Transmission of High Quality Audio over IP Networks

Erik Hellerud

Centre for Quantifiable Quality of Service in
Communication Systems – *Centre of Excellence*
Department of Electronics and Telecommunications
Norwegian University of Science and Technology

A thesis submitted to the
Norwegian University of Science and Technology
Faculty of Informatics, Mathematics, and Electrotechnics

for the degree of
Doctor of Philosophy

April 2009

Abstract

This thesis deals with the transmission of high quality audio over packet switched networks, such as the internet. Unlike other transmission media, no guarantees are given regarding bandwidth, delay, or loss ratios, and all these factors will typically be time-varying.

For interactive, or two-way, applications it is important to keep a low end-to-end delay in order to have a high Quality of Experience (QoE). In this scenario, retransmission of lost packets is not possible simply due to the delay restrictions, thus one has to expect and prepare for periods with packet loss.

The varying bandwidth is also a challenge. A desired feature for media streams is scalability, or the ability to easily adjust the sending rate to the current conditions. Furthermore, the receivers may have a wide range of equipment and capabilities. Ideally the same media stream, or possibly a subset of it, should be used for all receivers. With a scalable format the use of transcoding is avoided, which will increase the complexity and possibly the delay as well.

Multiple aspects of audio transmission are handled in this thesis. Techniques for Error Protection (EP) and robust transmission, lossless and perceptual compression, and the quality experienced by the end users are investigated. A novel EP scheme where a perceptual criterion is used to select which parts of the audio that needs protection is presented. The quality improvement as a function of the added redundancy can then be quantified. Simulations show that adding moderate amounts of redundancy can result in a substantial quality improvement. The use of a network architecture supporting service differentiation has also been examined. A layered, lossless, and low delay codec is presented. The base layer is perceptually compressed, but when both layers are received the original audio stream is recovered. By transmitting the base layer at a high priority, the probability that at least this layer is received is increased and the use of regular Error Concealment (EC) is minimized. Moreover, the overhead from using the layering is very low when compared with regular lossless coding.

Higher Order Ambisonics (HOA) is a naturally scalable multichannel format for wave field reproduction. In this work, both lossless and perceptual compression of this format is examined. Perceptual coding results in an error in the reproduced wave field which is very low in the sweet spot, but increases as a function of the distance from the center. For the lossless compression scheme, the focus is

on maintaining the scalability and keeping a very low delay for both encoding and decoding. The compression of synthetic signals, and signals recorded with a spherical microphone array, is evaluated, and the results show that the total rate is significantly reduced by exploiting the inter-channel correlation.

The perceived quality of packet loss distorted audio is also investigated. Effects of different packet loss processes resulting from simulations of two network architectures are compared using subjective testing. The test revealed that the less bursty loss process lead to higher perceived quality for the lowest packet loss ratio. Packet sizes were also considered in the simulations and it was found that using small packets may be beneficial. A new type of subjective test has also been used, where the users' acceptability for packet loss is evaluated. As expected, the acceptability decreases quickly as the packet loss ratio increases, and it is seen that music exposed to a packet loss ratio of 1.5% can no longer be said to have acceptable quality.

Preface

This dissertation is submitted in partial fulfillment of the requirements for the degree of PhD at the Norwegian University of Science and Technology (NTNU). This work has been done within the integrated PhD program and started in August 2004. During the two first years of this program, 50% of the time is spent on finishing the MSc-thesis, while the remaining time is used to start the PhD education.

Since August 2004 I have been funded and hosted by the Centre for Quantifiable Quality of Service in Communication Systems (Q2S), but the work has officially been conducted at the Department of Electronics and Telecommunications.

I have, as seen from the included papers, worked with several people. I would especially like to thank my supervisor professor U. Peter Svensson for interesting discussions and insightful feedback. The help from all my co-authors has also been very valuable and I am grateful for the assistance of Jerome Daniel.

Erik Hellerud
Trondheim, Norway
April 2009

Contents

Abbreviations

AAC	Advanced Audio Coding
AAC-ELD	Enhanced Low Delay AAC
AAC-LD	Low Delay AAC
AF	Assured Forwarding
ALS	Audio Lossless Coding
APES	Amplitude and Phase Estimation
ATH	Absolute Threshold of Hearing
AU	Access Unit
BE	Best Effort
CD	Compact Disc
DCCP	Datagram Congestion Control Protocol
DD+	Dolby Digital Plus
DiffServ	Differentiated Services
DVD-A	DVD Audio
DTS	Digital Theatre Systems
DWT	Discrete Wavelet Transform
EC	Error Concealment
ECN	Explicit Congestion Notification
EF	Expedited Forwarding
EP	Error Protection

FEC	Forward Error Correction
FFT	Fast Fourier Transform
FLAC	Free Lossless Audio Codec
GAPES	Gapped-data Amplitude and Phase Estimation
GMT	Global Masking Threshold
HE-AAC	High Efficiency AAC
HOA	Higher Order Ambisonics
IP	Internet Protocol
ITU	International Telecommunication Union
LMS	Least Mean Squares
LP	Linear Prediction
LPC	Linear Predictive Coding
MD	Multiple Description
MDCT	Modified Discrete Cosine Transform
MLP	Meridian Lossless Packing
MOS	Mean Opinion Score
MPEG	Moving Pictures Expert Group
MP3	MPEG-1 Layer 3
MUSHRA	Multiple Stimuli with Hidden Reference and Anchor
MTU	Maximum Transmission Unit
ODG	Objective Difference Grade
PEAQ	Perceptual Evaluation of Audio Quality
PESQ	Perceptual Evaluation of Speech Quality
PHB	Per-Hop Behavior
PLR	Packet Loss Ratio
QoE	Quality of Experience
QoS	Quality of Service

RED	Random Early Detection
RLC	Run Length Coding
RTCP	RTP Control Protocol
RTP	Real-time Transport Protocol
SBR	Spectral Band Replication
SDG	Subjective Difference Grade
SLA	Service Level Agreement
SLS	Scalable Lossless Coding
SNR	Signal to Noise Ratio
SPL	Sound Pressure Level
TCP	Transmission Control Protocol
TFRC	TCP Friendly Rate Control
UDP	User Datagram Protocol
UEP	Unequal Error Protection
VoIP	Voice over IP
WCLMS	Weighted Cascaded Least Mean Squares
WFS	Wave Field Synthesis
WMA9	Windows Media Audio 9

Summary of Papers

Seven papers are included in part II of this thesis, and a summary of each is presented in this chapter. Some of the included papers have undergone minor typographical changes since they were published.

Paper A: ***Perceptually Controlled Error Protection for Audio Streaming over IP Networks***

This paper proposes an Error Protection (EP) scheme for audio streaming over Internet Protocol (IP) networks. The idea is to simulate packet losses at the sender side and evaluate the performance of the Error Concealment (EC) algorithm using a perceptual criterion. The sender compares the EC with a low bit rate encoded version of the same frame, and if the low bit rate frame is found to be better than the error concealment by a certain margin, it is sent as error protection. In this work the comparison between the low bit rate version and the error concealment is performed with the full reference metric Perceptual Evaluation of Audio Quality (PEAQ).

Paper B: ***Robust Transmission of Lossless Audio with Low Delay over IP Networks***

A layered low-delay codec is presented. The base layer is perceptually encoded using a pre- and postfilter and the difference between the base layer and the original signal is the enhancement layer. The total algorithmic system delay is only 256 samples. When both layers are received the original signal is recovered. By using the Differentiated Services (DiffServ) architecture the base layer can be transported with high priority, thus increasing the probability that at least this layer is received. Also, the layering increases the total rate only slightly when compared to regular lossless coding.

Paper C: ***Encoding Higher Order Ambisonics with AAC***

In this paper it is evaluated how encoding Higher Order Ambisonics (HOA) with Advanced Audio Coding (AAC) distorts the reproduced sound field. It is found that the error is increasing as a function of the distance from the sweet spot, and

this error is also frequency-dependent. Furthermore, it is shown that by using more bits for the lower order signals the error in the sweet spot becomes very low, and that the spatial information is improved even if very low bit rates are used for the higher order signals.

Paper D: ***Spatial Redundancy in Higher Order Ambisonics and its Use for Low Delay Lossless Compression***

The spatial redundancy for synthetic HOA signals is evaluated using a low delay lossless codec. By exploiting the inter-channel correlation the total bit rate is significantly reduced, but this reduction is dependent on both the number of sources and also the sources location. The presented codec has a total system delay of 256 samples, and preserves the desired HOA features such as the scalability and the ability to use arbitrary loudspeaker layouts for reproduction.

Paper E: ***Lossless Compression of Spherical Microphone Array Recordings***

The same compression system as in paper D is evaluated for real recordings from a spherical microphone array, the "Eigenmike"®. It is found that the microphone signals are highly correlated, and that a significant rate reduction can be achieved by exploiting the inter-channel correlation. The signals converted to the spherical harmonics domain are also highly correlated, but it is more difficult to achieve high compression gains for these signals.

Paper F: ***Influence of Sender Parameters and Network Architecture on Perceived Audio Quality***

Network simulations are used to evaluate the differences between the Best Effort (BE) and DiffServ architectures for audio transmission. It is found that DiffServ makes the packet loss process less bursty, which leads to a higher perceived quality for the lowest loss ratio. From the network simulations it was also seen that using small packets may be beneficial since it leads to a lower packet loss ratio.

Paper G: ***How Much is Too Much? – On the Acceptability of Packet Loss Distorted Audio***

An acceptability test is used to evaluate the perceived quality of packet loss distorted audio. It is found that the acceptability quickly decreases as the packet loss ratio increases. Also, the results show that there is no significant difference between 64 and 128 kbps AAC (stereo) in situations with packet loss. Clip lengths ranged from 25 to 28 seconds, and by having all packet losses in either the first or last half of the clip, it can be seen that packet losses early in the clip are perceived as more acceptable.

Papers not Included

These papers are related to the thesis, but not included.

- J.E. Voldhaug, E. Hellerud and U.P. Svensson. "Evaluation of Packet Loss Distortion in Audio Signals," in *the 120th AES Convention*. Preprint 6855. Paris, France. May 2006.
- J.E. Voldhaug, E. Hellerud, A. Undheim, E. Austreim, U.P. Svensson and P.J. Emstad. "Effects of Network Architecture on Perceived Audio Quality," In *Proceedings of the 2nd ISCA Tutorial and Research Workshop on Perceptual Quality of Systems*. Berlin, Germany, Sep, 2006.[1]
- A. Solvang, U.P. Svensson and E. Hellerud. "Quantization of 2D Higher Order Ambisonics Wave Fields." In *the 124th AES Convention*. Preprint 7370. Amsterdam, the Netherlands. May 2008.

[1] This paper is an early version of paper F.

Part I

Background and Motivation

Chapter 1

Introduction

Transmission of high quality audio over the current internet is a challenging task. The traditional telephone lines are line switched, which means that the user reserves this resource when it is used. In a packet switched network, like the internet, multiple users will share the same transmission medium. While this is a resource-efficient approach, this will typically make the available bandwidth and network delay time varying, thus these networks can become highly resource constrained.

Regular stereo with Compact Disc (CD) quality results in an uncompressed bit rate of 1.41 Mbps, and for multichannel audio this rate will increase further. To reduce the data rate the signals can be compressed. With lossless compression the rate will typically be reduced with 20-60% [1], depending on the signal content. With perceptual coding, the data reduction can be more than 90%, while still keeping an almost "transparent" audio quality [2]. The high compression gain makes perceptual compression interesting in situations where a quality loss is acceptable, but lossless compression is the natural choice for applications that require maximum quality and possibly further processing.

Given a compressed audio stream, an important aspect is how the transmission will influence the end-users perceived quality. For interactive applications, such as video conferencing or distributed music plays, the end-to-end delay is identified as an important factor. For speech applications it is advised to keep the total delay below 150 ms [3]. In a distributed music play, the quality will begin to reduce already at 60 ms delay, and the effect is noticed already with 25 ms delay [4, 5]. The encoder itself may contribute significantly to the total delay since the encoding delay can often be in the range of several hundred milliseconds for general purpose encoders [6].

Packet loss is also a serious problem for real-time audio streaming [7]. For applications that are not delay-sensitive, a long playout buffer makes it possible to retransmit lost packets and receive them before their scheduled playout time. For real-time applications, the playout buffer has to be so short that it makes

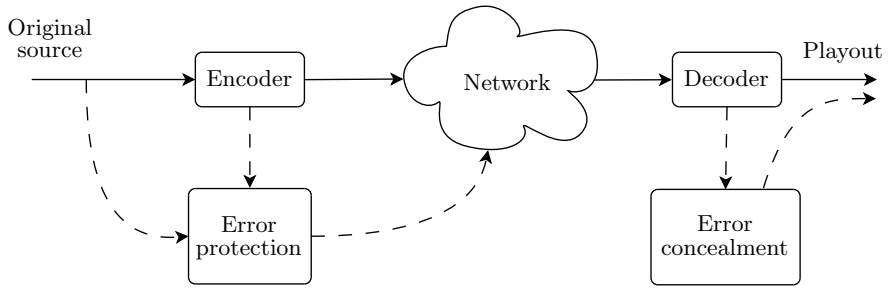


Figure 1.1. Overview of audio transmission over the internet.

retransmissions impossible. In these situations the content of the lost packet has to be replaced by the decoder, a process known as Error Concealment (EC). A typical packet may contain 5 – 60 ms of audio and a packet loss will then lead to one long or multiple short gaps, depending on the packetization, in the stream [8]. This concealment can be improved, or simplified, if the sender employs some sort of Error Protection (EP). In general, error protection can be said to add redundancy to or modify the audio stream, in order to make it easier to handle transmission errors. How a packet loss will influence the perceived quality depends strongly on the content of the packet, the error concealment, and also the application. This has been thoroughly studied for narrowband speech, but to a lesser extent for music.

For streaming media it is an advantage to be able to adapt to the varying bandwidth, both to avoid packet losses but also to be fair against the other users of the same transmission medium. For multichannel audio, due to the potentially high bit rate, this can be said to be very important. Also, it is advantageous if the same media stream can be used for all receivers, even if they have different kinds of playback equipment. Without this ability the media stream has to be transcoded into different formats, which will increase the complexity at the sender side and possibly also the delay.

This thesis deals with many aspects of transmission of high quality audio over the internet. From the sender side, with a focus on error protection, compression, and scalability, to the receiver side with focus on the resulting perceived quality. The first part of this thesis contains necessary background theory and motivation for this work. Chapter 2 handles the history and theory of multichannel audio with special focus on Higher Order Ambisonics (HOA), a multichannel wave field reproduction technique, which is used in three of the included papers. A brief introduction to lossless and perceptual coding is given in chapter 3 and networking theory and techniques for coping with network problems are presented in chapter 4. Subjective and objective quality evaluation for audio is covered in chapter 5 and concluding remarks are given in chapter 6. The second part of this thesis consists of the seven included papers.

