, but increases as function of distance from the center. Encoding the higher order signals with a low bit rate does not lead to a reduced audio quality. The spatial information is improved when higher-order channels are included, even if these are encoded with a low bit rate.

## 1. INTRODUCTION

Higher Order Ambisonics (HOA) is a technique for reproducing a complete soundfield, either a complete three-dimensional representation or in just two dimensions; in this work the latter is considered. The extension of the area over which an accurate representation is achieved is proportional to the order $N$ [1]. For a two-dimensional representation and an order of $N$, $2N + 1$ channels are necessary. If regular CD-quality is used, with a bit depth of 16 and 44.1 KHz sampling rate, the total rate becomes $(2N + 1) * 16 * 44.1$ kbps. For an order of 7, which is the highest order used in this work, this would mean a rate of more than 10 Mbps.

10 Mbps is such a high rate that both storage and network transmission can be problematic, but, as with regular compression, audio signals contain significant redundancy which can be removed without sacrificing perceptual quality.

*Centre for Quantifiable Quality of Service in Communication Systems, Centre of Excellence appointed by The Research Council of Norway, funded by the Research Council, NTNU and UNINETT. http://www.q2s.ntnu.no/

The authors have not found prior studies that specifically look at the compression of HOA, but several techniques for compressing other multichannel audio formats exist. The last decade has seen the development of several new codecs. For instance, the recently standardized MPEG Surround [2] can give remarkably good quality for some test items [3] for bit rates as low as 64 kbps (for a traditional ITU 5.1 layout). However, it is found in [3] that a bit rate of 448 kbps is needed for the more sensitive test items regardless of the chosen codec. MPEG Surround works by downmixing the signal in the encoder to stereo and encoding the spatial information using parametric cues. One advantage of this scheme is that the encoded format is stereo compatible. The 5.1 format can be called a sweet-spot technique, that is, the format does not generate an extended sound field. Therefore, the spatial distribution of the quantization error is more or less irrelevant. For the HOA format, on the other hand, the spatial distribution of the quantization error is highly relevant, and that is the scope of this paper. A more theoretical approach analyzing the effects of quantizing Higher Order Ambisonics signals is given in a companion paper [4].

Higher Order Ambisonics is described briefly in section 2, and AAC is presented in section 3. In section 4 numerical results from encoding both the Ambisonics B- and D-format are given, and also some comments from informal listening. Conclusions and suggestions for further work are offered in sections 5 and 6.

## 2. HIGHER ORDER AMBISONICS

Here, only a short introduction to HOA will be given since the complete theory is thoroughly presented in [5, 6]. The theory behind Ambisonics was developed in the 1970s, and although it has not gained any significant commercial interest it is still one of the few methods for reproducing complete sound fields. The other significant alternative is Wave Field Synthesis (WFS) [7]. A horizontal sound field can be expressed in terms of its cylindrical harmonics decomposition [8]:

$$p(r,\theta) = B_{00}^{+1} J_0(kr) + \sum_{m=1}^{\infty} J_m(kr) B_{mm}^{+1} \sqrt{2} cos(m\theta) + \sum_{m=1}^{\infty} J_m(kr) B_{mm}^{-1} \sqrt{2} sin(m\theta), \quad (1)$$

where $J_n$ is the n'th order Bessel function and $k$ is the wave number ($k = \frac{2\pi}{\lambda} = \frac{\omega}{c}$). The coefficients $B_{mm}^{\pm 1}$ are the so-called B-format signals in Ambisonics. As seen from eq. 1, these coefficients describe the sound field for all angles and radii.

For practical use, the infinite sums in eq. 1 must be truncated to a maximum order $N$. Then, the B-format coefficients form the HOA representation of order $N$. The B-format coefficients can be found either by encoding each virtual source's signal individually [5] or by using a multi-element microphone that extracts the B-format signals by processing the microphone signals [9].

To derive the loudspeaker signals (the D-format) from the B-format, a simple matrix multiplication is used [5]. The parameters for this decoding matrix is given by the order and loudspeaker locations. For regular loudspeaker layouts the decoding matrix is given as

$$D = \frac{\sqrt{2}}{N} \begin{bmatrix} \frac{1}{\sqrt{2}} & \dots & \frac{1}{\sqrt{2}} \\ cos(\phi_1) & \dots & cos(\phi_M) \\ sin(\phi_1) & \dots & sin(\phi_M) \\ cos(2\phi_1) & \dots & cos(2\phi_M) \\ sin(2\phi_1) & \dots & sin(2\phi_M) \\ \dots & \dots & \dots \\ cos(K\phi_1) & \dots & cos(K\phi_M) \\ sin(K\phi_1) & \dots & sin(K\phi_M) \end{bmatrix}^T, \quad (2)$$

where $K = 2N+1$, $M$ is the number of loudspeakers, and $\phi_i$ is the angle of the $i$'th loudspeaker.

