SPATIAL REDUNDANCY IN HIGHER ORDER AMBISONICS AND ITS USE FOR LOW DELAY LOSSLESS COMPRESSION

Erik Hellerud, Audun Solvang, and U. Peter Svensson

Centre for Quantifiable Quality of Service in Communication Systems
Norwegian University of Science and Technology
Trondheim, Norway
{erih, auduso, svensson}@q2s.ntnu.no

ABSTRACT

When Higher Order Ambisonics (HOA) is used to represent a sound field, the channels might contain a lot of redundancy in some cases. This redundancy can be exploited in order to provide more efficient network transmission and storage. In this work the amount of inter-channel redundancy for Higher Order Ambisonics is investigated. Furthermore, lossless compression techniques that build on this redundancy are studied, with a focus on low-delay algorithms for real-time, or two-way, applications. The presented encoding scheme results in a delay of 256 samples, but with a rather high computational complexity both for encoding and decoding. The system also preserves the desired features of the HOA format, such as the scalability and the ability to reproduce over arbitrary loudspeaker layouts.

Index Terms— audio coding, ambisonics, lossless coding, IP networks

1. INTRODUCTION

Currently, only two formats exists for reproducing complete wave fields, Wave Field Synthesis (WFS) [1] and Higher Order Ambisonics (HOA) [2], but both formats have very limited commercial use. Recent developments for microphones may change this. For instance, the "Eigenmike" presented in [3], is a microphone with 32 elements recording Ambisonics up to the fourth order in the spherical harmonics transform domain. Such a microphone can be a possible component in systems for broadcasting musical performances, but it can also become a key component in e.g. video conferencing systems in order to provide a better spatial separation between the participants.

A recording with such a microphone results in up to 25 channels in the transformed spherical harmonics domain, and if we consider regular CD quality audio, with 16 bit per sample and a sampling rate of 44.1 kHz, the result is a bit rate of 17.6 Mbps. This data rate is so high that transmission, but also storage, is a difficult task. For musical performances and also videoconferencing situations, it may be desirable to keep the full quality of the recording, meaning that lossy coding must be avoided. Lossless coding will of course result in a much higher bit rate than lossy coding, but it can still reduce the bit rate significantly. Therefore, our focus in this work is on lossless coding. Another important factor for the end-to-end quality is the total delay. For regular audio coders, the encoding delay can be quite high. In this paper a backward-adaptive prediction algorithm is used to reduce the encoding delay to 256 samples.

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://q2s.ntnu.no/

With HOA the inter-channel correlation can in many situations be very high, and by exploiting this correlation and removing the redundancy the resulting compression efficiency may be much higher than for e.g. regular stereo coding. In the extreme situation with only one source and no reverberation, the difference between all channels will simply be a multiplication factor. With more sources this will of course change, but the correlation between channels can still be quite high.

Ambisonics is also a format that is well suited for network transmission. The HOA-format is naturally scalable, meaning that the transmission rate can be adapted to the available network bandwidth. By using the newer Differentiated Services (DiffServ) [4] network architecture it is also possible to transmit the channels at different priority levels. With this approach the probability that the most important channels are received can be increased or even guaranteed. Another interesting feature of the HOA-format is that it is independent of the loudspeaker layout. The receiver can have an arbitrary number of speakers and positions, the decoder only has to be matched to the current loudspeaker configuration. In order to preserve these properties it is necessary that the codec supports this type of hierarchical data.

Meridian Lossless Packing (MLP) [5] is also a codec supporting hierarchical audio such as HOA, however this codec does not focus on low delay, but rather to normalize the peak data rate in order to play back from media such as DVD-A.

2. HIGHER ORDER AMBISONICS

Higher Order Ambisonics (HOA) is an extension to regular Ambisonics, which was originally developed during the 1970s. For a three dimensional reproduction this system is based on decomposing the wavefield into spherical harmonics. A HOA reproduction of order N will require $(N+1)^2$ channels. A complete derivation for 3D reproduction can be found in [2].