Chapter 2

Multichannel Audio

Numerous techniques for multichannel audio reproduction exist. Regular surround systems such as 5.1 and 7.1 are sweet-spot techniques. This means that the 5.1 or 7.1 channels are mastered such that the correct sound reproduction only happens at one point in space, but there is a reasonable effect also outside the sweet-spot. In order to achieve correct sound reproduction over an extended area more advanced approaches are needed.

There are not many systems available for complete wave field reproduction, and none has really achieved any commercial success. One of the systems that has arguably gained most interest is Wave Field Synthesis (WFS) [9,10]. WFS is based on Huygens' principle, which says that each point on a wave front can be seen as the starting point of a new radiating wave. The pressure and particle velocity on a closed surface can then be sampled and played back using loudspeakers in the same positions. The upper frequency limit for correct reproduction will then be determined by the physical distance between the samples.

A different approach for sound field reproduction is Higher Order Ambisonics (HOA), which is based on a spherical harmonics decomposition of the wave field. While this is a more complicated approach than WFS, it results in a format which is easily scalable and thus more appropriate for network transmission.

Due to the requirement for a high number of loudspeakers, the most important application for wave field reproduction systems has historically been cinemas. However, the number of homes with 5.1 or 7.1 systems is rapidly increasing, meaning that it is possible to utilize these systems for wave field reproduction as well. For instance, third order, two dimensional Ambisonics is used in two games for the Sony Playstation 3 [11]. Another important use is videoconferencing, since the use of HOA can allow for better spatial separation of the participants, which improves the speech intelligibility.

HOA is used in papers C – E and is presented briefly in the following section.

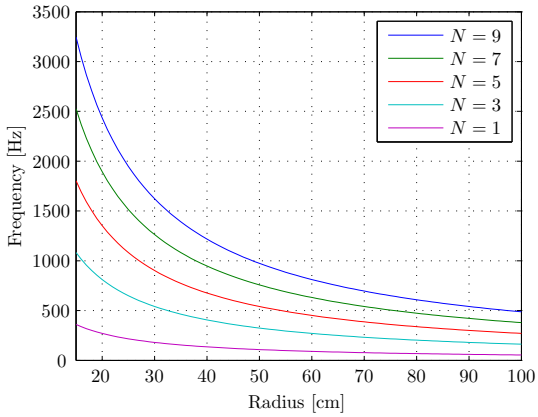


Figure 2.1. Frequency limit as function of reconstruction volume/area radius and order.

2.1 Higher Order Ambisonics

Ambisonics is based on work as far back as the 1970s [12]. For three dimensional reproduction this system is based on spherical harmonics decomposition, while cylindrical harmonics decomposition is used for two dimensional reproduction [13]. A three dimensional first order representation is equivalent to the signals recorded with one omni-directional and three figure-of-eight microphones. This theory was further developed in [14, 15] into so-called HOA.

The number of channels is determined by the order of the decomposition. For a three dimensional reproduction the number of channels, M, is given as $M = (N+1)^2$, where N is the order in the spherical harmonics domain. For the two dimensional case is $M = 2N + 1$.

The spatial reproduction is only accurate up to a frequency and within an order dependent radius. The normalized truncation error, ϵ, has been calculated in [16], and the error is given as

$$\epsilon_N(kr) = 1 - \sum_{n=0}^{N}(2n+1)(j_n(kr))^2, \tag{2.1}$$

where k is the wave number defined as $k = \frac{2\pi}{\lambda}$, λ being the wave length, r is the distance from the center, and j_n is the n'th order spherical Bessel function. It is found that if $N < kr$ the error is less than 4% which should be acceptable for most practical applications. For a two-dimensional representation the expression for the error is somewhat different, but the relationship $N < kr$ still holds. In figure 2.1 this frequency limit is shown as a function of order and radius.

Figure 2.2 shows an example of plane wave reproduction, and it can be clearly

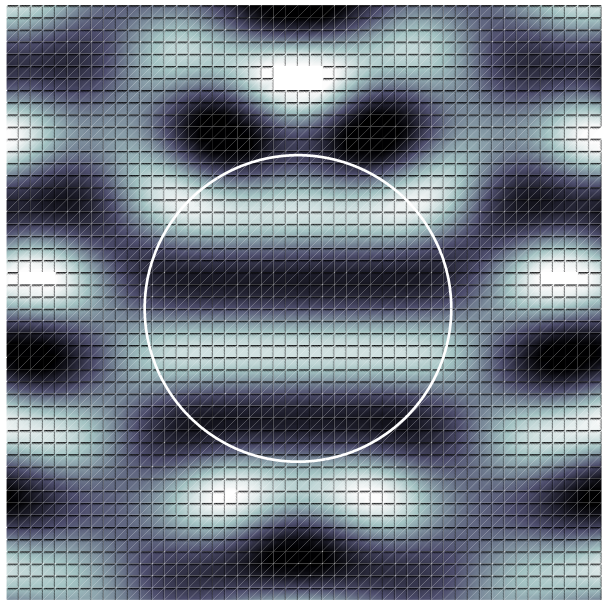


Figure 2.2. Plane wave reproduced with two dimensional 7-th order HOA. The frequency is 3000 Hz and the circle illustrates $r = \frac{N}{k} = 12.8$ cm.

seen that within the reconstruction radius a plane wave is achieved. Outside the radius the wave is distorted. The waves radiating from the loudspeakers are not added correctly, leading to noticeable interference effects.

By reducing the order, the radius of the volume, or area, with good reconstruction is decreased. This can be seen as a scalability in the spatial domain. Also, a relatively fine granularity is achieved. Reducing the order by one will remove two channels for a two dimensional reproduction. For three dimensions the reduction is order dependent as seen in figure 2.3.

Both synthesizing and decoding HOA signals is relatively straight-forward: The operations are simply matrix multiplications [14]. The encoding matrix is dependent on the locations of the sources, while the decoding matrix is dependent on the loudspeaker layout. The number of loudspeakers, L, is restricted by $L \geqslant M$, but using more loudspeakers than the minimum will not necessarily increase the quality [17]. The fact that the decoding matrix depends only on the loudspeaker layout is especially interesting. This means that the same audio signals can be used for all receivers and the decoder adapts the channels to the current layout. Also, the number of channels received can be matched against the number of loudspeakers available. Special techniques for decoding to 5.1 [18], stereo [19], and even WFS [20] are available.

In contrast to synthesizing, recording a complete soundfield is more compli-

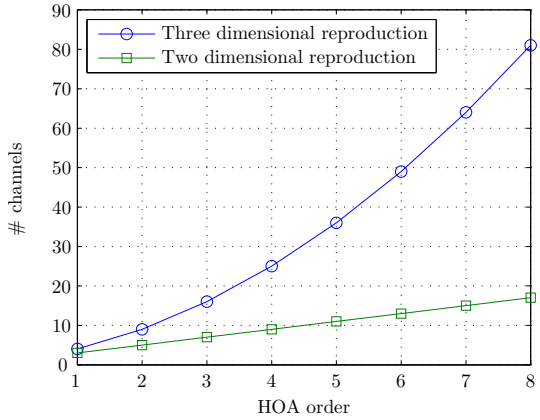


Figure 2.3. Number of channels needed as a function of order for both two and three dimensional reproduction.

cated. The Sound Field microphone [21] has four microphone capsules and is able to record first order Ambisonics. In [22] a spherical microphone array was presented. This array has 32 microphone capsules, and the recordings can be converted into fourth order HOA [23]. Recently, this microphone has become commercially available under the name "Eigenmike"® [24].

While the use of HOA is interesting from a networking point of view, due to the scalability and the flexibility, a high order will result in a very high bit rate. In paper C it is evaluated how encoding the HOA-signals with a perceptual codec will affect the spatial distortion. Applying a perceptual codec to HOA-signals has, to our knowledge, not been studied earlier. Papers D and E focus on the lossless compression of both synthetic signals and natural recordings. In these papers a lossless low delay codec is presented. The main difference between this encoder and other multichannel compression techniques [25, 26] is that it supports the hierarchical structure of the HOA-signals while still keeping a very low delay for both encoding and decoding. By restricting the delay this codec can be used for applications with real-time requirements such as video conferencing and distributed music plays. This codec also exploits the inter-channel correlation in order to achieve higher compression gains than what is possible when each channel is encoded independently.

Chapter 3

Audio Coding

On a regular Compact Disc (CD) audio is represented using a sampling frequency of 44.1 kHz and 16 bits per sample. For a monaural channel this leads to an uncompressed bit rate of 706 kbps. Audio compression is used in order to reduce the bit rate necessary for transporting or storing an audio signal. In general, there can be said to be two types of compression. When using lossless compression, as the name indicates, the original source signal should be recovered exactly after decoding, meaning that only redundancy is removed during the compression process. Lossless compression is the natural choice when maximum quality is desired and when one does not have strict storage or bandwidth restrictions. Furthermore, having the original signals is also advantageous in situations where processing is planned at a later stage.

On the other hand, with perceptual coding the goal is not that the decoder reconstructs the original signal, but rather reconstructs a signal that is perceptually close to the original audio signal. Here, the encoder tries to remove irrelevance in addition to redundancy. The irrelevance consists of the parts of the audio signal that humans are not able to perceive. This type of compression is useful in order to greatly reduce the audio data rate, such as for portable music players, or for streaming applications where some quality degradation is acceptable.

Perceptual and lossless coding will be covered in the two next sections.

3.1 Perceptual Coding

When a perceptual codec is given a target bit rate the goal is that the encoded audio should be perceived as close to the original source quality as possible, and ideally the compressed signal should be perceived as identical to the original source. To achieve this it is necessary to exploit the limitations of the auditory system.

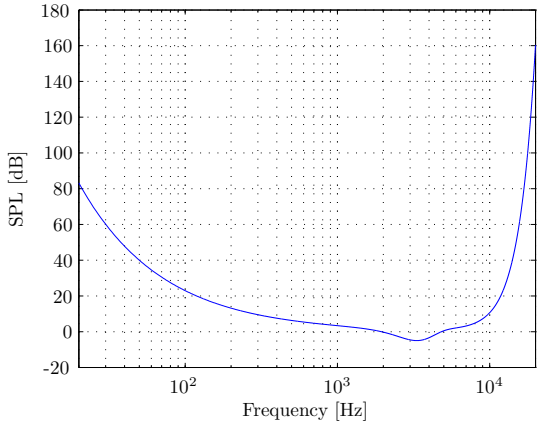


Figure 3.1. Absolute threshold of hearing (from [29]).

3.1.1 Psychoacoustics

Although the human hearing is very sophisticated it has some limitations that can be exploited in communication. One of the more important factors for audio coding is the Absolute Threshold of Hearing (ATH). The ATH defines the minimum Sound Pressure Level (SPL) for hearing a pure tone, and this threshold is highly frequency dependent (figure 3.1). Although this definition is for pure tones it is also used as an indication of how much noise the quantizer can add within a frequency band without it being noticeable. For instance, above 20 kHz and below 20 Hz our hearing has severely reduced sensitivity [27, 28], which simply means that a perceptual codec can remove most of the information in this range.

The ear functions as a series of overlapping band pass filters, known as auditory filters, and the width of these filters is increasing with frequency [27, 30]. Given noise and a pure tone, the pure tone is more easily detected if it is located in a different auditory filter than the noise. Masking occurs when signal components cannot be perceived, and in a perceptual codec these components can be removed. A pure tone may also mask weaker signals, both other pure tones and noise. In these situations one has to investigate the spreading of the pure tone. This asymmetrical triangular spreading function is related to the auditory filters and it has a much steeper slope towards the lower frequencies than the higher frequencies.

Non-simultaneous masking is known as temporal masking. If a pure tone is audible for a limited period of time, it can actually mask other tones both before its onset (pre-masking) and after it has been removed (post-masking). However, post-masking is more dominant than pre-masking [31]. Temporal masking is exploited in several codecs, such as MPEG-1 [32].

In order to create a psychoacoustical model one has to investigate the signal in time frames, and combine the results from the masking analysis with the ATH and derive the Global Masking Threshold (GMT). The encoder can then quantize the signal such that the quantization noise stays below the GMT, and in theory the noise should then be inaudible.

The importance of these psychoacoustical principles is shown in [33] and described there as the "13-dB miracle". The masking threshold is calculated using the MPEG-1 Psychoacoustic Model 2 and noise shaped according to this threshold is introduced. While this results in a signal to noise ratio as low as 13 dB, the noise is barely audible. Given white noise with the same signal to noise ratio the noise is easily audible.

3.1.2 Perceptual Audio Coders

Perceptual coders can be divided into four categories [34]: Sinusoidal, subband, linear predictive, and transform coders.

A sinusoidal coder will encode the audio signal using a sum of sinusoidals with different frequency, phase and amplitude. This approach is typically most successful for very low bit rates (6-24 kbps) [35, 36].

Subband coders use a filterbank with bandpass filters. Based on the results from the psychoacoustic model, the output of each filter is quantized before it is entropy coded. MPEG-1 Layer 1 and 2 are classical subband coders [32]. These codecs use a filterbank with 32 uniform bandpass filters for signal decomposition, but the calculation of the psychoacoustical parameters is performed using a 512-point Fast Fourier Transform (FFT) for Layer 1 and a 1024-point transform for Layer 2. High quality is achieved at 192 kbps per channel for Layer 1 and 128 kbps per channel for Layer 2. More modern subband approaches based on a Discrete Wavelet Transform (DWT) decomposition have also been presented [37].

Linear Prediction (LP) coders have traditionally been used mostly for speech coding, both narrowband [38] and wideband [39]. LP analysis is used to extract the spectral envelope, and the prediction residual is encoded together with the prediction parameters. Recently LP-coding has also been used for compression of wideband music. In [40] a codec based on LP is presented, where good audio quality is achieved around 100 kbps for regular monaural CD-quality signals. This work was followed up in [41–44] where this technique was used to implement a low delay codec. The work finally presented in [45] is a codec with 256 samples delay which results in good audio quality for 32 kHz signals at 64 kbps (monaural). The key to keeping the delay low is the use of backwards adaptive prediction, since this allows the encoder to continue the prediction over frame boundaries without resetting the predictors. In paper B of this thesis this codec is modified into a layered codec in order to provide lossless compression with a perceptually encoded base layer. By using a layered approach, the base layer can be transmitted using a high priority, thus increasing the probability that at least this layer is received and the use of regular Error Concealment (EC) is minimized. In situations with

no packet loss, the original signal is recovered. Dividing the signal into two layers increases the total bit rate only slightly when compared to regular lossless compression.