For an HOA encoding/decoding of order $N$ and reproduced over $2N+1$ loudspeaker in a circular, regular array, the resulting wave field error stays below -15 dB [10] as long as the relationship

$$N = kr \quad (3)$$

is fulfilled, where $r$ is the radius of the reproduction area.

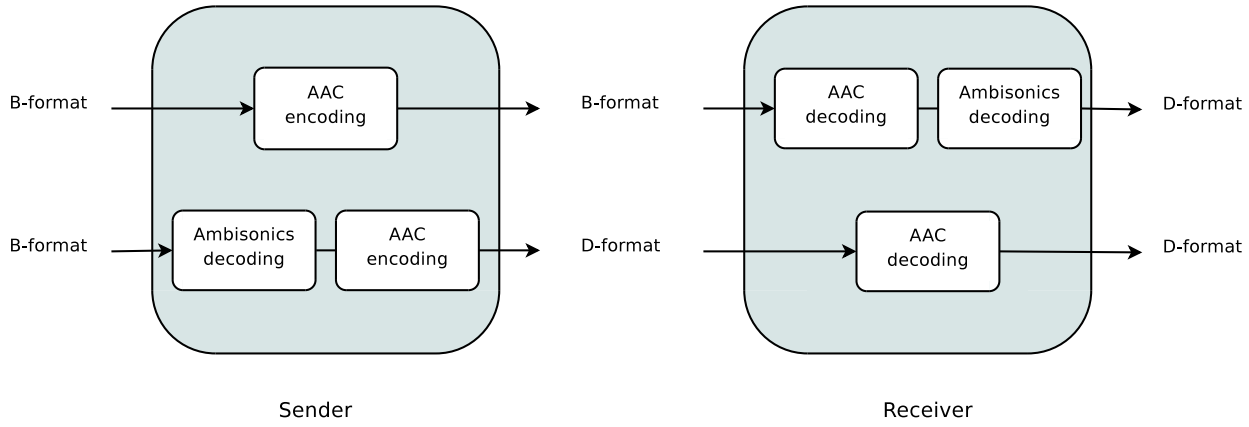



**Fig. 1:** Encoding schemes: Upper signal path shows encoding the B-format signals, lower shows encoding the D-format signals (loudspeaker signals).

One very interesting feature of HOA is the flexibility of the format. Given a representation of order $N$ with $2N + 1$ channels, a subset using channels 1 to $2M + 1$ ($M \leq N$) can be used to decode the source to an order $M$ representation. Decoding only a subset of the channels will only affect the spatial resolution. This makes the HOA format ideal for network transmission since the number of channels transmitted can be adapted to the receiver's setup, and the transmission rate can also be adapted to the available network bandwidth. It should also be mentioned that this format is very suitable for future network architectures such as Differentiated Services (DiffServ) [11]. With DiffServ there are several priority levels in the network, so the most important data can be transmitted at a high priority level, with a significant lower probability for data loss than in the current Best Effort network architecture. Using the DiffServ architecture it would be natural to transmit the lower order components with a high priority, while the higher orders could be transmitted using regular priority, thus increasing the probability that at least a lower order representation is received.

In addition to the scalability and layered structure of the format, it is also very flexible from the reproduction perspective. A signal in the HOA B-format can can be decoded to an arbitrary loudspeaker configuration, including the common 5.1 and 7.1 configurations [12]. In the encoding approach presented here, the scalability is not removed from the format. Also, due to the low bit rates of the higher order channels, a relatively fine granularity is achievable.

One solution for compressing HOA is to decode the B-format to loudspeaker signals (D-format) and encode each individual signal (Figure 1, lower path). This approach leads to a uniform error across the listening area. If the loudspeaker signals' amplitudes differ much, e.g., as caused by one dominant source direction, fewer bits could be assigned to the weaker-amplitude channels [13]. However, such bit distributions might lead to unwanted spatial distribution effects, so that distortions that are masked at the central listening position get unmasked in non-central listening positions.

