An Investigation into Embedded Audio Coding Using An AAC Perceptually Lossless Base Layer

Kevin Adistambha, Christian H Ritz, Jason Lukasiak, Ian S Burnett

Whisper Labs
School of Electrical, Computer, and Telecommunications Engineering
University of Wollongong,
NSW, Australia
ka07@uow.edu.au

Abstract

Embedded lossless audio coding attempts to combine the higher compression ratios of perceptual coding with the perfect reconstruction of the original signal provided by lossless coding. This paper examines the residual signal of a perceptual audio coding base layer and considers its usage as an embedded bitstream for an embedded lossless coder. It is shown that the residual signal of a lossy perceptual audio coder retains correlation that can be exploited in lossless compression. This allows an embedded stream to be provided in a lossless coder with approximately 6% overhead over pure lossless audio coding. A 6% overhead appears to be a minimal cost for the backward compatibility and scalability afforded by embedded streams.

1. Introduction

Over the past decade there has been a surge of consumer interest in audio coding; this has particularly focused on perceptual audio coding such as MPEG-1 Layer III (mp3) (Painter and Spanias 2000) and, more recently, MPEG-2/4 Advanced Audio Coding (AAC) (Herre and Purnhagen 2003). These two coders use psychoacoustic techniques designed to remove perceptually irrelevant sections of the original signal in order to achieve very high compression ratios. While the resulting perceptually compressed signal sounds near perfect to the human ear, there is a significant amount of audio data that is not recoverable following compression.

Lossless compression, on the other hand, attempts to achieve objectively lossless compression by preserving every bit of data of the source signal. Examples of lossless compression coders are LZ77 based gzip (http://www.gzip.org), Monkey's Audio (http://www.monkeysaudio.com) and more recently MPEG-4 Audio Lossless Coding (Liebchen, Reznik, Moriya, and Yang 2004). Lossless coding is important in any environment that cannot tolerate any loss of data such as mastering and archiving. The problem with lossless compression is that the resultant compressed file size is significantly larger than a comparable perceptually compressed signal. For comparison, lossless coding can achieve compression ratios of about 5 bits per sample (Hans and Schafer 2001) if the original signal is 16 bits per sample, while perceptual coding can achieve compression ratios of 1-2 bits per sample (Painter and Spanias 2000) while retaining the sonic clarity of the original 16 bit signal due to psychoacoustic processing. This leads to difficulties in transferring the losslessly encoded signal through a limited bandwidth environment such as the Internet.

Lossless scalable embedded coding is a relatively new coding paradigm (Raad 2002; Yu, Lin, Rahardja, and Ko 2004) that attempts to combine the advantages of perceptual coding and lossless coding. This approach can achieve bit-by-bit perfect reconstruction of the original signal, given sufficient bitrate, while maintaining the ability to scale to lower bitrates without the need to re-encode the signal. Embedded coding achieves this by embedding a lossy signal in the lossless bitstream, so the lossless bitstream can be truncated to reveal the lossy information only. There is significant work in this area using different approaches such as Advanced Audio Zip (AAZ) (Yu et al. 2004) which is the reference implementation of the upcoming MPEG-4 Scalable Lossless Coder standard, (Moriya, Iwakami, Jin, and Mori 2000), (Moriya, Jin, Mori, Ikeda, and Kaneko 2003), (Geiger, Herre, Schuller, and Sporer 2003) and Set Partitioning in Hierarchical Tree (SPIHT) based embedded scalable coder (Raad 2002). Within these examples, there are coders that work within the existing MPEG framework of audio compression such as AAZ and (Moriya et al. 2000), while works like (Raad 2002) are entirely new coders based on the embedded paradigm.

Embedded coding is a more flexible approach to delivering digital content than perceptual coding or lossless coding individually. One aspect of flexibility that is particularly relevant to content providers is: there is no need to encode a source signal multiple times in different bitrates to accommodate different bandwidth requirements. The most attractive way to deliver an embedded audio bitstream is, we believe, to encode the original signal with a perceptual coder first then encode the residual signal using entropy coding approaches to achieve objective lossless quality.