Transform coding is currently giving the best results for high quality audio. Here, the time-domain signal is transformed into the frequency domain for psychoacoustical analysis and quantization. A common transform is the Modified Discrete Cosine Transform (MDCT) [46]. This transform is used in the second stage of MPEG-1 Layer 3 (MP3) [47] and in MPEG Advanced Audio Coding (AAC) [48–50]. AAC uses a 2048-point MDCT for signal decomposition, or 8 shorter frames of 256 samples if transients are detected within the frame. This rather long frame size will give a good frequency resolution for the psychoacoustical analysis, but it also comes with a downside; the frame size contributes significantly to the total system delay [6]. Furthermore, AAC uses a large bit reservoir which introduces additional delay. These factors make regular AAC not suitable for real-time, or two-way, applications.

MPEG-4 Low Delay AAC (AAC-LD) [51] mitigates this shortcoming by reducing the window size to maximum 1024 samples and minimizing the use of a bit reservoir. AAC-LD has a delay of less than 20 ms at 48 kHz sampling rate. In [52] the MPEG-4 Enhanced Low Delay AAC (AAC-ELD) is presented. Here, the perceptual quality is increased by using Spectral Band Replication (SBR) [53–56] while still keeping the delay around 20 ms. It is common that the correlation between the high and low frequencies is significant, and with SBR the high frequencies are represented using the low frequencies. To achieve low bit rates without SBR the encoder will typically have to reduce the bandwidth of the signal, or accept that the quantization noise is above the GMT. With SBR it is possible to keep the full bandwidth of the signal even for very low bit rates.

Scalability is an important concept for audio coders used for network streaming. With scalability, a subset of the encoded bit stream can be extracted and used to decode a representation with lower perceived quality. MPEG-4 Scalable Lossless Coding (SLS) [57] is one of the latest developments for this class of coders. Using this codec the full representation is lossless, but if a subset is extracted this will be an AAC encoded version of the original source. This leads to a perceptual scalability, increasing the bit rate will also increase the perceived quality.

In [58] a new framework for audio coding is presented, where the goal is to use the same codec for all types of input and output. By combining a transform, sinusoidal, and a predictive coder, the perceptual rate-distortion optimization module automatically determines how the bits should be distributed among the different coding techniques. The end result is a codec that can adapt to a wide range of transmission conditions and application requirements.

One of the latest developments for compression of multichannel audio is MPEG Surround [59, 60]. This encoder downmixes the input signals down to either a mono or stereo signal, and the spatial information is transmitted as side information. The downmix can be encoded using any codec, as long as the receiver supports it. The spatial information consists of information about the

level differences between the channels, prediction parameters for predicting a channel from others, and also prediction errors. The receiver will then combine the downmix with the spatial cues and generate e.g. 5.1 output. Using this technology it is possible to transmit multichannel audio with a bit rate as low as 48 kbps. However, in [61] several multichannel codecs are evaluated. Among these is MPEG Surround with the downmix encoded with High Efficiency AAC (HE-AAC), which is AAC extended with SBR, at bit rates of 64 and 96 kbps. Here it is found that while this on average gives a good quality, it results in poor quality for some critical test clips such as applause. The only codecs in this test that achieve excellent sound quality for all test clips are Digital Theatre Systems (DTS) [62] at 1.5 Mbps, and Dolby Digital Plus (DD+) [63] and Windows Media Audio 9 (WMA9) at 448 kbps. This indicates that rather high bit rates are needed to have high quality multichannel audio.

A different system for compressing 3/2 surround is investigated in [64, 65]. Here, the audio channels are transformed into a hierarchical domain where the channels are ordered after their perceptual importance. The goal here is not to be able to remove channels but rather to bandlimit the available channels such that the most important channel keeps the maximum bandwidth while the less important channels can be bandlimited. This system is mostly intended as a pre-processing step before a general low bit rate multichannel codec is applied.

Perceptual coding of wave field reproduction formats has not been examined much previously, due to the complexity of applying a perceptual model for a complete wave field. However, in paper C it is evaluated how encoding the HOA-signals with AAC affects the reproduced wave field.

3.2 Lossless Coding

The goal with lossless coding is to compress a signal as much as possible, but still be able to reconstruct the original signal after decoding. Several tools exist for compressing general data files such as *gzip* [66] and *bzip/bzip2* [67]. But these programs are rarely successful for audio signals; they are more suited for compressing documents and other text files.

Lossless audio coders are typically less complex than perceptual coders, simply because they do not need to implement neither a perceptual model nor quantization. Most coders are based on linear prediction [26, 68–70] where the current sample is estimated and the prediction residual $e(n)$ is entropy coded (eq. 3.1).

$$e(n) = x(n) - \sum_{i=1}^{N} a_i x(n-i) \tag{3.1}$$

It is also possibly to use a transform in a lossless codec, as long as the transform has an exact inverse. In MPEG-4 SLS an integer MDCT [71] is used to

achieve the scalability between the perceptually encoded base layer and lossless compression.

Coding is employed to remove redundancy from the residual signal entropy. Huffman and Run Length Coding (RLC) are two common and well known alternatives. Rice coding (also known as Golomb-Rice coding) is also a common choice for lossless audio coders, and this scheme is a variant of Golomb codes [72]. When the prediction residual is Laplacian distributed it has been shown that Rice coding is actually equal to Huffman coding [73].

To illustrate the difference between a general purpose and an audio compressor a clip with classical music has been used. The uncompressed file is 73.5 MB, and *gzip* is able to reduce the file size to 63.0 MB or 85% of the original size. While this still is a significant improvement, Free Lossless Audio Codec (FLAC) [69] is able to reduce the file size to 44.6 MB, which is only 60% of the original size.

For multichannel audio one can of course use a regular lossless codec for each channel. However, there can be situations where there are significant amounts of inter-channel redundancy to remove. In MPEG-4 Audio Lossless Coding (ALS) inter-channel redundancy is removed using a simple three-tap filter [26]. Meridian Lossless Packing (MLP) is used for DVD Audio (DVD-A), and uses a matrix-transformation followed by prediction to remove inter-channel redundancy [25]. MLP supports hierarchical data such as HOA, but this codec does not focus on low delay, but rather to minimize the peak data rate to allow playback from bandlimited media such as DVD-A.

In papers D and E a lossless compression scheme for multichannel audio is presented. The most important difference between this codec and previous work is that it maintains a very low delay for both encoding and decoding, which makes this codec suitable for applications with real-time requirements. This codec exploits the inter-channel correlation in order to remove additional redundancy, while keeping the hierarchical structure of the HOA format. The argument for doing this is to preserve the scalability and also keeping the ability to use arbitrary loudspeaker layouts for reproduction. Both synthetic and natural recordings are used to evaluate the system.

Chapter 4

Audio Transmission over IP Networks

In order to successfully transmit audio over the internet, the bit stream has to be split into packets. This process is known as packetization. Each packet can then be transmitted to the receiver using a wide range of different protocols. Network protocols, packetization, and how to handle network problems are described in this chapter.

4.1 Network Protocols

The internet is based on a layered architecture where each layer provides an interface for the layer above [74]. Although this model originally had seven layers, it is most common to use the simplified model with a stack of five layers. From the bottom up they are the physical layer, data link layer, network layer, transport layer, and application layer. The physical and data link layer only deal with connections between two neighboring network nodes, while the layers from the network layer and up handle end-to-end connections. Only the protocols that handle end-to-end connections are of interest for this work, thus they are the only ones covered here.

4.1.1 Network Layer

The most common protocol on the network layer is the Internet Protocol (IP) [75]. This is a connectionless protocol, which means that a packet can just be transmitted to the receiver. The connection does not need to be set up beforehand. Furthermore, IP is an unreliable protocol and no guarantees are provided regarding whether or not a transmitted packet actually arrives at its destination. This means that a packet can be dropped anywhere in the network,

for instance due to congested routers or simply due to link failures, and the sender will not be informed. Packet losses have to be handled at a higher layer.

4.1.2 Transport Layer

The Transmission Control Protocol (TCP) is often used on the transport layer for regular data transport [76]. This is a connection-oriented protocol, meaning that the connection has to be set up before transmission begins. TCP is also a reliable protocol, and it is guaranteed that the sent data is actually received. The receiver will transmit back an acknowledgment for each packet received. If the sender does not receive this acknowledgment within a certain time frame the packet will be retransmitted. The received packets are reordered and delivered to the layer above in the correct order. Another important feature with TCP is that this protocol adjusts its sending rate to the current network conditions. TCP will increase the sending rate as long as acknowledgments are received. When an acknowledgment is missing, it is assumed that this has happened due to congestion and the sending rate will be reduced. By doing this all TCP flows will be competing for the network resources on equal terms.

However, for applications with real-time requirements, like videoconferencing or regular Voice over IP (VoIP), TCP is not suitable. The retransmission and reordering system in TCP, which are appropriate for file transfers, can result in problems for real-time audio traffic. When a packet is lost, the receiver will not be able to use this or any future packets until the lost packet is received. The receiver has no choice but to stop playback while waiting for the retransmission. This problem can be overcome by using a rather large playout buffer, and this is perhaps the most suitable solution for transmissions without strict real-time requirements. For internet radio, buffering a few hundred milliseconds before playback starts will not create any problems for most users, but for real-time applications this approach is not appropriate since this playout buffer will introduce too much delay. Instead, the User Datagram Protocol (UDP) [77] is often used on the transport layer for media traffic.

UDP is a simple connectionless protocol. It basically adds a checksum of the content such that the receiver can check that the correct data has been received. This protocol does not include any mechanisms for retransmission and it does not guarantee that all packets are received. This simple approach is more suitable for real-time applications, because it allows the application to handle packet loss in a more efficient manner without the need of a large playout buffer.

However, this protocol does also have some rather large disadvantages. One of the key features with TCP is the regulation of transmission rate, and UDP does not provide anything like this. This can result in an UDP flow taking more than its fair share of the available bandwidth. In [78] a flow is said to be reasonably fair "if its sending rate is generally within a factor of two of the sending rate of a TCP flow under the same conditions." Ideally, when using UDP the sending rate should be controlled by a different mechanism in order to be TCP-friendly.

Another, and perhaps more appropriate, protocol for media streaming is the Datagram Congestion Control Protocol (DCCP) [79]. DCCP is a replacement for UDP and comes with three different congestion control mechanisms, e.g. TCP Friendly Rate Control (TFRC) [78]. However, the use of congestion control requires that the sender is actually able to change its transmission rate. In a situation with only one receiver the encoder can change the target bit rate. In situations with multiple participants a scalable encoder is a more elegant solution, since a subset of the bit stream can be sent to the receivers who require less bandwidth, instead of reducing the quality for all participants.

4.1.3 Application Layer

A widely used protocol on the application layer for multimedia traffic is the Real-time Transport Protocol (RTP) [80]. This protocol provides the stream with sequence numbers and a timestamp which tells the receiver about the scheduled playback time. The protocol on the layer below may deliver packets out of order and packet losses will still be detected easily. The RTP Control Protocol (RTCP), also defined in [80], specifies a protocol for transmitting feedback about the quality of the data distribution. This feedback can be used to adjust sending rate in order to avoid bandwidth problems.

4.2 Differentiated Services

Currently, the architecture for the internet is named Best Effort (BE). BE gives no guarantees and treats all traffic equally. While this is a reasonable solution for regular data traffic and file transfer, multimedia traffic could benefit from a service differentiation due to the requirements for low delay and a low packet loss ratio.

Differentiated Services (DiffServ) [81] provides this by dividing traffic into classes. Guarantees are only provided for a class and not an individual flow. Each class has a Per-Hop Behavior (PHB), which says how packets should be forwarded in the network. The default PHB is typically meant for best effort traffic while Expedited Forwarding (EF) [82] is meant for traffic that require low loss and delay, which makes it a suitable class for real-time audio and video. In addition the Assured Forwarding (AF) [83] class is defined. This class is further divided into four new groups, with different packet drop probabilities. When this differentiation is used, traffic will mostly be affected by traffic within the same class.

Given these two architectures it is important to evaluate the performance benefits from using the more advanced approach. The delay and jitter for voice traffic in DiffServ and BE environments are compared in [84], and it is found that DiffServ leads to better performance, especially for the codecs using a high bit rate. In paper F network simulations are used to generate packet loss traces for audio traffic in both BE and DiffServ environments. The differences between

these traces are evaluated using a subjective test, and it is shown that the packet loss process resulting from DiffServ is less bursty, and leads to a slightly higher perceived quality for lower packet loss ratios.

4.3 Packetization

Most audio coders process a stream in blocks and each block will usually result in one independently decodeable unit, but how these units are packetized are up to the sender. For MPEG-4 these units are known as an Access Unit (AU).

The Maximum Transmission Unit (MTU) is the maximum size of a packet that can be transmitted without being fragmented. For regular IP networks this limit is 1500 bytes, and a natural packetization choice is to buffer AUs until they reach the MTU and transmit them together. However, this approach has some disadvantages. Buffering packets will increase the total delay, since the sender has to wait before a packet is transmitted. Furthermore, when multiple consecutive AUs are contained within one packet, a single packet loss will result in a large gap in the audio stream (figure 4.1a). To combat this, it is possible to use a technique named interleaving [8]. Here, AUs are transmitted out of order, thus a packet loss will create multiple small gaps in the audio stream instead of a large one (figure 4.1b). Unfortunately, this will increase the system delay further, making it less suitable for real-time communication.

A different approach is to transmit each AU as soon as it has been processed. This introduces minimal delay, but it will increase the network overhead. The headers added by IP/UDP/RTP have a total size of 40 bytes, and with a small payload, e.g. 200 bytes, the packet efficiency is as low as 83%.

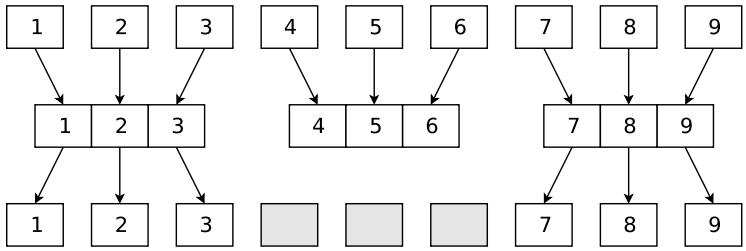
Paper F investigates different techniques for packetization and also different packet sizes. Network simulations show that a small packet size is beneficial when considering the packet loss ratio alone. However, it has to be considered whether or not this approach justifies the increased rate.