Another disadvantage with this approach is that some of the desired features of the B-format are no longer available. Encoding the D-format signals means that the receiver has to use a fixed loudspeaker setup, it is not longer possible to use an arbitrary loudspeaker configuration. Also, the scalability has been removed; the sender has to transmit all channels, making this encoding scheme less suited for network transmission.

Due to the reasons presented above, encoding the B-format signals (Figure 1, upper path) seems like a more reasonable solution.

## 3. ADVANCED AUDIO CODING

MPEG-4 Advanced Audio Coding (AAC) [14] is currently one of the best stereo audio encoders. Transparent quality can result from bit rates as low as 64 kbps for a stereo signal [15].

An Advanced Audio Coding (AAC) encoder splits the signal into frames of 2048 samples (or 256 samples if transients are detected) overlapping with 50%, and transforms each frame into the frequency domain using the Modified Discrete Cosine Transform (MDCT). From a psychoacoustic analysis the quantization threshold for each subband is selected and finally, the resulting coefficients are entropy coded.

For very low bit rates it has been shown that it can be beneficial to represent the high frequency content in a parametric way derived from the low frequency content. This technique is called Spectral Band Replication (SBR) [16], and this is used in the encoder selected for this work [17]. SBR is a part of MPEG-4 High Efficiency AAC (HE AAC). For very low bit rates in standard AAC, it is unavoidable that the noise will be above the masked threshold if the whole frequency range is encoded. However, there will be a significant quality reduction if the signal is simply low-pass filtered. By using SBR the high frequency range in the signal is maintained, but it is encoded using only a few bits. By utilizing the correlation between the low and high frequencies, an estimate of the high frequency content can be given from the transmitted low frequency content. By using SBR the quantization noise will ideally be below the masked threshold for the lower frequency range. This has been shown to increase the perceived quality significantly when very low bit rates are used.

## 4. ENCODING HOA SIGNALS

The technique used in this paper is very simple; each B-format channel is encoded independently using AAC. One advantage of the scheme is that both the scalability and flexibility of the format is intact, and it is also easy to use varying bit rates for the different channels/orders. Using a lower bit rate for the higher order components will be shown to be an essential technique for maintaining the perceived quality as well as the spatial resolution, even with a very low total bit rate.

Encoding channels will lead to distortions, so several configurations have been tested in this work to minimize the distortions. One difference from regular stereo is that in addition to good sound quality, it is also desirable that the compression does not introduce spatial distortion, meaning that sources are perceived to originate from a different direction than in the original clip, or that the direction are perceived as less distinct.

Given a total bit budget there are numerous options for how the bits could be distributed between channels. The most obvious solution is to use the same bit rate for all channels. A different option is to vary the bit rate between channels, either using a lower bit rate for the higher or for the lower order components. Also, it is useful to consider whether a low order representation consisting of channels with a high bit rate is preferable over a higher order encoding at lower bit rates. Reducing the Ambisonics order will reduce the spatial resolution, but if the gain in sound quality is significant it may be worth it.

To analyze the error in the reproduction area, the HOA signals were decoded to loudspeaker signals (D-format), and the sound pressure calculated for loudspeakers radiating plane waves:

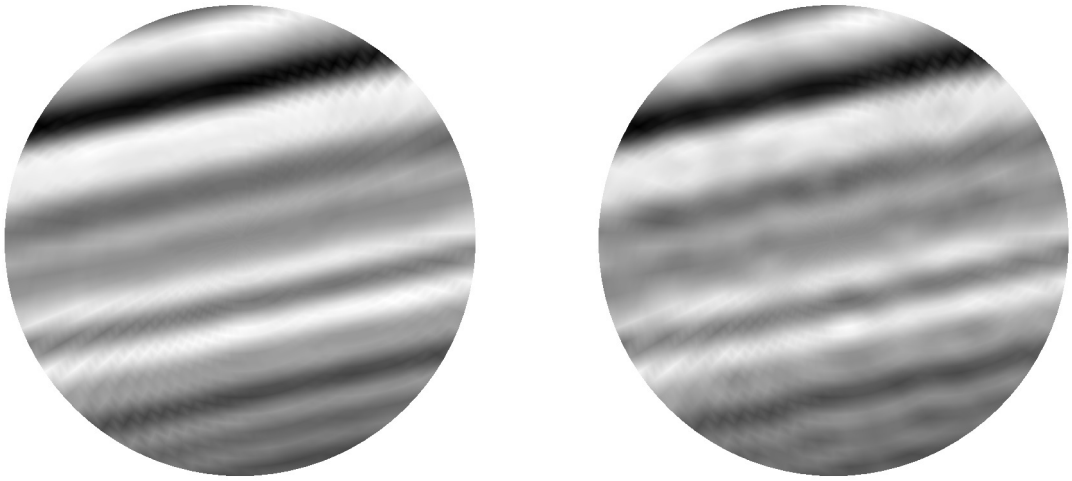
$$p(r,k,\theta) = \sum_{m=1}^{M} e^{jkr\cos(\theta_m-\theta)} l_m(0,k,\theta_m), \qquad (4)$$