In this paper we will examine the feasibility of working with MPEG-4 AAC as the base layer and using different coding methods to compress the residual (the lossless enhancement layer), which we define as the sample-by-sample difference between the base layer and the original signal. Our goal is to achieve the lowest overall base layer plus enhancement layer bitrate by altering the operating bitrate of the base layer coder. Since not all perceptual

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

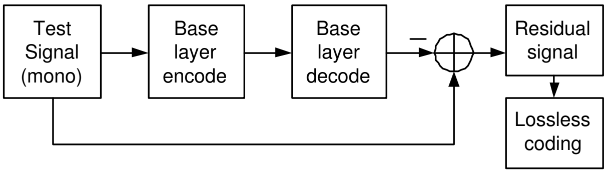

Figure 1: *Experiment block diagram.*

Table 1: *Genre breakdown of test signals.*

Genre	Number of files
Ambient	3
Blues	1
Classical	5
Electronic	35
Jazz	2
Pop	31
Rock	11

coders are created equal, e.g. there are some coders that are tailored to perform best at specific bitrate requirements, we will also present results for mp3 and open source Ogg Vorbis (http://www.vorbis.com) as the base layer.

2. Methodology

The experiment was performed using the structure shown in Figure 1. As described in Figure 1, the base layer was first encoded from the original signal. The resulting perceptually compressed signal was then decoded and subtracted from the original signal to obtain the residual signal. This was performed using appropriate delays to ensure time synchronization of the input and encoded signals.

In order to gain statistically significant results, a large test signal set consisting of 88 music files, varying from 9 seconds to 7 minutes 10 seconds in length with an average length of 2 minutes 52 seconds overall, was employed in determining the residual signal depicted in Figure 1. The music files were obtained from Q-Music (http://www.q-music.co.uk) with 44.1 KHz sampling rate, 16 bits per sample stereo. Before encoding, the files were reduced to mono by taking the left channel and discarding the right channel. Mono test signals are chosen to simplify the experiment, since correlation between channels is not significant and therefore multichannel lossless coding generally compresses each channel independently (Hans and Schafer 2001). The genre breakdown of the test signals is listed in Table 1.

Blues, classical, jazz, rock and pop categories are self-explanatory. Ambient is music with no beat or any percussion. Electronic is music where the instruments used to create the sounds are not traditional instruments, but rather electronically created by using devices such as a synthesizer.

For the base layer four different perceptual coders were tested: Nero AAC (http://www.nero.com), Psy-Tel AAC (http://www.rarewares.org), Lame mp3 encoder (http://lame.sourceforge.net) and Ogg Vorbis (http://www.vorbis.com). The base layers in this experiment were encoded at bitrates of 64, 96, 128, 160,192, 224 and 256 kbps.

The tested coders operate using the same basic principles. The original signal is first divided into frames using windowing functions specific to each coder. Each frame is then transformed from time domain into frequency domain using the Modified Discrete Cosine Transform (MDCT). Psychoacoustic techniques are then used to determine which parts of the signal inside the frame can be coarsely quantized, while keeping the overall perceptual quality of the encoded signal relatively high. This judgement is based on the target bitrate and is achieved by keeping quantization noise below the masking threshold. The quantized parameters from these steps are then entropy coded and multiplexed with side information to allow correct decoding of the resulting bitstream. Since the signals to be encoded in this experiment are mono, coding techniques relevant to the perceptual base layers such as joint stereo or mid/side coding are not used. In the case of AAC coders, the AAC Low Complexity profile is used. Perceptual Noise Substitution (PNS) and other low bitrate coding tools associated with MPEG-4 AAC (Herre and Purnhagen 2003) are not used in this paper.

The perceptually encoded signal is decoded and subtracted from the original signal to create the residual signal. The residual signal will then contain the quantization noise that the base layer coder tries to keep below the masking threshold. An analysis of the residual signal thus indirectly measures the psychoacoustic characteristics of the base coder.