4.4 Error Protection and Concealment

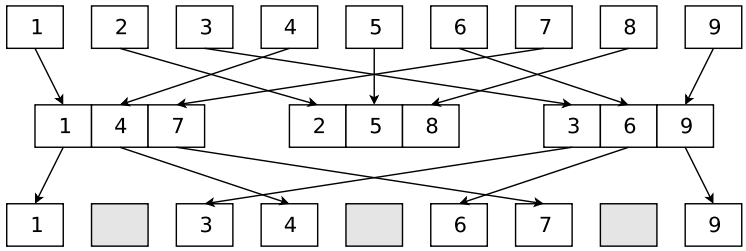
Packets transmitted over the internet may be lost during transmission and strict delay requirements make it impossible to wait for packet retransmissions. Error mitigation techniques can be divided into two categories, Error Concealment (EC) and Error Protection (EP), which are covered in the following sections.

4.4.1 Error Concealment

Error concealment is, as the name indicates, techniques that the receiver uses in order to conceal that any packets have been lost. If packet N is lost, the receiver cannot just jump directly from packet $N-1$ to $N+1$ since this will lead to time synchronization issues. The gap that is supposed to be filled with content



(a) Without interleaving: Three consecutive AUs are lost.



(b) Interleaving: Lost AUs are spread out.

Figure 4.1. Illustration of the consequences of packet loss both with and without interleaving. Figure taken from paper F.

from packet N, needs to be replaced with something. Various, relatively simple, EC techniques for audio signals are surveyed in [85]. The simplest solution, replacing lost packets with silence gives, not surprisingly, very poor results. What is interesting is that playing the previous received packet twice gives a rather large increase in perceived quality. If fading is used between the frame boundaries to remove discontinuities the quality is increased further.

However, many more complicated techniques have been proposed. In [86] time-scale modifications are used to "stretch" the next received packet to cover the gap from the lost packet. Several algorithms for interpolation in the frequency domain have also been proposed [87–92]. Yet another class of concealment is based on sinusoidal modeling of the lost content [93–95], and methods for smarter repetition of frames is investigated in [96, 97]. Also, a special error concealment method for the predictive low-delay codec in [45] is presented in [98].

In [99] a special EC-technique for multichannel audio is presented. This method is based on using other channels for interpolating the missing content, and packetizing the channels such that a single packet loss will only affect a subset of the channels.

4.4.2 Error Protection

Error protection is used as a term for techniques used by the sender to minimize the effect of packet loss, or to help the receiver with reconstructing missing data. Interleaving, as described in the previous section, can be considered an EP technique, since this type of packetization is meant to simplify the EC process for the receiver.

Another possibility is to exploit that certain parts of the compressed bit stream are more important than others. Some parts of the bit stream are crucial for decoding, while other parts can be predicted from previous packets [100–103]. By including the important parts in multiple packets, one is able to increase the probability that these important parts are received. One clear downside of this approach is that parts of the stream are transmitted in later packets which means that both the bit rate and delay are increased.

It is also possible to investigate the signal itself and transmit extra data to ease the reconstruction of selected parts of the signal [104, 105]. An example of this can be seen in [106]. Here, drumbeats (and other transients) are encoded and transmitted separately since transients can be very difficult to reconstruct.

A different alternative is the use of Multiple Description (MD) coding [107]. By using this scheme the encoded representation of a frame is divided into two parts. If only one of the parts is received, the decoder is able to reconstruct a low quality version of the frame. If both parts are received the highest quality is achieved. Examples of this can be seen in [108] where the low delay codec presented in [45] is encoded using multiple descriptions, and in [109] MD is used in a transform codec.

Paper A proposes a novel error protection scheme where the success of the error concealment is evaluated at the sender side using a perceptual criterion. By doing this the sender can determine whether or not a packet needs to be protected. In our implementation the error protection is simply a low bit rate representation of the same frame. This system is implemented for MPEG-4 AAC, but it should be noted that this system can be used for any type of encoder. The advantage of this method is that the receiver will only have to use EC on frames where this provides a reasonable audio quality, and since the remaining frames are protected the result is a system which is very robust against transmission errors.

Chapter 5

Evaluation of Audio Quality

Evaluating perceived audio quality can either be performed using objective or subjective tests. In a subjective test, participants are asked to rate the quality of the presented clip, either with or without listening to a reference. These tests are very time consuming, and they have to be carefully planned to avoid bias effects [110].

Objective tests, on the other hand, use a parametric approach where features of the audio signals, and possibly transmission systems, are used in order to estimate the quality experienced by the users. They are fast and easy to perform, but the mapping from an objective model to a subjective score is not trivial. A full-reference metric compares an original audio signal with the impaired version and calculates the quality, while a no-reference metric uses only the impaired signal to estimate the quality. Also completely parametric methods exist which do not need an audio signal at all. These metrics calculate the impairments from factors such as the chosen codec and transmission system properties like delay and loss characteristics.

For speech there has been much work on objective models, both for coding artifacts but also for general transmission systems, including packet based. For music, on the other hand, there has been some work on objective, full-reference metrics, but little work has been done for impairments resulting from packet based transmission.

5.1 Subjective Quality Evaluation

In a subjective quality evaluation, audio is played back to the listeners and subjects are asked to rate the quality. Depending on the application, a choice of test methodology has to be made, and numerous standards for this exist. There are also a large number of other problems involved, such as the room used, equipment, and also how the subjects are instructed [111].

For evaluating small impairments, typically resulting from coding artifacts,

ITU BS.1116 [112] is a common choice. Here the subject is presented with 3 stimuli, labeled as A, B, and reference. Either A or B should be identical to the reference, what is known as a hidden reference. The subject should assess the impairments of A and B relative to the reference based on a scale from 1 (very annoying) to 5 (imperceptible). However, since there is a hidden reference, at least one of the stimuli must be rated as imperceptible.

ITU BS.1534, Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) [113], is a test intended for evaluating larger distortions than BS.1116. With MUSHRA, the subject is presented with multiple test items simultaneously and one item labeled as reference. One of the test items should be a hidden reference, and there should at least be one band-limited anchor. All items should be rated from 0 (bad) to 100 (excellent). In paper F this standard is used to evaluate the perceived quality of packet loss distorted audio. Here, the focus is on evaluating different temporal loss distributions resulting from two different network architectures. This test was performed in order to investigate whether or not it is beneficial to use a more advanced architecture instead of the relatively simple model used today. It is found that the new architecture leads to a less bursty loss process which results in a slightly higher perceived quality.

In [114] a number of arguments against the ITU standards are presented. One of the key issues raised is the interpretation of the quality labels. It is shown that the perceived difference between the grades in the ITU-R 5-grade impairment scale are unequal and even more problems arise if the labels are translated into a different language. It is not necessarily true that results from tests using two different languages can be directly compared, since the translation can affect the interpretation of the quality labels. In [115] a so-called acceptability test is used to evaluate multimedia content for mobile devices. This test is a binary test, the subject is given a scenario and is asked to evaluate whether or not the quality of the test item is acceptable. This type of test is fast and easy to perform for the subjects. In paper G, an acceptability test is used to evaluate packet loss distorted music within an internet radio streaming scenario. Previous work has often used packet loss ratios as high as 10 – 15%. Here, much lower ratios are used in order to identify when the packet loss ratio leads to unacceptable quality. It is found that the audio quality is not acceptable when the stream is exposed to a packet loss ratio of 1.5%, and that the acceptability is severely reduced already with 0.5% packet loss. Also, it is shown that there is no difference between clips encoded at 64 and 128 kbps in situations with packet loss, which indicates that a bit rate reduction is a reasonable option during periods with packet loss.

5.2 Objective Quality Evaluation

The E-model [3, 116] is a completely parametric approach for evaluating narrow-band speech quality, originally intended for network planning. The E-model is an additive model, meaning that the distortions due to several factors are simply

added. Examples of parameters are the codec in use, the delay, signal to noise ratio, and the talker echo. By adding these impairment factors the transmission rating (R) results which can be converted into a Mean Opinion Score (MOS). This model has been created by performing a large number of subjective tests evaluating the different impairments, and then using these results to calculate the actual impairment factor. This model has later been extended to handle impairments from packet loss [117–119]. Here, a codec's ability to handle packet loss, the packet loss ratio, and also the burstiness of the packet loss process are identified as important parameters for the quality. Recently this model has undergone some changes in order to estimate impairments for wideband speech, including impairments from packet based transmission [120, 121]. The same method for estimating quality has been applied to video transmission as well [122, 123].

Perceptual Evaluation of Audio Quality (PEAQ) [124, 125] and Perceptual Evaluation of Speech Quality (PESQ) [126, 127] are full-reference metrics. These systems use a perceptual model to evaluate the differences between an impaired signal and the original audio signal. PESQ is designed to handle impairments resulting from packet losses. PEAQ has been designed to only evaluate coding distortions, which it does with a very high accuracy, especially when the results are averaged over several test clips [128]. An evaluation of how PEAQ handles packet loss distorted audio can be found in [7]. Development of objective quality evaluation for multichannel audio has also started, and preliminary results are presented in [129]. However, this system is only intended for sweet spot reproduction, not complete wave fields.

No-reference metrics and parametric approaches can easily be employed in operational networks since they only require the impaired signal or characteristics of the transmission medium. Full-reference metrics have traditionally been mostly used in laboratory settings since they require both the original and the impaired signal. However, as seen in paper A they can also be used on the sender side in a transmission system if the sender is able to simulate the possible impairments resulting from the transmission.

Chapter 6

Conclusions and Contributions

This thesis covers multiple aspects of audio transmission over IP networks. Papers A – E focus on sender side techniques to improve network transmission and papers F and G investigate aspects of the end user's perceived quality.

In paper A an Error Protection (EP) scheme for audio streaming is presented. The most important contribution here is the use of a perceptual criterion for determining whether an audio frame needs to be protected or not. In earlier work EP is used to assist the receiver with reconstructing parts of the signal, but here it is actually investigated whether or not the receiver can reconstruct the signal satisfactorily without any assistance. In this work the perceptual analysis is performed in a separate module, but this analysis could also be incorporated into the perceptual model in the encoder. The implementation used in this paper introduces additional delay. However, it can easily be modified in order to avoid this, thus making it suitable for real-time communication. Although AAC is used in this implementation, this approach is independent of the chosen codec. The amount of redundancy added to the packet stream is determined by the quality criterion selected. Depending on the type of audio signal, the increase in quality can be very significant using only small amounts of redundancy.

Paper B presents a layered codec which has a delay of only 256 samples, and when both layers are received the original lossless signal is recovered. This codec is intended for a network supporting service differentiation in order to increase the probability for receiving the base layer. The base layer is a perceptually encoded representation of the original signal, while the enhancement layer contains the difference between the original signal and the base layer. The base layer is compressed using a well known codec, but the extension to a low delay lossless codec made in this work is novel. In situations where the enhancement layer is lost, the base layer can be used alone and still result in a relatively high perceived quality. Even though this codec is restricted to have a very low delay the resulting bit rate is on average only 56% of the original bit rate, for the base and enhancement layer combined. Also, it is shown that the layering introduces

a very low overhead when compared with regular lossless compression.

Papers C – E deal with the compression and transmission of Higher Order Ambisonics (HOA). In paper C it is investigated how compressing the cylindrical harmonics components with AAC affects the reproduced sound field. To our knowledge, this is the first study within this topic. It is found that encoding HOA with AAC results in an error that is very low in the sweet spot, but it increases as a function of the distance from the center. How steep this increase is, is dependent on the spectrum of the content. It is also found that it is beneficial to have an unequal distribution of bits between the orders in the cylindrical harmonics domain. By using most bits on the lowest orders, good sound quality is achieved and the spatial information is still improved by including the higher order channels.

In paper D the spatial redundancy for two-dimensional synthesized HOA-signals is investigated. A low delay, lossless compression scheme that exploits the inter-channel correlations is presented. Previous work on lossless multichannel compression does not focus on the coding delay which means that they are not suitable for real-time communication. In addition to a very low delay for encoding and decoding, the presented codec preserves the hierarchical HOA-format in order to keep the stream scalable and to allow reproduction over arbitrary loudspeaker layouts. It is found that for synthetic signals, without any added reverberation, the bit rate is significantly reduced when the inter-channel redundancy is removed. However, this reduction is highly dependent on the number of sources and also their locations.

The work from paper D is continued in paper E. Here, the same compression system is applied on signals from a newly developed spherical microphone array, the "Eigenmike"®. It is found that the microphone signals are highly correlated and the rate is typically reduced with more than 10% when the inter-channel correlation is exploited. When the signals are converted into the HOA domain, the rate reduction is severely reduced. However, this does not mean that these signals are not correlated, but rather that the correlation is more difficult to exploit. The reason for this is that the transformation into the spherical harmonics domain typically results in many channels with very low energy, thus they are easily compressed by only removing the intra-channel correlation. Several options for reducing the complexity of this encoder without increasing the bit rate significantly are also investigated.

The perceived quality experienced by the end user is evaluated in papers F and G. These subjective tests are the first steps towards developing objective metrics for estimating the perceived quality. In the first paper, network simulations are used to investigate the importance of packet sizes and network architecture for audio streaming. It is found that the packet loss ratio is significantly lower with smaller packets, but the downside is that the total rate also increases due to the number of extra packets transmitted. The network simulations were also used to create models for the packet loss processes resulting from the BE and DiffServ architectures. DiffServ is an architecture with substantially higher

complexity than BE, and it is important to evaluate the advantages from using this architecture and at the same time investigate how different temporal loss distributions affect the perceived quality. From the subjective test, it is found that the less bursty DiffServ loss distribution leads to higher perceived quality at a low packet loss ratio.

In paper G the acceptability of packet loss distorted audio is evaluated using a very fast and easy subjective test. This test is important since it actually reveals at which loss rates the subjects finds the quality too low to be acceptable. It is seen that the audio quality can no longer be said to be acceptable when the packet loss ratio is 1.5%. Averaged over all clips the acceptability is less than 40%. The difference between stereo signals encoded with AAC at 64 and 128 kbps is also evaluated, and in situations with packet loss there are no significant differences between these encoding rates.

References

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

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

[3] ITU-T Recommendation G.107, "The E-Model, a Computational Model for Use in Transmission Planning," 2003.

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

[5] S. Farner, A. Sæbø, A. Solvang, and U.P. Svensson, "Ensemble Hand-Clapping Experiments under the Influence of Delay and Various Acoustic Enviroments," in *The 121st AES Conv.*, October 2006, Preprint 6905.