where $l_m$ is the signal from the $m$'th loudspeaker, which is placed in the angle $\theta_m$.

The sound fields resulting from the original and encoded clips were then compared. The error, $\epsilon$, is defined here as

$$\epsilon(r,\theta,t) = 10 * log_{10} \frac{(p_c(r,\theta,t) - p_r(r,\theta,t))^2}{p_r(r,\theta,t)^2}, \qquad (5)$$

where $p_r$ is the reference sound field, and $p_c$ is the sound field resulting from the encoded signals. The error is calculated time-sample by time-sample, and averaged over time. It should be noted that this error measure does not take any perceptual aspects into account. To find the average at a given radius the error was averaged across all angles. Furthermore, the error can also be averaged across the entire reproduction area, i.e., for radii up to the edge of the reproduction circle.



(a) Original (HOA order 7). (b) Channels 12-15 compressed.

**Fig. 2:** Wave field snapshot for a signal with only one source and reproduced with order 7. The compressed channels are encoded to 64 kbps. The radius of the circles is 14 cm.

Surprisingly, compressing HOA with AAC seems to work remarkably well. To analyze the effects from compression, both wave field analysis and casual subjective evaluation were used.

### 4.1. Wave field analysis

A wave field analysis was performed by comparing an original soundfield with a processed soundfield over a reproduction area (a circle of radius 14 cm). The original soundfield consists of either a single virtual source, or four virtual sources (spread to four different source angles), all positioned at infinite distance, and with no room reflections or reverberation included. This corresponds to the HOA processing of an extremely dry mono-microphone recording which is arguably the most critical case for detecting the direction of a source. The processing includes HOA encoding to a reproduction order $N$ of each source signal at the desired virtual source angle, AAC encoding/decoding the $2N+1$ HOA B-format signals, and applying a basic HOA decoding as given by eq. 2 for a regular, circular array of $2N+1$ loudspeakers at infinite distance. The HOA decoding yields the D-format signals, i.e., the loudspeaker signals, $l_m$. The loudspeaker signals were transformed using a DFT and eq. 4 was applied one frequency at a time, and an inverse DFT gave the time-domain signal. A final wave field snapshot was then generated by plotting the instantaneous wave field across the reproduction area.

One such example is shown in figure 2 where a single virtual source, emitting a dry drum beat recording, was HOA encoded to order 7, and reproduced over 15 loudspeakers. The last four B-format channels were compressed with AAC to 64 kbps. From the wave field analysis it can be seen that the difference between the original and the compressed audio is

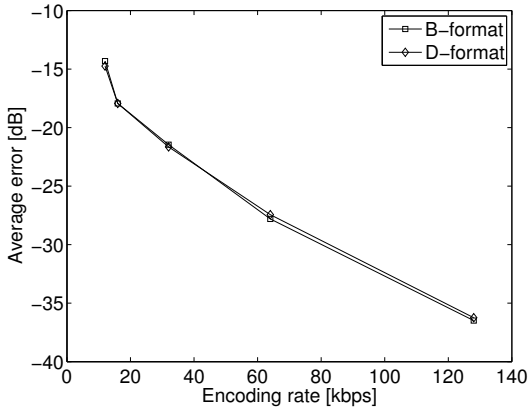



**Fig. 3:** Error averaged across the entire reproduction area. The HOA order was 7, and all channels were encoded with the same rate.