To analyse the residual we used three methods. The first method is a calculation of the first order entropy of the residual using:

$$H = -\sum_{i=1}^{N} p_i \, log_2 \, p_i \tag{1}$$

where *H* is the entropy of a signal, *N* is the number of symbols and p_i is the probability of an i^{th} symbol occurring in the signal.Also called Shannon entropy (Sayood 2000), *H* provides the lowest possible compression in bits per sample that can be theoretically achieved. Shannon entropy assumes that the source signal is *independent identically distributed (i.i.d.)*.

The second analysis method used in this experiment calculates the Signal to Noise Ratio (SNR) between the residual and the original signal so as to measure the level of quantization noise present in the residual signal. SNR is used instead of Signal to Mask Ratio (SMR) (Painter and Spanias 2000) because lossless coding is not concerned with perceptual masking. The SNR measurement is therefore used in this paper to show the objective scalability of the base layer coders across tested bitrates.

The third method involves compression of the residual using gzip and Monkey's Audio. Gzip was chosen to represent a universal lossless compression tool based on the well known LZ77 algorithm. Monkey's Audio was chosen to represent the state of the art in audio specific lossless coding (Liebchen et al. 2004) that involves decorrelation steps before entropy coding.

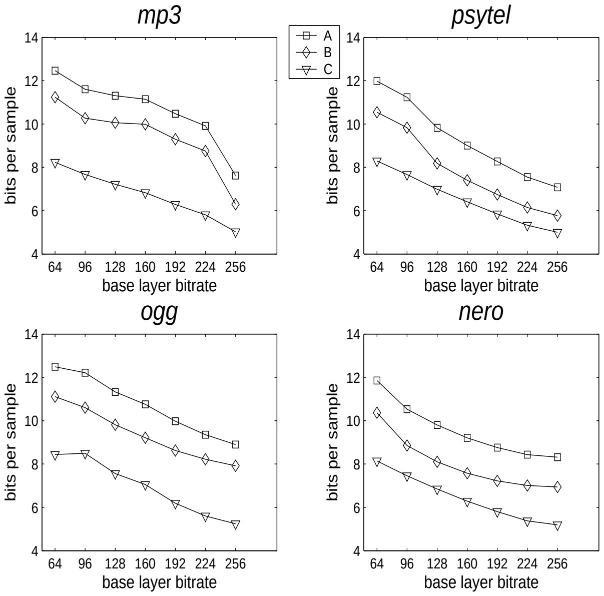

Figure 2: *Gzip compressed, entropy and Monkey's Audio compressed residual bits per sample comparison. Curve A represents the bitrate of the residual compressed with gzip, curve B represents the entropy of the residual signal calculated using (1) and curve C represents the bitrate of the residual compressed with Monkey's Audio.*

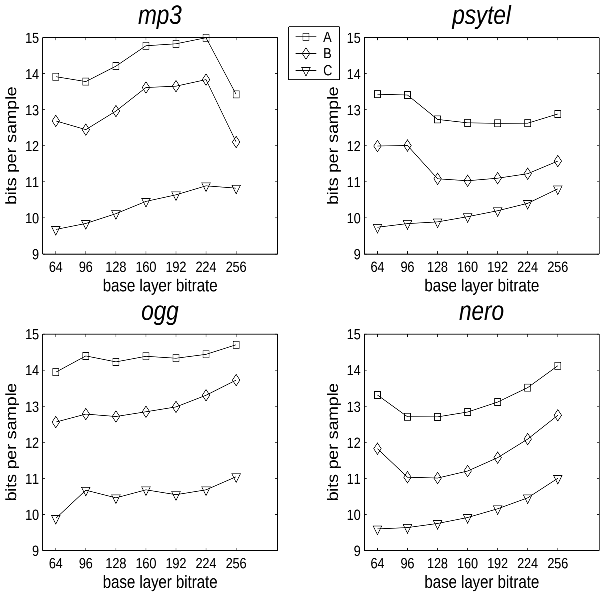

Figure 3: *Total bitrate comparison. Curve A represents the bits per sample of the lossy base layer+residual compressed by gzip, curve B is lossy base layer+residual entropy and curve C is lossy base layer+residual compressed with Monkey's Audio.*

3. Results

3.1. Entropy and compression of the residuals

To determine the characteristics of the residual across various base layer rates, the residual coding methods detailed in Section 2 were applied to the test signals shown in Table 1. The results are shown in Figure 2.