The system is somewhat simpler for a 2-dimensional reproduction. Here, the wavefield is decomposed into cylindrical harmonics. The pressure at a point in the plane with radius r and angle θ can be expressed as

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

where J_n is the n-th order Bessel function and k is the wave number defined as $k = \frac{2\pi}{\lambda}$. By truncating these sums to N, we have a

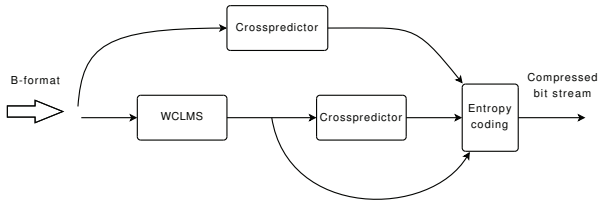


Fig. 1. Overview showing the three different signal paths to the entropy coder.

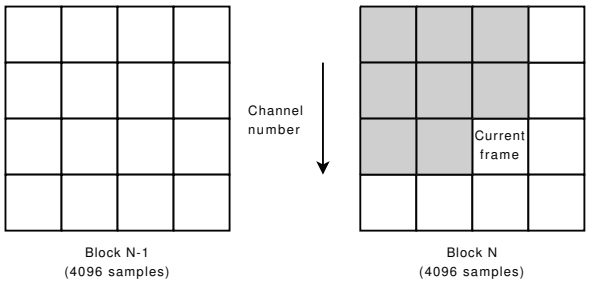


Fig. 2. The possible reference frames for the current coding frame are shown in gray. A low number of channels and frames are used for simplicity.

representation of order N. For this frequency-domain formulation, the coefficients $B_{mm}^{\pm 1}$ form the HOA channels (B-format). These coefficients can be derived from a microphone array using processing of the microphone signals [3], or from synthesis/panning, where a single-source recording is positioned at the desired angle using HOA encoding [2]. As seen from the formula this will require $2N + 1$ channels. For a signal with order N, near-perfect reproduction can result inside a circle with radius $r = \frac{N}{k}$, i.e. the radius is frequency-dependent.

The loudspeaker signals are found by multiplying the B-format signals with a decoding matrix. The parameters for the matrix depend only on the loudspeaker locations and the order.

3. LOSSLESS COMPRESSION WITH LOW DELAY

Lossless compression is usually based on prediction of the time domain signal [6], and entropy coding of the prediction residual. The correlation between channels is usually very low for stereo signals, but a cross predictor is included in the MPEG-4 ALS codec [7] in order to exploit this correlation when it is present. For HOA the correlation between channels can potentially be very large, and this can be used to decrease the bit rate significantly more than for regular stereo signals.

In this work the constraint that the system should have very low delay is added, which may reduce the coding gain compared to more unrestricted implementations. The importance of keeping the delay low can be seen in [8], where it is found that the delay should be kept below 40-60 ms for musical performances. Voice applications have more relaxed requirements, it is recommended that the end-to-end delay should be less than 150 ms in [9]. The implementation presented in this work results in an algorithmic delay of 256 samples, but at the expense of high encoding and decoding complexity. It should also be stated that the delay of the presented codec can easily be lowered, but we have found that 256 samples is a reasonable trade-off between packet transmission efficiency and error propagation.

In a basic mode, when this codec does not exploit the correlation between channels, a Weighted Cascaded Least Mean Squares (WCLMS) [10] predictor is used. This predictor uses filters of decreasing order to produce three estimates of the current sample. The first estimate is from a regular LMS predictor, and this prediction error is fed to the second predictor in order to produce the second estimate. The third predictor is given the prediction error from the second predictor as input. The final estimate is a weighting of the three estimates, and the weighting is determined by evaluating the accuracy of the 3 predictors for the previous samples. By cascading multiple predictors different filter lengths can be used, and the prediction estimate is improved compared with only one predictor. In this work the filter lengths have been 80, 30 and 5.

The advantage of using adaptive prediction compared to the more traditional Linear Predictive Coding (LPC), is that the filter coefficients do not have to be transmitted to the receiver, the disadvantage is that the prediction result can be slightly worse than compared to regular LPC. However, for low delay audio coding regular LPC can be problematic. In order to keep the delay low the frame sizes must be kept short, and since it is necessary to transmit filter coefficients for each frame, the data rate for transmitting the filter coefficients may become significant.

In order to avoid error propagation when data is lost, the prediction algorithm is reset for every 4096 samples, but data is transmitted for each frame (256 samples).