[6] M. Lutzky, M. Gayer, G. Schuller, U. Kraemer, and S. Wabnik, "A Guideline to Audio Codec Delay," in *The 116th AES Conv.*, May 2004, Preprint 6062.

[7] J.E. Voldhaug, E. Hellerud, and U.P. Svensson, "Evaluation of Packet Loss Distortion in Audio Signals," in *The 120th AES Conv.*, May 2006, Preprint 6855.

[8] B. Grill, T. Hahn, D. Homm, K. Krauss, G. Ohler, W. Sörgel, and C. Spitzner, "Characteristics of Audio Streaming over IP Networks within the ISMA Standard," in *The 112th AES Conv.*, April 2002, Preprint 5514.

[9] A.J. Berkhout, "A Holographic Approach to Acoustic Control," *J. Audio Eng. Soc*, vol. 36, no. 12, pp. 977–995, December 1988.

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

[11] S.N. Goodwin, "3D Sound for 3D Games – Beyond 5.1," in *The 35th AES International Conference*, 2009.

[12] M.A. Gerzon, "Periphony: With-Height Sound Reproduction," *J. Audio Eng. Soc*, vol. 21, no. 1, pp. 2–10, February 1973.

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

[14] 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.

[15] J. Daniel and S. Moreau, "Further Study of Sound Field Coding with Higher Order Ambisonics," in *The 116th AES Conv.*, May 2004, Preprint 6017.

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

[17] A. Solvang, "Spectral Impairment of Two-Dimensional Higher Order Ambisonics," *J. Audio Eng. Soc*, vol. 56, no. 4, pp. 267–279, April 2008.

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

[19] M.A. Gerzon, "Hierarchical Transmission System for Multispeaker Stereo," *J. Audio Eng. Soc*, vol. 40, no. 9, pp. 692–705, September 1992.

[20] R. Nicol and M. Emerit, "3D-Sound Reproduction Over an Extensive Listening Area: A Hybrid Method Derived from Holophony and Ambisonic," in *Proc. of The AES 16th International Conference: Spatial Sound Reproduction*, March 1999, Paper Number 16-039.

[21] M.A. Gerzon, "The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound," in *The 50th AES Conv.*, March 1975, Preprint L-20.

[22] 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 (ICASSP02)*, 2002.

[23] S. Moreau, S. Bertet, and J. Daniel, "3D Sound Field Recording with Higher Order Ambisonics - Objective Measurements and Validation of Spherical Microphone," in *The 120th AES Conv.*, May 2006, Preprint 6857.

[24] mh acoustics, "em32 Eigenmike microphone array," `http://mhacoustics.com/page/page/2949006.htm`, [Online].

[25] 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.

[26] 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.

[27] H. Fletcher, "Auditory Patterns," *Modern Physics*, vol. 12, pp. 47–66, 1940.

[28] B. Moore, *An Introduction to the Psychology of Hearing*, Emerald Group Publishing Ltd, 5 edition, Jan. 2003.

[29] E. Terhardt, "Calculating Virtual Pitch," *Hearing Research*, vol. 1, pp. 155–182, 1979.

[30] B. Moore, "Masking in the Human Auditory System," in *Collected Papers on Digital Audio Bit-Rate Reduction (AES)*, May 1996.

[31] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer-Verlag, 1990.

[32] K. Brandenburg and G. Stoll, "ISO/MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio," *J. Audio Eng. Soc*, vol. 42, no. 10, pp. 780–792, October 1994.

[33] K. Brandenburg and T. Sporer, "-NMR- and -Masking Flag-: Evaluation of Quality Using Perceptual Criteria," in *Proc. of the 11th AES International Conf.*, April 1992.

[34] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding*, Wiley-Interscience, 2007.

[35] B. Edler, H. Purnhagen, and C. Ferekidis, "ASAC – Analysis/Synthesis Audio Codec for Very Low-Bit Rates," in *The 100th AES Conv.*, May 1996, Preprint 4179.

[36] H. Purnhagen and N. Meine, "HILN – the MPEG-4 Parametric Audio Coding Tools," in *IEEE International Symposium on Circuits and Systems (ISCS00)*, 2000.

[37] M. Deriche and D. Ning, "A Novel Audio Coding Scheme Using Warped Linear Prediction Model and the Discrete Wavelet Transform," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2039–2048, 2006.

[38] ITU-T Recommendation G.729, "Coding of Speech at 8 kbit/s Using Conjugate-structure Algebraic-code-excited Linear Prediction (CS-ACELP)," 2007.

[39] ITU-T Recommendation G.722, "7 kHz Audio-coding within 64 kbit/s," 1988.

[40] A. Harma, U.K. Laine, and M. Karjalainen, "WLPAC – A Perceptual Audio Codec in a Nutshell," in *The 102nd AES Conv.*, February 1997, Preprint 4420.

[41] B. Edler and G. Schuller, "Audio Coding Using a Psychoacoustic Pre- and Post-filter," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP00)*, 2000.

[42] B. Edler, C. Faller, and G. Schuller, "Perceptual Audio Coding Using a Time-Varying Linear Pre- and Post-Filter," in *The 109th AES Conv.*, August 2000, Preprint 5274.

[43] S. Dorward, D. Huang, S.A. Savari, G. Schuller, and B. Yu, "Low Delay Perceptually Lossless Coding of Audio Signals," in *Data Compression Conference (DCC01)*, 2001.

[44] G. Schuller and A. Harma, "Low Delay Audio Compression Using Predictive Coding," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP02)*, 2002.

[45] 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.

[46] J. Princen and A. Bradley, "Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation," *IEEE Trans. Speech Audio Process.*, vol. 34, no. 5, pp. 1153–1161, 1986.

[47] S. Shlien, "Guide to MPEG-1 Audio Standard," *IEEE Trans. Broadcast.*, vol. 40, no. 4, pp. 206–218, 1994.

[48] ISO/IEC 14496-3, "Coding of Audio-visual Objects – Part 3: Audio," 1998.

[49] M. Bosi, K. Brandenburg, S. Quackenbush, S. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, "ISO/IEC MPEG-2 Advanced Audio Coding," *J. Audio Eng. Soc*, vol. 45, no. 10, pp. 789–814, October 1997.

[50] B. Grill, "The MPEG-4 General Audio Coder," in *Proc. of the 17th AES Conf.*, August 1999.

[51] E. Allamenche, R. Geiger, J. Herre, and T. Sporer, "MPEG-4 Low Delay Audio Coding based on the AAC Codec," in *The 106th AES Conv.*, April 1999, Preprint 4929.

[52] M. Schnell, M. Schmidt, M. Jander, T. Albert, R. Geiger, V. Ruoppila, P. Ekstrand, M.Lutzky, and B. Grill, "MPEG-4 Enhanced Low Delay AAC - A New Standard for High Quality Communication," in *The 125th AES Conv.*, 2008, Preprint 7503.

[53] 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.

[54] E. Larsen, R.M. Aarts, and M. Danessis, "Efficient High-frequency Bandwidth Extension of Music and Speech," in *The 112th AES Conv.*, 2002, Preprint 5627.

[55] S. Meltzer, R. Bohm, and F. Henn, "SBR Enhanced Audio Codecs for Digital Broadcasting Such as "Digital Radio Mondiale" (DRM)," in *The 112th AES Conv.*, 2002, Preprint 5559.

[56] T. Ziegler, A. Ehret, P. Ekstrand, and M. Lutzky, "Enhancing MP3 with SBR: Features and Capabilities of the New MP3PRO Algorithm," in *The 112th AES Conv.*, 2002, Preprint 5560.

[57] 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.

[58] N.H. v. Schijndel, J. Bensa, M.G. Christensen, C. Colomes, B. Edler, R. Heusdens, J. Jensen, S.H. Jensen, W. B. Kleijn, V. Kot, B. Kövesi, J. Lindblom, D. Massaloux, O.A. Niamut, F. Nordén, J.H. Plasberg, R. Vafin, S.v.d. Par, D. Virette, and O. Wübbolt, "Adaptive RD Optimized Hybrid Sound Coding," *J. Audio Eng. Soc*, vol. 56, no. 10, pp. 787–809, October 2008.

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

[60] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, and K.S. Chong, "MPEG Surround – The ISO/MPEG Standard for Efficient and Compatible Multichannel Audio Coding," *J. Audio Eng. Soc*, vol. 56, no. 11, pp. 932–955, November 2008.

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

[62] S.M.F. Smyth, W.P. Smith, M.H.C. Smyth, M. Yan, and T. Jung, "DTS Coherent Acoustics Delivering High-Quality Multichannel Sound to the Consumer," in *The 100th AES Conv.*, May 1996, Preprint 4293.

[63] R.L. Andersen, B.G. Crockett, G.A. Davidson, M.F. Davis, L.D. Fielder, S.C. Turner, M.S. Vinton, and P.A. Williams, "Introduction to Dolby Digital Plus, an Enhancement to the Dolby Digital Coding System," in *The 117th AES Conv.*, October 2004, Preprint 6196.

[64] Y. Jiao, S. Zielinski, and F. Rumsey, "Hierarchical Bandwidth Limitations of Surround Sound – Part I: Psychoacoustically Hierarchical Transform," *J. Audio Eng. Soc*, vol. 56, no. 12, pp. 1057–1068, December 2008.

[65] Y. Jiao, S. Zielinski, and F. Rumsey, "Hierarchical Bandwidth Limitation of Surround Sound—Part II: Optimization of Bandwidth Allocation Strategy," *J. Audio Eng. Soc*, vol. 57, no. 1/2, pp. 5–15, January 2009.

[66] gzip, "`http://gzip.org`," [Online].

[67] bzip, "`http://bzip.org`," [Online].

[68] A. Robinson, "SHORTEN: Simple Lossless and Near-Lossless Waveform Compression," Technical report CUED/F-INFENG/TR.156, Cambridge University Engineering Department, 1994.

[69] Free Lossless Audio Codec, "FLAC," `http://flac.sourceforge.net/`, [Online].

[70] Monkey's Audio, "`http://www.monkeysaudio.com/`," [Online].

[71] R. Geiger, J. Herre, J. Koller, and K. Brandenburg, "IntMDCT - A Link Between Perceptual and Lossless Audio Coding," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP02)*, 2002.

[72] S. Golomb, "Run-length Encodings (Corresp.)," *IEEE Trans. Inf. Theory*, vol. 12, no. 3, pp. 399–401, 1966.

[73] A.A.M.L. Bruekers, W. Oomen, R.J.v.d. Vleuten, and L.M.v.d. Kerkhof, "Lossless Coding for DVD Audio," in *The 101st AES Conv.*, November 1996, Preprint 4358.

[74] A. Tanenbaum, *Computer Networks*, Prentice Hall Professional Technical Reference, 2002.

[75] J. Postel, "Internet Protocol," RFC 791, Sept. 1981.

[76] J. Postel, "Transmission Control Protocol," RFC 793, Sept. 1981.

[77] J. Postel, "User Datagram Protocol," RFC 768, Aug. 1980.

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

[79] E. Kohler, M. Handley, and S. Floyd, "Datagram Congestion Control Protocol (DCCP)," RFC 4340, Mar. 2006.

[80] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," RFC 3550, July 2003.

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

[82] B. Davie, A. Charny, J.C.R. Bennet, K. Benson, J.Y. Le Boudec, W. Courtney, S. Davari, V. Firoiu, and D. Stiliadis, "An Expedited Forwarding PHB (Per-Hop Behavior)," RFC 3246, Mar. 2002.

[83] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, "Assured Forwarding PHB Group," RFC 2597, June 1999.

[84] J.K. Muppala, T. Bancherdvanich, and A. Tyagi, "VoIP Performance on Differentiated Services Enabled Network," in *IEEE International Conference on Networks (ICON00)*, 2000.

[85] C. Perkins, O. Hodson, and V. Hardman, "A Survey of Packet Loss Recovery Techniques for Streaming Audio," *IEEE Network*, vol. 12, no. 5, pp. 40–48, 1998.

[86] H. Sanneck, A. Stenger, K. Ben Younes, and B. Girod, "A New Technique for Audio Packet Loss Concealment," in *Global Telecommunications Conference (GLOBECOM96)*, 1996.

[87] E. Kurniawati, E. Kurniawan, C.T. Lau, B. Premkumar, J. Absar, and S. George, "Error Concealment Scheme for MPEG-AAC," in *The Ninth International Conference on Communications Systems (ICCS04)*, 2004.

[88] H. Ofir and D. Malah, "Packet Loss Concealment for Audio Streaming Based on the GAPES Algorithm," in *The 118th AES Conv.*, May 2005, Preprint 6334.

[89] S-U. Ryu and K. Rose, "A Frame Loss Concealment Technique for MPEG-AAC," in *The 120th AES Conv.*, 2006, Preprint 6662.

[90] P. Lauber and R. Sperschneider, "Error Concealment for Compressed Digital Audio," in *The 111th AES Conv.*, November 2001, Preprint 5460.

[91] S. Quackenbush and P.F. Driessen, "Error Mitigation in MPEG-4 Audio Packet Communication Systems," in *The 115th AES Conv.*, September 2003, Preprint 5981.

[92] A. Floros, M. Avlonitis, and P. Vlamos, "Frequency-Domain Stochastic Error Concealment for Wireless Audio Applications," *Mob. Netw. Appl.*, vol. 13, no. 3-4, pp. 357–365, 2008.

[93] J. Lindblom and P. Hedelin, "Packet Loss Concealment Based on Sinusoidal Extrapolation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP02)*, 2002.

[94] S-U. Ryu and K. Rose, "Advances in Sinusoidal Analysis/Synthesis-based Error Concealment in Audio Networking," in *The 116th AES Conv.*, May 2004, Preprint 5997.

[95] H. Hou and W. Dou, "Real-Time Audio Error Concealment Method Based on Sinusoidal Model," in *International Conference on Audio, Language and Image Processing (ICALIP08)*, 2008.

[96] W.K. Ng, J. Absar, S. George, C.T. Lau, and B. Premkumar, "An Enhanced Smart Copying (ESC) Method for the Reconstruction of Missing Audio Packets," in *IEEE International Conference on Communications, Circuits and Systems (ICCCS02)*, 2002.

[97] M. Niediwiecki and K. Cisowski, "Smart Copying – A New Approach to Reconstruction of Audio Signals," *IEEE Trans. Speech Audio Process.*, vol. 49, no. 10, pp. 2272–2282, 2001.

[98] S. Wabnik, G. Schuller, J. Hirschfeld, and U. Kraemer, "Packet Loss Concealment in Predictive Audio Coding," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA05)*, 2005.