In Figure 2, the results are presented in bits per sample to decouple them from the sampling rate. From Figure 2 it is clear that compressing the residual using Monkey's Audio (curve C) consistently yields lower bits per sample across all base layer bitrates. Although the entropy computation should give the lowest possible bits per sample achievable, the measure used here assumes an independent and identically distributed property for the signal. The lower bit rate results of Monkey's Audio suggest that correlation still exists in the residual signal.

Note that all tested base layer coders perform differently across the same range of bitrates. This indicates that there is no general best rate at which any lossy base layer should operate when only considering compression of the residual signal, i.e. the results are specific to the psychoacoustic model employed by the base layer coder.

3.2. Total bitrate

The total bitrate of an embedded stream resulting from adding the base layer stream to the respective lossless residual streams is presented in Figure 3.

The total bits per sample have an increasing trend as the base layer bitrate increases for all coders tested, even though in Figure 2 the residual bits per sample have a decreasing trend with increasing base layer bitrate. This is due to the fact that the base layer increases by ≈ 0.73 bits per sample for each 32 kbps increase in bitrate, which means that we have to find a balance between increasing base layer bitrate, decreasing bits per sample of the residual and increasing SNR of the residual.

3.3. Signal to noise ratio

Figure 4 shows the SNR of the residuals and the residual's entropy. As expected, and because there is less quantization error present as the base layer bitrate is increased, the SNR of the residual becomes higher with increasing base layer bitrate. Figure 4 shows the scalability of the Nero AAC encoder across tested bitrates in comparison with the other coders, due to the regularity of which it increases its SNR with the corresponding bitrate.

3.4. Comparison of embedded coding and lossless only coding

There are two methods available to achieve embedded lossless coding. The first method, as presented in this paper, involves encoding the base layer using a lossy coder and appending the residual signal. The second method involves encoding of the original signal losslessly and appending a perceptually coded version of the signal to the lossless stream. While the second method does not initially appear valid, it is less complex than the first method because there is no decoding stage required at the encoder.

To compare the resultant file sizes of these two methods, both methods were implemented. Table 2 is the result using the Nero AAC as the base layer encoder. Column A is the average file size increase when a losslessly encoded signal has a perceptually encoded signal appended. Column B is the average file size increase of embedded coding, using a perceptually encoded base layer and losslessly encoded residual signal, over a straight lossless encoding. For

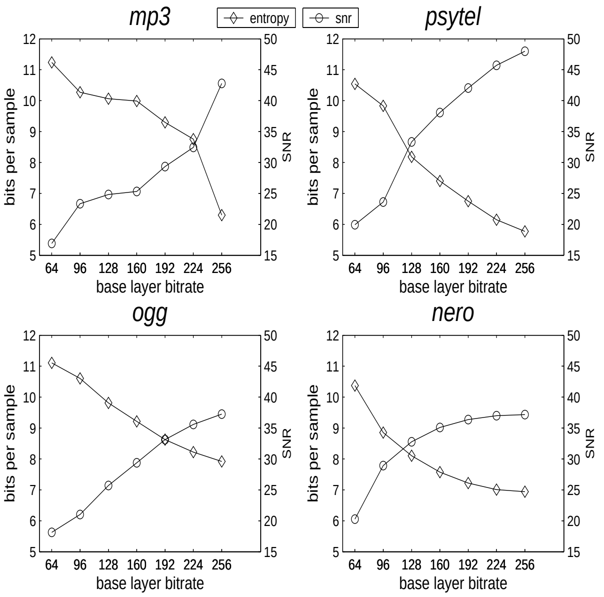

Figure 4: *Residual entropy and residual SNR.*

Table 2: *Percentage of filesize increase of embedded coding compared to pure lossless coding with Nero AAC as the perceptual layer and Monkey's Audio as the lossless layer.*

Nero AAC bitrate	A (lossless+perceptual)	B (lossy+residual)
64 kbps	16.02%	5.88%
96 kbps	24.02%	6.37%
128 kbps	32.03%	7.55%
160 kbps	40.04%	9.44%
192 kbps	48.05%	12.04%
224 kbps	56.05%	15.42%
256 kbps	64.06%	21.33%

the purposes of these figures we performed straight lossless coding using Monkey's Audio.