3.1. Exploiting inter-channel redundancy

To exploit the correlation between the B-format channels, this codec investigates two possibilities. The first option is to look directly at the input signals (B-format), and to estimate the coding channel (x^c) from the reference channel (x^r) using a simple three-tap filter (figure 1, upper signal path)

$$\tilde{x}^c(n) = x^c(n) - \sum_{i=-1}^{1} \gamma_i \cdot x^r(n-i), \tag{2}$$

where the prediction parameters (γ_i) are found using the procedure in [7].

The reason for having this option available is that a channel can be a scaled version of the reference channel, and this system will predict such channels easily. To find the reference channel, the correlation between x_c and the possible reference channels are calculated, and the highest correlated channel is chosen as the reference.

The reference frame is restricted to be a previously processed frame from a lower or the same order as the current coding frame. However, the reference frame has to be a part of the same block as the coding block of 4096 samples in order to avoid error propagation in situations with packet loss. The possible reference frames are shown in figure 2. In practice the best reference frame is usually one of the frames of lower order, but from the same time slot.

This also means that a frame in the first channel (denoted W) could be predicted from earlier frames within the same block, however this has very rarely been found to be successful, thus W is always encoded independently.

The reason for restricting the possible reference frames is to preserve the scalability of the B-format. Higher order channels are only useful if all the lower order channels have been received. With this approach each channel will only have dependencies to channels with lower order, and higher order channels can be omitted without affecting the lower order channels if this is necessary.

The other option for crossprediction is to use the prediction residual ($e(n)$) from the WCLMS predictor,

$$\tilde{e}^c(n) = e^c(n) - \sum_{i=-1}^{1} \gamma_i \cdot e^r(n-i). \quad (3)$$

This setup will work better in those situations where the correlation is lower than compared to the first situation. The same system for selecting the reference frame as for the first case is used here.

If there is no coding gain from using the reference channel, the encoder uses the signal directly from the WCLMS, thus no crossprediction is used. Unlike the adaptive filter coefficients, the crossprediction coefficients must be transmitted to the receiver, in addition to a reference to which frame that has been used as a reference.

3.2. Entropy Coding

For lossless coding Golomb-Rice coding has been shown to be near optimal [6] since the residual after linear prediction is usually close to Laplace distributed. With Golomb-Rice each integer is represented with 1 sign bit, m bits indicating the lower order bits, and the remaining higher order bits are encoded unary. The parameter m is found as:

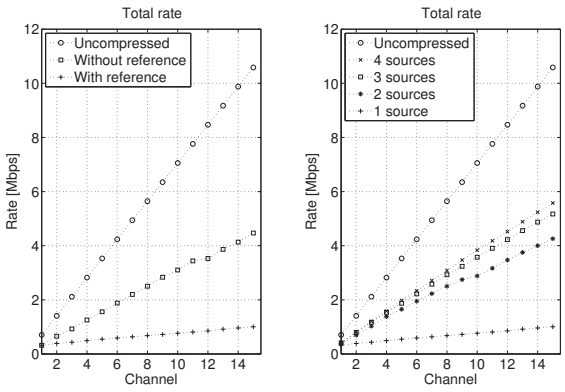
$$m = \left\lfloor log_2 \left(log_e(2) \cdot \frac{\sum_{n=1}^{256} |e(n)|}{256} \right) \right\rfloor, \quad (4)$$

where $e(n)$ is the prediction residual, and m is the only parameter that has to be transmitted to the decoder.

However, for highly correlated signals the content of the coding channel can often be close to zero when a reference channel is used. With Golomb-Rice coding all samples must be coded with a minimum number of 2 bits per sample, which is unnecessarily high in these situations. When the prediction residual contains mostly zeros, the encoder switches to Adaptive Huffman [11] coding. This algorithm is reset at the same time as the WCLMS prediction, meaning that blocks of 4096 samples are processed before the algorithm is reset. Another reason for including Adaptive Huffman is that for synthetic signals some B-format channels can be empty, and for these situations 2 bit per sample is also clearly unnecessary. Another possibility would perhaps be to use Run Length Coding (RLC), but preliminary results indicate that this will only outperform Adaptive Huffman for completely empty channels. For the signals containing round-off errors, the Adaptive Huffman algorithm will result in a lower data rate than RLC.