[99] R. Sinha, C. Papadopoulos, and C. Kyriakakis, "Loss Concealment for Multi-Channel Streaming Audio," in *International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV03)*, 2003.

[100] J. Korhonen, "Error Robustness Scheme for Perceptually Coded Audio based on Interframe Shuffling of Samples," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP02)*, 2002.

[101] J. Korhonen and Y. Wang, "Schemes for Error Resilient Streaming of Perceptually Coded Audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP03)*, 2003.

[102] J. Korhonen, Y. Wang, and D. Isherwood, "Toward Bandwidth-Efficient and Error-Robust Audio Streaming over Lossy Packet Networks," *Multimedia Systems*, vol. 10, no. 5, pp. 402–412, August 2005.

[103] Y. Wang, W. Huang, and J. Korhonen, "A Framework for Robust and Scalable Audio Streaming," in *Proceedings of the 12th ACM international conference on Multimedia (Multimedia04)*, 2004.

[104] S-U. Ryu, E. Choy, and K. Rose, "Encoder Assisted Frame Loss Concealment for MPEG-AAC Decoder," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP06)*, 2006.

[105] M. Chen and M.N. Murthi, "Optimized Unequal Error Protection for Voice over IP," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP04).*, 2004.

[106] Y. Wang and S. Streich, "A drumbeat-pattern based error concealment method for music streaming applications," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP02)*, 2002.

[107] V.K. Goyal, "Multiple Description Coding: Compression Meets the Network," *IEEE Sig. Proc. Magazine*, vol. 18, no. 5, pp. 74–93, 2001.

[108] G. Schuller, J. Kovacevic, F. Masson, and V.K. Goyal, "Robust Low-Delay Audio Coding Using Multiple Descriptions," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1014–1024, 2005.

[109] R. Arean, J. Kovacevic, and V.K. Goyal, "Multiple Description Perceptual Audio Coding with Correlating Transforms," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 140–145, 2000.

[110] S. Zielinski, F. Rumsey, and S. Bech, "On Some Biases Encountered in Modern Audio Quality Listening Tests – A Review," *J. Audio Eng. Soc*, vol. 56, no. 6, pp. 427–451, June 2008.

[111] S. Bech and N. Zacharov, *Perceptual Audio Evaluation - Theory, Method and Application*, Wiley, 2006.

[112] ITU-R Recommendation BS.1116, "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," 1997.

[113] ITU-R Recommendation BS.1534, "Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems," 2003.

[114] A. Watson and M.A. Sasse, "Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications," in *Proceedings of the sixth ACM international conference on Multimedia (Multimedia98)*, 1998.

[115] M.A. Sasse and H. Knoche, "Quality in Context - An Ecological Approach to Assessing QoS for Mobile TV," in *ISCA Tutorial and Research Workshop on Perceptual Quality of Systems*, Sep 2006.

[116] S. Möller, *Assesment and Prediction of Telephone-Speech Quality*, Kluwer Academic Publishers, 2000.

[117] A. Raake, *Speech Quality of VoIP: Assessment and Prediction*, Wiley, 2006.

[118] A. Raake, "Short- and Long-Term Packet Loss Behavior: Towards Speech Quality Prediction for Arbitrary Loss Distributions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1957–1968, 2006.

[119] A. Raake, "Predicting Speech Quality under Random Packet Loss: Individual Impairment and Additivity with other Network Impairments," *Acta Acustica united with Acustica*, vol. 90, no. 6, pp. 1061–1083, Nov/Dec 2004.

[120] S. Möller, A. Raake, N. Kitawaki, A. Takahashi, and M. Waltermann, "Impairment Factor Framework for Wide-Band Speech Codecs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1969–1976, 2006.

[121] M. Waltermann and A. Raake, "Towards a New E-model Impairment Factor for Linear Distortion of Narrowband and Wideband Speech Transmission," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP08)*, 2008.

[122] A. Raake, M.N. Garcia, S. Möller, J. Berger, F. Kling, P. List, J. Johann, and C. Heidemann, "T-V-model: Parameter-based Prediction of IPTV Quality," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP08)*, 2008.

[123] M.N. Garcia, A. Raake, and P. List, "Towards content-related features for parametric video quality prediction of IPTV services," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP08)*, 2008.

[124] ITU-R Recommendation BS.1387-1, "Method for Objective Measurements of Perceived Audio Quality (PEAQ)," 2001.

[125] T. Thiede, W.C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J.G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, "PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality," *J. Audio Eng. Soc.*, vol. 48, no. 1/2, pp. 3–29, March 2000.

[126] ITU-T Recommendation P.862, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs," 2001.

[127] A.W. Rix, M.P. Hollier, J.G. Beerends, and A.P. Hekstra, "PESQ – The New ITU Standard for End-to-End Speech Quality Assessment," in *The 109th AES Conv.*, September 2000, Preprint 5260.

[128] W.C. Treurniet and G.A. Soulodre, "Evaluation of the ITU-R Objective Audio Quality Measurement Method," *J. Audio Eng. Soc*, vol. 48, no. 3, pp. 164–173, March 2000.

[129] I. Choi, B.G. Shinn-Cunningham, S.B. Chon, and K-M. Sung, "Objective Measurement of Perceived Auditory Quality in Multichannel Audio Compression Coding Systems," *J. Audio Eng. Soc*, vol. 56, no. 1/2, pp. 3–17, January 2008.

Part II

Included Papers

Paper A

Perceptually Controlled Error Protection for Audio Streaming over IP Networks

Erik Hellerud, Jan Erik Voldhaug, and U. Peter Svensson

In proceedings of the International Conference on Digital Telecommunications (ICDT06), Cap Esterel, France, August 2006.

The thesis author had the original idea for this paper. The co-authors participated in scientific discussions.

Is not included due to copyright

Paper B

Robust Transmission of Lossless Audio with Low Delay over IP Networks

Erik Hellerud and U. Peter Svensson

In proceedings of the International Symposium on Signal Processing and Information Technology (ISSPIT07), Cairo, Egypt, December 2007.

The thesis author had the original idea for this paper. The co-author participated in scientific discussions.

Is not included due to copyright

Paper C

Encoding Higher Order Ambisonics with AAC

Erik Hellerud, Ian Burnett, Audun Solvang, and U. Peter Svensson

In proceedings of the 124th AES Convention, Amsterdam, The Netherlands, May 2008. Preprint 7366.

The second author proposed to use AAC to compress Higher Order Ambisonics. The thesis author implemented the system, analyzed the results, and wrote the paper. The third and fourth authors participated in scientific discussions.

Encoding Higher Order Ambisonics with AAC

Erik Hellerud, Ian Burnett, Audun Solvang, and U. Peter Svensson

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.

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]:

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

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

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

where $K = 2N + 1$, M is the number of loudspeakers, and ϕ_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

$$N = kr \tag{3}$$

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

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.

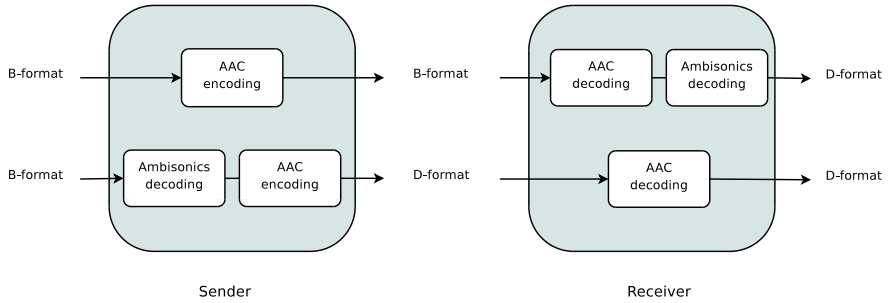


Figure 1. Encoding schemes: Upper signal path shows encoding the B-format signals, lower shows encoding the D-format signals (loudspeaker signals).

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:

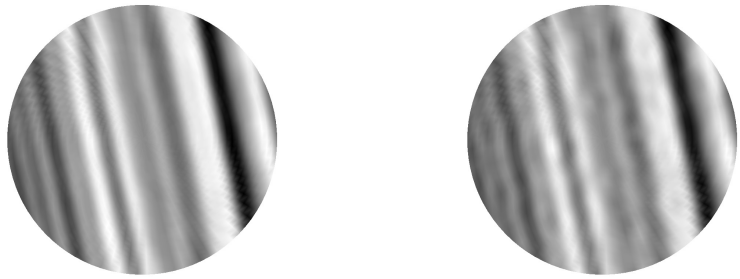
$$p(r,k,\theta) = \sum_{m=1}^{M} e^{jkr\cos(\theta_m-\theta)} l_m(0,k,\theta_m), \tag{4}$$

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

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

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

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.

Figure 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 that the wave becomes more blurry,

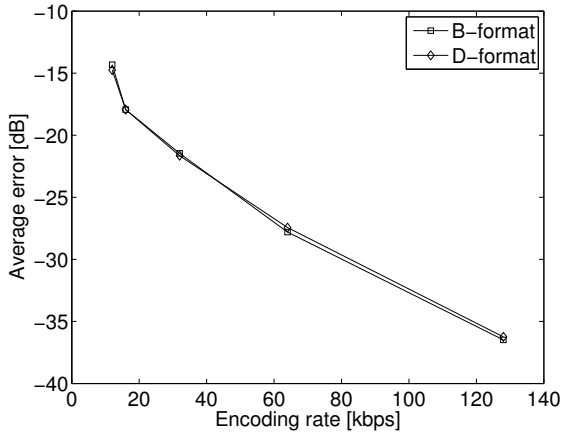


Figure 3. Error averaged across the entire reproduction area. The HOA order was 7, and all channels were encoded with the same rate.

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.

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 a 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

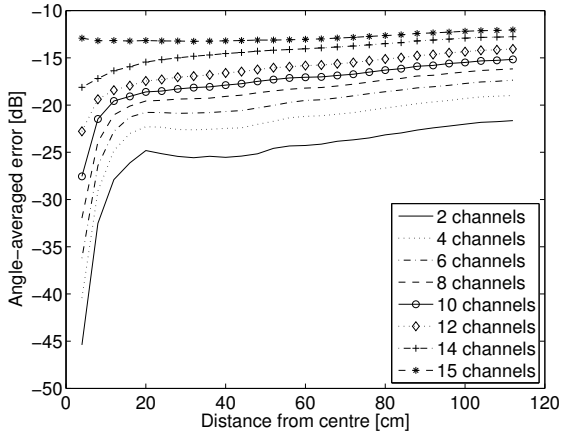


Figure 4. Angle-averaged error as a function of distance from the centre for a clip containing four sources. The HOA order was 7, and a subset of the B-format signals were encoded to 12 kbps. "2 channels" means that channels 14-15 were compressed to 12 kbps, "4 channels" means that 12-15 were compressed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

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

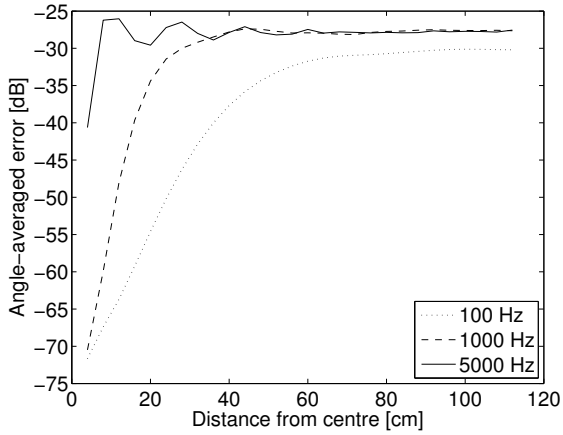


Figure 5. Angle-averaged quantization error as a function of distance and frequency. A single sinusoidal with frequency 100, 1000, or 5000 Hz were reproduced with order 7.

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. Støfringsdal 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," *IEEE Multimedia*, 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, Preprint 7370.

[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, June 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 (ICASSP02)*, 2002.

[10] D.B. Ward and T.D. Abhayapala, "Reproduction of a Plane-wave Sound Field Using an Array of Loudspeakers," *IEEE Trans. Speech Audio Process.*, 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].

Paper D

Spatial Redundancy in Higher Order Ambisonics and its Use for Low Delay Lossless Compression

Erik Hellerud, Audun Solvang, and U. Peter Svensson

In proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP09), Taipei, Taiwan, April 2009.

The thesis author had the original idea for this paper. The co-authors participated in scientific discussions.

Is not included due to copyright

Paper E

Lossless Compression of Spherical Microphone Array Recordings

Erik Hellerud and U. Peter Svensson

In proceedings of the 126th AES Convention, Munich, Germany, May 2009. Preprint 7668.

The thesis author had the original idea for this paper. The co-author participated in scientific discussions.

Lossless Compression of Spherical Microphone Array Recordings

Erik Hellerud and U. Peter Svensson

Abstract

The amount of spatial redundancy for recordings from a spherical microphone array is evaluated using a low delay lossless compression scheme. The original microphone signals, as well as signals transformed to the spherical harmonics domain, are investigated. It is found that the correlation between channels is, as expected, very high for the microphone signals, in two different acoustical environments. For the signals in the spherical harmonics domain, the compression gain from using inter-channel prediction is reduced, since this conversion results in many channels with low energy. Several alternatives for reducing the coding complexity are also investigated.

1. Introduction

The "Eigenmike" [1,2] is a spherical microphone array with a total of 32 elements. A recording from this microphone can be used to recreate the complete three dimensional sound field using techniques such as Higher Order Ambisonics (HOA) [3, 4]. When HOA is used, the 32 microphone signals have to be transformed into the spherical harmonics domain. Given the number of elements on this microphone, the maximum possible order in the spherical harmonics domain is four [5], which results in a total of 25 channels.

The transmission of 25 channels over a network can be problematic since it will create a very high data rate. No bandwidth guarantees are given with the current internet and the available bandwidth may also vary over time. One advantage with HOA under these conditions is that the format is naturally scalable. This means that the sender can adapt the transmission rate to the current network conditions. Down scaling a HOA signal will not reduce the basic audio quality, only the spatial resolution is affected [6].

For delay-sensitive applications such as distributed music plays, where musicians are located at geographically different locations, and also video conferencing, it is important to minimize the end to end delay in order to maximize the Quality of Experience (QoE). It is advised that the delay is kept below 150 ms for voice applications [7], but for music applications this limit is as low as 60 ms [8]. The network delay is difficult to control, but the delay introduced by the encoder and decoder can often contribute significantly to the total delay [9]. Regular audio codecs will typically use rather large blocks of samples to increase the accuracy of the perceptual model and to increase the compression gain, but this will also increase the delay. Several perceptual codecs with low delay have been presented [10, 11], however in situations where the maximum quality is desired, lossless compression is a preferred choice.