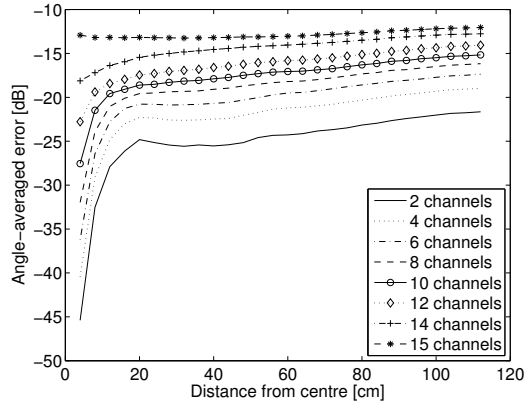



**Fig. 4:** Angle-averaged error as a function of distance from the centre for a clip containing four sources. The HOA order was 7, and a subset of the B-format signals were encoded to 12 kbps. "2 channels" means that channels 14-15 were compressed to 12 kbps, "4 channels" means that 12-15 were compressed and so on.

that the wave becomes more blurry, meaning that the wave loses some of its distinct contours. This can be seen in figure 2b.

### 4.2. Error as a function of encoding rate

Figure 3 shows the average error in the listening area as a function of encoding rate for both encoding of the B-format and D-format signals. The clip used here has four virtual sources in four different locations. As expected this results in an approximately linear decrease, but it is interesting to see that the distortion is more or less equal whether it is the B- or D-format signals that have been encoded. It should be noted that HOA reproduction of a broadband signal over a circular reproduction area will introduce wave field distortions at high frequencies, which makes an averaging of the error across frequencies problematic. However, here we use the HOA encoded/decoded wave field as reference, and consequently, our error is only the one that is introduced by the AAC compression.

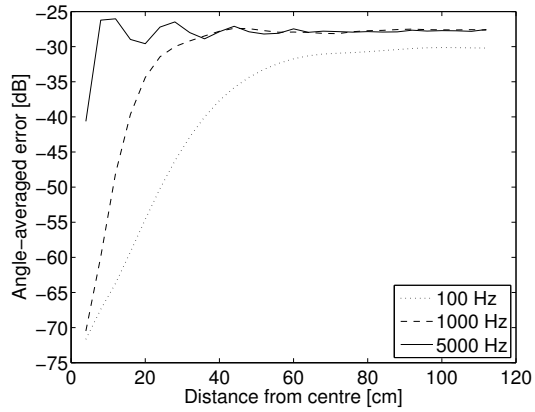



**Fig. 5:** Angle-averaged quantization error as a function of distance and frequency. A single sinusoidal with frequency 100, 1000, or 5000 Hz were reproduced with order 7.

Also, note the large difference between 16 and 12 kbps in figure 3. SBR is used for the lowest bit rate, and the use of SBR results in large sound field distortions since the high frequency content is only estimated, but perceptually the difference is not as big as it appears from the figure.

### 4.3. Error as function of distance

One important aspect of HOA is that the channels affect the listening area differently. The W component ($B_{00}$) affects the whole listening area, while the higher orders mostly affect the area further away from the centre of the listening area. This is illustrated in figure 4. This graph is generated from a wave field analysis resulting from a complex music clip with 4 sources from different directions reproduced with order 7.

As seen from the figure the resulting error in the centre is very low when only a few of the higher orders are compressed. If all channels are compressed to the same bit rate, the resulting error is uniform across the whole listening area, as indicated by the curve "15 channels" in figure 4.

The radial extent of the area with the lowest error depends on the frequency. In figure 5, this frequency dependence is illustrated by evaluating the angle-averaged error for single-frequency wave field of frequency 100, 1000 or 5000 Hz. In this case, the signal was transformed using the MDCT and quantized. As can be seen in figure 5, the higher the frequency is, the smaller is the area with a lower error.

The perfect reconstruction radius for a frequency of 1000 Hz and order 7 is 38 cm (Equation 3), and from figure 5 it can be seen that for the 1000 Hz sinusoidal the maximum distortion is achieved at approximately that distance.

### 4.4. Perceived quality

To evaluate the perceived quality casual subjective evaluation has been used in this initial study. Preliminary results indicate that higher order components do not affect the perceived audio quality, even if they are compressed to quite low bit rates. Using 12 kbps for the highest orders does not reduce the quality significantly, even though the individual channels have a very reduced sound quality. However, the spatial resolution is significantly improved when these low bit rate channels are included, compared with a lower order representation.