3.3. Network Transmission

The hierarchical structure of the Ambisonics B-format makes it very suitable for network transmission, and this layering is perhaps the most interesting feature from a network point of view. A traditional media stream will be transmitted using the User Datagram Protocol (UDP) and the sending rate will be decided by the encoder. However, as the amount of UDP traffic is increasing, this is far from an ideal solution since this may lead to unfair behavior towards TCP flows, which regulates its own sending rate depending on the available bandwidth. A flow is said to be reasonably TCP-friendly if its sending rate is within a factor 2 of the sending rate of a TCP flow competing in equal conditions [12]. With HOA the sending rate can be regulated by simply selecting the appropriate number of channels to transmit. Although scalable extensions to many existing audio formats have been developed [13], these formats will reduce the audio quality. The B-format is scalable in the spatial domain; reducing the number of channels will reduce the spatial resolution and not the audio quality.



(a) 1 source, compressed with and without a reference channel.

(b) Multiple sources, minimum bit rates only.

Fig. 3. Bit rates for synthetic clips without reverberation.

The use of an adaptive prediction algorithm introduces dependencies between frames, meaning that a packet loss will render the remaining packets until the next prediction reset useless. However, an advantage is that short frames can be used, thus reducing the total end-to-end delay. Another thing to consider is the packet size itself, in this work a packet is transmitted for every 256 samples. This can result in rather small packets, thus reducing the efficiency when considering the headers added from the network and transport layer. However, as seen in [14], the small packets resulting from short frames may actually lead to a lower packet loss ratio. When a router is congested it is more likely that a small packet will fit in the queue than a much larger ordinary sized packet.

4. RESULTS AND DISCUSSION

The coding scheme has only been tested with synthetic signals. However, signals that simulate a more natural recording have been created by adding reflections and filtering with room impulse responses. The signals have been synthesized up to order 7 and restricted to a 2 dimensional reproduction (15 channels). With a sampling rate of 44.1 kHz and 16 bit per sample, this results in an uncompressed bit rate of 10.58 Mbps.

4.1. Direct sound

For a clip with a single source and no reverberation, channels 2-15 will be equal to channel 1 (W) multiplied with a constant which is dependent on the source location. For these clips channels 2-15 will will be best predicted by using only the crosspredictor (the upper path in figure 1). This will lead to a prediction residual containing mostly zeros, but also some samples with values ± 1 due to round-off errors. For this prediction residual the Adaptive Huffman algorithm is used, which results in a rate of a bit more than 1 bit per sample, compared to the minimum of 2 bits per sample for Golomb-Rice coding. In figure 3a it can be seen how the use of a reference channel improves the coding gain. Due to the hierarchical structure of the B-format, the first channel is encoded without using a reference channel. The total bit rate for this clip becomes just above 1.1 Mbps when a reference channel is used, which means that the total bit rate has been reduced by close to 90%.

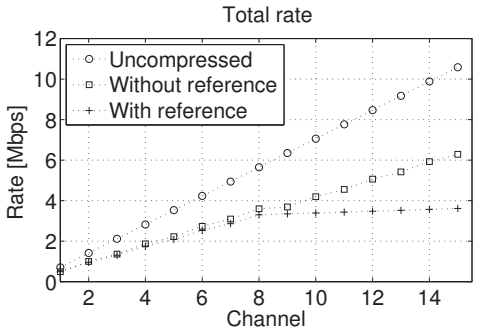
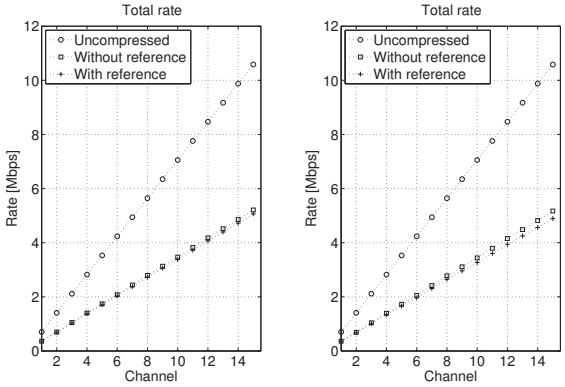


Fig. 4. 4 sources, placed in 0, 90, 180, and 270 degrees.



(a) Speech. (b) Music.

Fig. 5. Bit rates for simulated recording in a reverberant room.

In figure 3b more sources are added, and the total bit rate increases. However, the improvement from using a reference channel is still significant (results not shown) but decreasing as the number of sources increase. For the clip with three sources the total rate is reduced with 870 kbps, while with 4 sources the decrease is 580 kbps when a reference channel is introduced.