The microphone elements on the "Eigenmike" are located very closely, the radius of the microphone is only 42 mm. This should result in highly correlated channels, and this redundancy can be removed in order to achieve more efficient transmission. However, the correlation will depend on the number of sources and the room characteristics. Signals converted to the spherical harmonics domain may also be highly correlated, but for the HOA signals the correlation will also depend on the number of orders in the spherical harmonics domain.

Section 2 gives an introduction to HOA and how the signals from the "Eigenmike" can be converted into the spherical harmonics domain and then to loudspeaker signals. In section 3 the lossless compression scheme is presented, and section 4 explains how this format should be transmitted. Sections 5 and 6 provides the results for synthetic signals and microphone signals, respectively, before the concluding remarks are given in section 7.

2. Higher Order Ambisonics

The complete theory behind HOA is thoroughly described in [3–5] so only a short introduction will be given here. The theory is based on that a complete sound field can be described by spherical harmonics Y_{mn}^{σ} [12], which here are described as:

$$\begin{aligned} Y_{mn}^{\sigma}(\theta,\varphi) =& \sqrt{2m+1}\sqrt{(2-\delta_n)\frac{(m-n)!}{(m+n)!}} \\ & \cdot P_{mn}(sin(\varphi)) \begin{cases} \cos(n\theta), & \sigma = +1 \\ \sin(n\theta), & \sigma = -1 \end{cases}, \end{aligned} \tag{1}$$

where δ_n is the Kronecker delta function and P_{mn} is the Legendre function of degree m and order n.

The sound pressure at a point in space can then be written as

$$p(r,\theta,\varphi) = \sum_{m=0}^{\infty} j^m j_m(kr) \sum_{\substack{0 \leqslant n \leqslant m \\ \sigma = \pm 1}} B_{mn}^{\sigma} Y_{mn}^{\sigma}, \tag{2}$$

where j is the imaginary unit, $j_m(kr)$ is the spherical Bessel function of order m, k is the wave number defined as $k = \frac{2\pi}{\lambda}$ and B_{mn}^{σ} are the so-called B-format coefficients. If m is truncated to M we have a representation of order M which, for a complete three dimensional sound field, will lead to a total of $(M+1)^2$ channels. If the reproduction is restricted to two dimensions, only $2M+1$ channels are needed.

2.1. HOA Encoding

The B-format signals are extracted from the microphone signals using a matrix multiplication followed by a filtering of the signals. The parameters of the matrix are given by the locations of the microphone elements on the sphere. Given a microphone with Q elements in positions (θ_q, φ_q), the HOA order M is restricted by $Q \geqslant (M+1)^2$. The encoding matrix $\mathbf{E}$ can then be found as $\mathbf{E} = (\mathbf{Y}^t \cdot \mathbf{Y})^{-1}\mathbf{Y}^t$, where

$$\mathbf{Y} = \begin{bmatrix} Y_{00}^1(\theta_1, \varphi_1) & \cdots & Y_{MM}^{-1}(\theta_1, \varphi_1) \\ \vdots & \ddots & \vdots \\ Y_{00}^1(\theta_Q, \varphi_Q) & \cdots & Y_{MM}^{-1}(\theta_Q, \varphi_Q) \end{bmatrix}. \tag{3}$$

After the conversion an equalization of the HOA signals is also required, which is described in detail in [5].

2.2. HOA Decoding

Decoding the HOA signals to loudspeaker signals is also based on a matrix multiplication. Given a three dimensional reproduction and order M, the number of loudspeakers N in positions (θ_n, φ_n) must be such that $N \geqslant (M+1)^2$. The decoding matrix is given as $\mathbf{D} = (\mathbf{C}^t \cdot \mathbf{C})^{-1}\mathbf{C}^t$ where

$$\mathbf{C} = \begin{bmatrix} Y_{00}^1(\theta_1, \varphi_1) & \cdots & Y_{00}^1(\theta_N, \varphi_N) \\ Y_{11}^1(\theta_1, \varphi_1) & \cdots & Y_{11}^1(\theta_N, \varphi_N) \\ Y_{11}^{-1}(\theta_1, \varphi_1) & \cdots & Y_{11}^{-1}(\theta_N, \varphi_N) \\ \vdots & \ddots & \vdots \\ Y_{MM}^{-1}(\theta_1, \varphi_1) & \cdots & Y_{MM}^{-1}(\theta_N, \varphi_N) \end{bmatrix}. \tag{4}$$

The radius of the volume with correct reproduction depends on both frequency and order, and is given by $r = \frac{M}{k}$ [13].

As seen from the formulas, only the decoding matrix is a function of the loudspeaker positions, which means that the receiver can have an arbitrary loudspeaker layout. Also, the number of channels transmitted can easily be adapted to the receivers capability since reducing the order will only reduce the area or volume of reproduction.

3. Low Delay Lossless Compression

One of the most important factors for the encoding delay is the size of the processed blocks [9]. Most regular audio coders typically use quite large block sizes. For perceptual audio coders this is important in order to increase the frequency resolution, but it will also increase the efficiency of the entropy coder and reduce the total rate.

The lossless codec used in this work is based on backwards adaptive prediction and presented earlier in [14]. The advantage with adaptive prediction compared

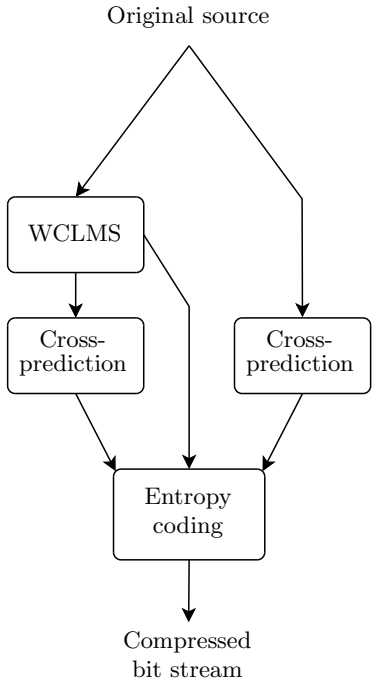


Figure 1. Illustration of the coding scheme.

to regular Linear Predictive Coding (LPC) is that it is possible to use a very short block size, but still have rather long prediction filters. The disadvantage is that the prediction accuracy may be reduced and thus the total rate may become higher than for LPC.

The HOA B-format is a hierarchical format, meaning that it is necessary to have channels $1 \rightarrow (M-1)$ to make use of channel M, and the presented codec supports this kind of structure. This is achieved by only allowing the encoder to predict channel content from reference channels with a lower channel number. A result of this is that the first channel has to be encoded independently, which corresponds to the middle signal path in figure 1. The removal of intra-channel redundancy is performed using a Weighted Cascaded Least Mean Squares (WCLMS) predictor [11]. This is based on cascading three Least Mean Squares (LMS) predictors with order 80, 30, and 5. The residual from each predictor is fed into the next predictor, and a weighting determines the final estimate. The weighting is based on prior performance of the three predictors. To avoid error propagation the predictors are reset for each block, which here is set to 4096 samples. The advantage of using adaptive prediction is that it is possible to transmit packets before the whole block has been processed. Here, a packet is

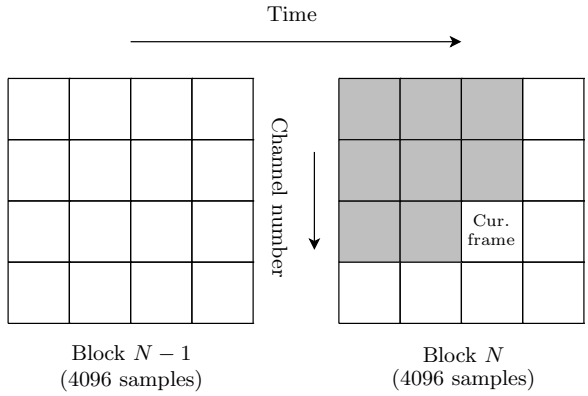


Figure 2. The possible reference frames for a given coding frame are shown with a gray background.

transmitted for every 256 samples (1 frame). However, it should be mentioned that if a packet is lost, the remaining packets until the next prediction reset are rendered useless.

For channels 2 and higher, channels with a lower number may be used to provide an even more accurate prediction. Here, the encoder can select between two alternatives. This inter-channel prediction can be done after the intra-channel prediction (left signal path) or it is possible to use the original signals for prediction directly (right signal path). Using the inter-channel prediction on the original signals is typically only successful for highly correlated signals, but this can easily occur, especially for synthetic HOA signals. Signals generated with only one source and no added reverberation will make channels $2 \rightarrow (M+1)^2$ equal to channel 1, but scaled with a factor depending on the location of the source.

This inter-channel prediction is based on a simple three tap filter where the coding channel (x^c) is estimated by 3 samples from the reference channel (x^r), and the prediction error is given as

$$e^c(n) = x^c(n) - \sum_{i=-1}^{1} \gamma_i \cdot x^r(n-i). \tag{5}$$

The filter coefficients γ_i are found using the procedure in [15].

For a given coding frame, there are a number of potential reference frames as illustrated in figure 2. To preserve the hierarchical structure of the B-format, only frames from the same channel or channels with a lower number than the current channel can be used as a reference. Also, the reference frame must be from the same or an earlier time slot, and it has to be from the same block as the coding frame to avoid error propagation across blocks.

The actual reference channel is found by checking the correlation between the current coding frame and its potential reference frames. A simple approach for reducing the computational load is to limit the potential number of reference frames. In this work it is evaluated how restricting the reference to be from the same time slot as the coding frame, which will correspond to the same column as the coding frame in figure 2, affects the total bit rate.

For entropy coding the encoder can select between Rice [16] and adaptive Huffman [17] coding. For a typical prediction residual from an audio signal, Rice coding has been shown to be near optimal. However, there are certain situations where Rice encoding is far from ideal. For synthetic signals the inter-channel correlation can be very high, such that the prediction residual becomes very small. The minimum bit per sample for Rice coding is 2, so in these situations the encoder can switch to adaptive Huffman in order to decrease the rate further.

4. Transmission over IP Networks

The transmission of multichannel audio is difficult, simply due to the high data rate and the requirements for both low delay and a low packet loss ratio. Regular data is usually transmitted with the Transmission Control Protocol (TCP), which is a reliable protocol and guarantees that all packets are received since lost packets are retransmitted automatically. This protocol also adjusts the sending rate in order to only use a fair share of the total bandwidth. Media streams, however, are often transmitted using the User Datagram Protocol (UDP) since the requirements for low delay makes retransmissions impossible. This protocol provides neither retransmissions nor congestion control. When using UDP it is important to adjust the sending rate, such that the stream does not use too much bandwidth. In [18] a flow is said to be TCP-friendly "if its sending rate is generally within a factor of two of the sending rate of a TCP flow under the same conditions". Under these conditions HOA has a clear advantage when compared to other multichannel formats since the format is naturally scalable. If the sender is required to reduce the data rate, it is possible to simply drop some of the higher order channels.

One problem with the hierarchical structure of HOA is that it is impossible to use the higher order channels if the lower order channels have been lost. Thus, it may be beneficial to transmit the lower order channels at a high priority. With the current internet this is not possible. However, with Differentiated Services (DiffServ), a newer network architecture, priority differentiation is implemented. When using this approach, the probability that the lower order channels are received can be increased or even guaranteed.

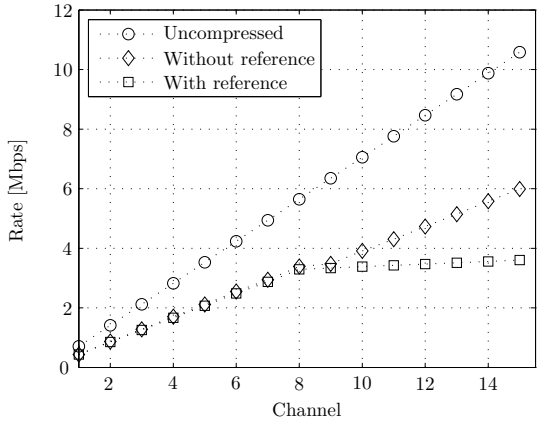


Figure 3. B-format: Rates for synthetic clip with four sources.

5. Synthetic Signals

In addition to signals recorded with the "Eigenmike", our approach has also been tested with synthetic signals, which can be said to correspond to a recording with flawless microphones in an ideal anechoic chamber, and without any noise. The B-format signals have been generated with the approach found in [3], but restricted to only two dimensions. The presented clip consists of four sources equally spaced around a circle, and encoded to order 7 which for a two dimensional representation leads to 15 channels. More results for synthetic signals can be found in [14].

Figure 3 shows that there is almost no increased rate for including channels 9-15 when the clip is compressed using reference channels. The reason for this can be seen in figure 4. The bubbleplot illustrates which channels that are used as a reference for any given coding channel. The area of the circle illustrates how often a channel has been used as reference for a given coding channel, and reference channel 0 means that the channel is compressed without the use of a reference. It can be seen that for channels 9-15 all frames are encoded using inter-channel prediction, and that only one channel has been used as reference for each coding channel. This indicates that e.g. channels 6 and 10 are highly correlated, and for this clip the difference between those two channels is in fact only a scaling factor.

For this clip we end up with a total rate of 3.60 Mbps for all 15 channels which is only 34% of the original uncompressed bit rate. To achieve this rate a full search for reference frames is used, i.e., all the gray frames in figure 2 are searched to find a reference frame. If we reduce the search to only include frames from the same time slot the computational complexity is greatly reduced. Interestingly, for this clip this does not affect the rate much, the new compressed

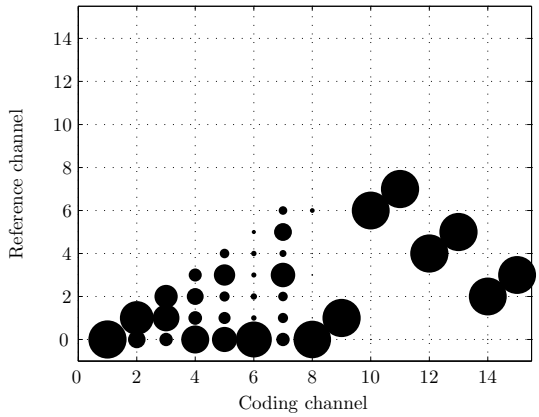


Figure 4. Bubbleplot showing which channels are used as reference for a synthetic clip with four equally spaced sources.

rate is 3.61 Mbps. This indicates that the best reference is almost always found within the same time slot as the current coding frame. This is not surprising for a synthetic clip where there are no echoes arriving later in a different time frame.