In Table 2, the advantage of embedded lossless coding from the file size point of view is evident. While the second coding method, labeled A in Table 2, implies less complexity, the increase in total file size does not justify its simplicity. Embedded coding using a base layer and a residual, labeled B in Table 2 however, results in less than a 10% increase in total file size when the base layer bitrates are below 192 kbps. This approach thus compares favourably to pure lossless coding while giving significant flexibility in the resulting bitstream.

4. Discussion & conclusions

In general, results indicate that higher base layer bitrates correspond to higher lossless compression ratios for the residual signal. This is expected as there is less quantization noise present in the higher base layer bitrate residual; this is evident in the SNR calculation. This residual characteristic is consistent with all perceptual coders tested in this paper. From Figures 2 and 3, the consistent lower bits per sample achieved by Monkey's Audio reveal that employing a decorrelation step in lossless residual coding is beneficial.

From Figure 2 and Figure 3, the Nero AAC encoder gives the lowest overall total bitrate amongst the coders that we tested. Based on empirical results, we have found that the best tradeoff is achieved when the base layer is operated at 96 kbps mono. This was demonstrated by an increase of total embedded file size of only 6.37% depicted in Table 2 compared to pure lossless coding, and the perceptual quality of 96 kbps Nero AAC approaches perceptually lossless synthesis.

While Monkey's Audio compression of the residual results in a lower bits per sample value than the computed entropy of the residual, Monkey's Audio was primarily designed to compress audio signals. Developing a lossless coder specifically tailored to compression of the residual signal would likely give better results in our coders.

5. Future work

For future work, we intend performing tests with different sampling rates (48 and 96 KHz) and bits per sample (24 or 32 bits). We also believe that it would be worthwhile considering the effect of MPEG-4 low bitrate coding tools such as Perceptual Noise Substitution and how those tools affect the characteristic of the residual signal. For encoding of the residual, novel methods such as SPIHT are also worth closer consideration.

References

Geiger, R., J. Herre, G. Schuller, and T. Sporer (2003). Fine grain scalable perceptual and lossless audio coding. *Proceedings of the ICASSP 2003*.

Hans, M. and R. Schafer (2001, July). Lossless compression of digital audio. *IEEE Signal Processing Magazine*, 21–32.

Herre, J. and H. Purnhagen (2003). General audio coding. In F. Pereira and T. Ebrahimi (Eds.), *The MPEG-4 Book*, pp. 487–544. IMSC Press.

Liebchen, T., Y. Reznik, T. Moriya, and D. Yang (2004). Mpeg-4 audio lossless coding. *Proceedings of the 116th AES*.

Moriya, T., N. Iwakami, A. Jin, and T. Mori (2000). A design of lossy and lossless scalable audio coding. *Proceedings of the ICASSP 2000*.

Moriya, T., A. Jin, T. Mori, K. Ikeda, and T. Kaneko (2003). Hierarchical lossless audio coding in terms of sampling rate and amplitude resolution. *Proceedings of the ICASSP 2003*.

Painter, T. and A. Spanias (2000, April). Perceptual coding of digital audio. *Proceedings of the IEEE 88*.

Raad, M. (2002). *Scalable and Perceptual Audio Compression*. Ph. D. thesis, University of Wollongong, Australia.

Sayood, K. (2000). *Introduction to Data Compression*. Morgan Kaufmann.

Yu, R., X. Lin, S. Rahardja, and C. Ko (2004). A scalable lossy to lossless audio coder for mpeg-4 lossless audio coding. *Proceedings of the ICASSP 2004*.