However, an important factor for the total bit rate is where the sources are located. In figure 4 the four sources are located with 90 degrees between them, and this results in a very high correlation between channels 9-15 and 1-8. The result is that there is a very low rate increase for including channels 9-15.

4.2. Reverberation

For microphone recordings the characteristics of the room will influence the recording, and the wall reflections will make the channels less correlated.

To simulate a more realistic recording, the mirror image [15] method is used to find the early part (80 ms) of the room impulse response and the late part is created by adding a decaying random tail. The signals have been filtered with this response to generate a signal resembling a microphone recording. A shoebox model with dimensions $4 \times 4 \times 2.4$ m and a reverberation time of 0.8 seconds was used and the distance between the source and receiver was set to be 2 m. Filtering the source with a room impulse response will function as adding noise to the signals, making it more difficult to take advantage of the inter-channel correlation and thus increasing the total rate compared with the signals containing only direct sound. It should be noted that rooms meant for music and speech communication typically have a quite low reverberation time, and the direct sound component might be dominant. Therefore, real-life scenarios might be somewhere between the two studied cases.

In figure 5 bit rates for two clips filtered with a room impulse response are shown. As seen from the figures the effect of using a reference channel is severely reduced. However, the reduction in bit rate can still be significant. For the clip containing music the use of reference channels reduces the bit rate with almost 279 kbps when all 15 channels are considered, which corresponds to a rate reduction of 5.4%. For the clip containing only speech the rate reduction is 2.8%.

5. CONCLUSION

A system for compressing and transmitting lossless Higher Order Ambisonics with low delay is presented. It is shown that by exploiting the inter-channel correlation between HOA-channels the rate can in certain situations be significantly lower when compared to compressing each channel independently. The cost of exploiting the inter-channel redundancy is a rather high computational complexity both for encoding and decoding. Since the algorithmic delay of the presented system is only 256 samples, it can be used for interactive applications such video conferencing and music performances.

6. REFERENCES

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

[2] 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.

[3] 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.

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

[5] M.A. Gerzon, P.G. Craven, J.R. Stuart, M.J. Law, and R.J. Wilson, "The MLP Lossless Compression System for PCM Audio," *J. Audio Eng. Soc*, vol. 52, no. 3, pp. 243–260, March 2004.

[6] M. Hans and R.W. Schafer, "Lossless Compression of Digital Audio," *IEEE Sig. Proc. Magazine*, vol. 18, no. 4, pp. 21–32, 2001.

[7] T. Liebchen, T. Moriya, N. Harada, Y. Kamamoto, and Y. A. Reznik, "The MPEG-4 Audio Lossless Coding (ALS) Standard - Technology and Applications," in *The 119th AES Conv.*, 2005, Preprint 6589.

[8] C. Chafe and M. Gurevich, "Network Time Delay and Ensemble Accuracy: Effects of Latency, Asymmetry," in *The 117th AES Conv.*, October 2004, Preprint 6208.

[9] ITU-T Recommendation G.114, "General Characteristics of International Telephone Connections and International Telephone Circuits: One-Way Transmission Time," February 1996.

[10] G. Schuller, B. Yu, D. Huang, and B. Edler, "Perceptual Audio Coding using Adaptive Pre- and Post-filters and Lossless Compression," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 379–390, 2002.

[11] J.S. Vitter, "Design and Analysis of Dynamic Huffman Codes," *J. ACM*, vol. 34, no. 4, pp. 825–845, 1987.

[12] M. Handley, S. Floyd, J. Padhye, and J. Widmer, "TCP Friendly Rate Control (TFRC): Protocol Specification," RFC 3448, Jan. 2003.

[13] R. Geiger, R. Yu, J. Herre, S. Rahardja, S-W. Kim, X. Lin, and M. Schmidt, "ISO/IEC MPEG-4 High-Definition Scalable Advanced Audio Coding," *J. Audio Eng. Soc*, vol. 55, no. 1/2, pp. 27–43, January/February 2007.

[14] J.E. Voldhaug, E. Hellerud, U.P. Svensson, A. Undheim, E. Austreim, and P.J. Emstad, "Influence of Sender Parameters and Network Architecture on Perceived Audio Quality," *Acta Acustica united with Acustica*, vol. 94, no. 1, pp. 1–11, 2008.

[15] J.B. Allen and D.A. Berkley, "Image Method for Efficiently Simulating Small-room Acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.