To evaluate this encoding scheme several clips were used, ranging from simple clips with a single source in a single direction to more complex clips with several sources in multiple locations. The setup consisted of 15 loudspeakers in a uniform layout, so the highest order possible for playback was 7. Several configurations of bit allocations between the channels were tested, and the most promising solutions seem to be to use a high bit rate for the W component, and reducing the bit rate for the channels as the order increases. Total bit rates as low as 256 kbps were tested, meaning a compression ratio of more than 41. Even at this bit rate, the sound quality was still very good, but some sound sources were moved away from their original location. By increasing the bit rate slightly, to e.g. 384 kbps, no spatial distortion was audible in casual evaluation.

Another effect of using very low bit rates on all B-format channels is that the AAC encoder may have to remove parts of the signal that are actually audible in order to reach the target rate.

Encoding the B-format channels with AAC is clearly not an optimal solution. The perceptual model is not matched against the reproduction, meaning that parts of the signal that the model determines audible may actually be inaudible due to spatial masking. Also, working on a single channel at a time makes it impossible to utilize the correlation between channels. For signals with only one source the channels are highly correlated, but even for the more complex clips there can be a very high correlation between some of the channels. This means that the bit rate could be reduced further if a more complex encoding approach was used.

## 5. CONCLUSION

From the presented results it can be seen that reasonably good sound quality can be achieved by encoding the Ambisonics B-format with AAC.

It was found that using more bits for the lower order signals resulted in an increasing error as a non-uniform function of distance from the centre. Also, it was found that the actual sound quality of the higher order components is not that important; even at a significantly reduced perceptual signal quality, these contribute significantly to the perceived spatial resolution.

## 6. FURTHER WORK

This work should be followed up with a more thorough subjective test to evaluate the performance of this encoding scheme. Also, it should be investigated how one could utilize the correlation between the channels to further reduce the bit rate.

## 7. REFERENCES

[1] B. Stofringsdal and U. P. Svensson, "Conversion of Discretely Sampled Sound Field Data to Auralization Formats," *J. Audio Eng. Soc*, vol. 54, no. 5, pp. 380–400, May 2006.

[2] S. Quackenbush and J. Herre, "MPEG Surround," *Multimedia, IEEE*, vol. 12, no. 4, pp. 18–23, Oct.-Dec. 2005.

[3] A. Mason, D. Marston, F. Kozamernikm, and G. Stoll, "EBU Tests of Multi-channel Audio Codecs," in *The 122nd AES Conv.*, 2007, Preprint 7052.

[4] A. Solvang, U. P. Svensson, and E. Hellerud, "Quantization of Higher Order Ambisoncs wave fields," in *The 124th AES Conv.*, 2008.

[5] J. Daniel, S. Moreau, and R. Nicol, "Further Investigations of High-Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging," in *The 114th AES Conv.*, February 2003, Preprint 5788.

[6] M. A. Poletti, "A Unified Theory of Horizontal Holographic Sound Systems," *J. Audio Eng. Soc*, vol. 48, no. 12, pp. 1155–1182, December 2000.

[7] M. M. Boone and E. N. G. Verheijen, "Multichannel Sound Reproduction Based on Wavefield Synthesis," in *The 95th AES Conv.*, October 1993, Preprint 3719.

[8] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, Academic Press, 1999.

[9] J. Meyer and G. Elko, "A Highly Scalable Spherical Microphone Array Based on an Orthonormal Decomposition of the Soundfield," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.

[10] D. B. Ward and T. D. Abhayapala, "Reproduction of a Plane-wave Sound Field Using an Array of Loudspeakers," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 6, pp. 697–707, Sep 2001.

[11] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services," RFC 2475, December 1998.

[12] M. Neukom, "Decoding Second Order Ambisonics to 5.1 Surround Systems," in *The 121st AES Conv.*, October 2006, Preprint 6980.

[13] A. Solvang and U. P. Svensson, "Removal of Spatial Irrelevancy in 3D Audio Utilizing Ambisonics and the Continuity Illusion," in *Proceedings of Norsk Symposium i Signalbehandling (NORSIG-05)*, Sep. 2005.

[14] ISO/IEC 14496-3, "Coding of audio-visual objects – Part 3: Audio," 1998.

[15] ISO/IEC JTC/SC29/WG11, "Report on the MPEG-2 AAC Stereo Verification Tests," MPEG1998/N2006, San Jose, USA, February 1998.

[16] M. Dietz, L. Liljeryd, K. Kjorling, and O. Kunz, "Spectral Band Replication, a novel approach in audio coding," in *The 112th AES Conv.*, 2002, Preprint 5553.

[17] Nero AAC Codec, "http://www.nero.com/," [Online].