6. "Eigenmike" Recordings

The output from the "Eigenmike" is 32 channels with a sampling frequency of 44.1 kHz and 16 bit per sample. The layout of the elements on the sphere is shown if figure 5. The microphone capsules are closely spaced, meaning that in principle there should only be small differences between the signals, especially for anechoic recordings. However, in practice, due to addition of noise and microphone imperfections the signals differ somewhat more than the theoretical values making it difficult to achieve the same compression ratios as for synthetic signals.

6.1. Microphone Signals

The correlation between the microphone signals should theoretically be highest for a recording in an anechoic chamber due to the lack of reflections. If we consider classical piano music recorded in this environment, we see that there is a significant coding gain from exploiting the inter-channel correlation. The uncompressed bit rate for 32 channels is 22.58 Mbps, while the lowest bit rate achieved here is 10.7 Mbps which is only 47% of the original uncompressed bit rate (figure 6). Although this is a significant rate reduction, it is not as high as for the synthetic signals. If all channels are encoded independently the total rate becomes 12.4 Mbps, meaning that the gain from using reference frames is

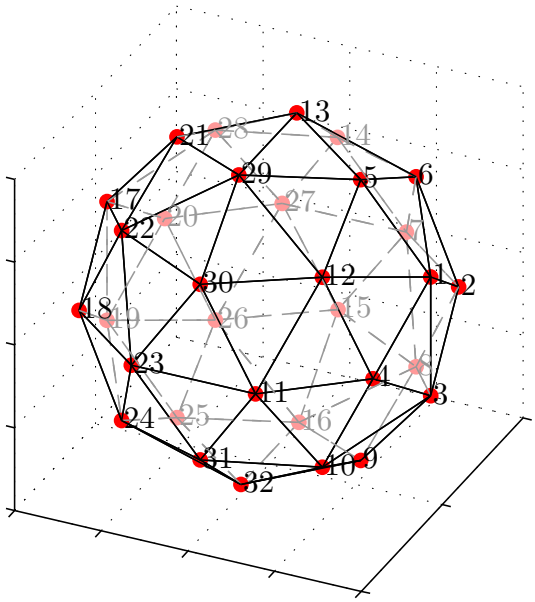


Figure 5. Location of the microphone elements on the "Eigenmike". Shaded elements are located on the back of the microphone. The radius of the microphone is 42 mm.

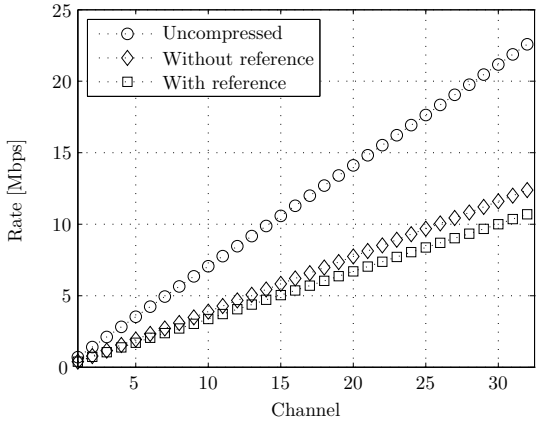


Figure 6. Microphone signals: Rates for music recorded in an anechoic chamber.

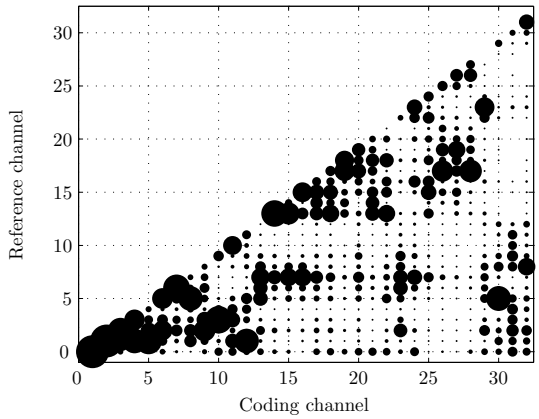


Figure 7. Bubbleplot showing relationship between coding- and reference channel for the microphone signals from music recorded in an anechoic chamber.

close to 14% which indicates that the microphone signals are highly correlated. Figure 7 shows that a lot of frames are encoded using inter-channel prediction. By comparing figure 7 and 5 it can be seen that the reference frame selected is often from one of the neighboring microphone elements, but there are also situations where the physical distance between the coding and reference frame is longer.

As the number of channels increases, the complexity for finding the best reference frames will also increase. If the search is limited to only frames from the same time slot the total rate is actually not affected at all, the best reference frame is always found in the same time slot as the coding frame.

The results for recorded speech is quite similar to the results for music. The minimum rate for all channels is only 8.79 Mbps as seen in figure 8, which is 39% of the uncompressed bit rate. Also for this clip the compression gain from using inter-channel prediction is significant, the rate is reduced with 9% when compared with encoding all channels independently.

If the search for a reference frame is limited to the same time slot as the current coding frame, the minimum rate is still 8.79 Mbps which indicates that all the reference frames are also here found within the same time slot as the coding frame.

For a recording from a regular room the wall reflections will affect the recording and these reflections can be seen as a distortion when compared with an anechoic recording. Surprisingly, the results for a regular room are very similar to the recordings from the anechoic chamber. Here, the experiments were done in a room with a mid-frequency reverberation time of approximately 0.2 s.

For a clip containing music the total rate (figure 9) is reduced to 9.59 Mbps, which is 16% less than the rate when each channel is compressed independently.

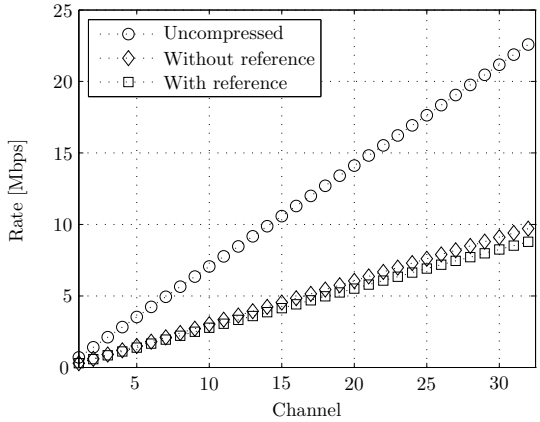


Figure 8. Microphone signals: Rates for speech recorded in an anechoic chamber.

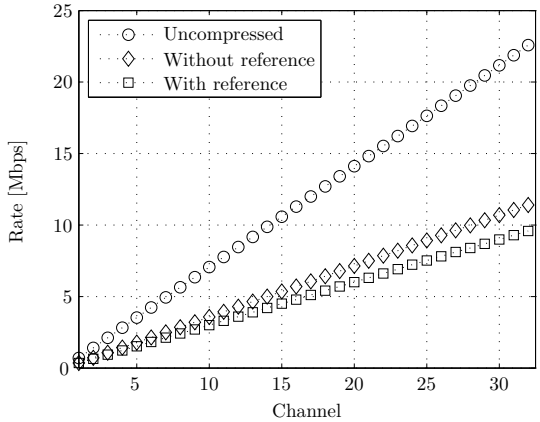


Figure 9. Microphone signals: Rates for music recorded in a regular room.

In this environment, with many delayed reflections it could have been expected that using a full search for reference frames would be beneficial, since the current coding frame can contain a strong reflection. However, reducing the search for reference frames has no effect on the total rate even in this situation.

6.2. Higher Order Ambisonics Signals

The microphone signals from the "Eigenmike" can be transformed into HOA B-format using the procedure from section 2. Given the 32 microphone elements, the maximum order is four, which leads to 25 B-format channels and the uncompressed

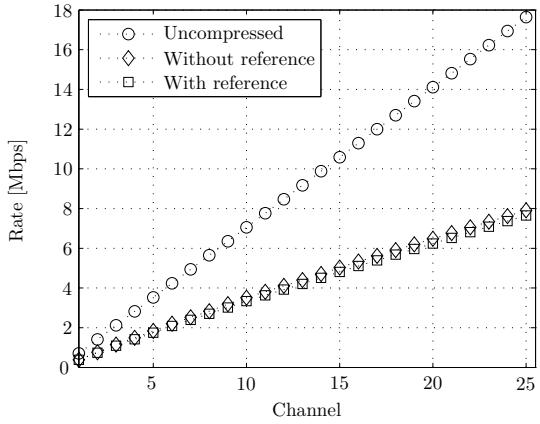


Figure 10. B-format: Rates for music recorded in an anechoic chamber.

bit rate becomes 17.64 Mbps. One important difference between the microphone signals and the B-format signals is that the energy is shifted between the channels. While the microphone signals typically contain about the same amount of energy, this is not true for the B-format channels. The first B-format channel (W), which is the average of all microphone signals, will always have a high energy, but some of the other channels will typically contain much less energy. The content of channels $2 \rightarrow (M+1)^2$ will be dependent on the location of the present sources.

For the channels with low energy, the intra-channel prediction will typically remove so much of the signal that the inter-channel prediction will not lead to an increased compression gain. However, as it will be shown, this does not mean the channels are uncorrelated.

If we first consider classical piano music recorded in an anechoic chamber we find that the total rate is reduced to 7.64 Mbps (figure 10) for all 25 channels, which is 43% of the uncompressed bit rate. However, now the gain from using inter-channel prediction is only a bit higher than 3% which is significantly less than for the microphone signals.

To investigate this inter-channel redundancy further, the energy of the current frame is evaluated prior to encoding. If the energy is below a given threshold, the frame is encoded using only the crosspredictor, i.e., using only the right path in figure 1. By using this approach for the same clip it is found that the rate is only slightly increased ($< 1\%$) even though more than 35% of the frames are predicted using only the crosspredictor. To decrease the complexity further, the search for a reference is restricted to the same time slot as the coding frame, and even in this situation the total rate is not increased.

Figure 11 shows the rates for a clip containing music recorded in a regular room, and for this clip the total rate is as low as 6.41 Mbps. Also for this

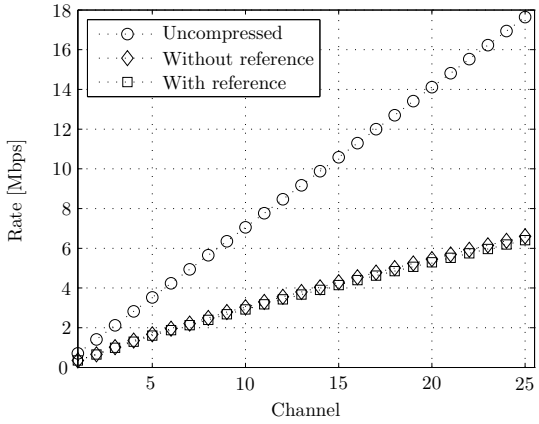


Figure 11. B-format: Rates for music recorded in a regular room.

clip the saving from using inter-channel prediction is 3%. However, if the same criterion for using only the crosspredictor as for the clip in the anechoic chamber is used here, the result is that 23% of the frames are encoded using only the crosspredictor, and also here the rate increases with less than 1%.

The WCLMS predictor has a high computational complexity for both the encoder and the decoder. By evaluating whether or not this predictor is needed prior to encoding the use of this predictor is minimized, and the computational load is reduced. As seen here the number of frames encoded without the WCLMS predictor is high for the HOA signals, but the total rate is only slightly increased, so this is a reasonable tradeoff when it is desired to save processing power.

7. Conclusions

In this work it is demonstrated that the inter-channel correlation is high for both the raw signals from a spherical microphone array and also for the signals converted to the spherical harmonics domain. The use of inter-channel predictors results in a high compression gain for the microphone signals, but the effect is severely reduced for the HOA signals. However, it is shown that the HOA signals are highly correlated even though the compression gain is lower. Furthermore, it is found that restricting the search for a reference frame to frames from the same time slot as the coding frame reduces the complexity, but it will not increase the bit rate significantly. Also, by minimizing the use of the WCLMS predictor, the computational load can be reduced further.

8. References

[1] 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 (ICASSP02)*, 2002.

[2] mh acoustics, "em32 Eigenmike microphone array," `http://mhacoustics.com/page/page/2949006.htm`, [Online].

[3] 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.

[4] J. Daniel and S. Moreau, "Further Study of Sound Field Coding with Higher Order Ambisonics," in *The 116th AES Conv.*, May 2004, Preprint 6017.

[5] S. Moreau, S. Bertet, and J. Daniel, "3D Sound Field Recording with Higher Order Ambisonics - Objective Measurements and Validation of Spherical Microphone," in *The 120th AES Conv.*, May 2006, Preprint 6857.

[6] E. Hellerud, I. Burnett, A. Solvang, and U.P. Svensson, "Encoding Higher Order Ambisonics with AAC," in *The 124th AES Conv.*, 2008, Preprint 7366.

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

[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] M. Lutzky, M. Gayer, G. Schuller, U. Kraemer, and S. Wabnik, "A Guideline to Audio Codec Delay," in *The 116th AES Conv.*, May 2004, Preprint 6062.

[10] M. Schnell, M. Schmidt, M. Jander, T. Albert, R. Geiger, V. Ruoppila, P. Ekstrand, M.Lutzky, and B. Grill, "MPEG-4 Enhanced Low Delay AAC - A New Standard for High Quality Communication," in *The 125th AES Conv.*, 2008, Preprint 7503.

[11] 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.

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

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

[14] E. Hellerud, A. Solvang, and U.P. Svensson, "Spatial Redundancy in Higher Order Ambisoncs and its Use for Low Delay Lossless Compression," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP09)*, 2009.

[15] 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.

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

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

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

Paper F

Influence of Sender Parameters and Network Architecture on Perceived Audio Quality

Jan Erik Voldhaug, Erik Hellerud, Astrid Undheim, Erling Austreim, U. Peter Svensson, and Peder J. Emstad

In Acta Acustica united with Acustica, Volume 94, pp. 1–11, 2008.

The first author had the original idea for this paper, but quit his PhD studies before the paper was finished. The thesis author finished the statistical analysis and wrote most of the paper. The third and fourth authors performed the network simulations and created the models for the packet loss processes.

Is not included due to copyright

Paper G

How Much is Too Much? - On the Acceptability of Packet Loss Distorted Audio

Jan Erik Voldhaug, Erik Hellerud, and U. Peter Svensson

In proceedings of the International Conference on Signal Processing and Communications (ICSPC07), Dubai, United Arab Emirates, November 2007.

The first author had the original idea for this paper, but quit his PhD studies before the paper was finished. The thesis author finished the statistical analysis and wrote most of the paper.

Is not included due to copyright