HIGH QUALITY LOW COMPLEXITY SCALABLE WAVELET AUDIO CODING

W. Kurt Dobson, J. Jack Yang, Kevin J. Smart, and F. Kathy Guo
U.S. Robotics Mobile Communications Corp.
605 North 5600 West
Salt Lake City, UT 84116
jyang@mhz.com

Abstract

This paper presents an audio coder for real-time multimedia applications. To achieve high quality at low bit rate, the audio coder uses a wavelet packet decomposition to transform the audio data into the wavelet domain, and a psychoacoustic model is used to minimize quantization noise. The wavelet packet decomposition tree structures were chosen in a way to closely mimic the critical bands in a psychoacoustic model. Instead of determining the masking thresholds in the Fourier domain, the wavelet coefficients are used to drive the psychoacoustic model directly. Most of the standard industrial sampling frequencies are supported by this coder. An efficient bit rate control scheme was designed such that the audio coder operates at virtually any desired bit rate level. The audio coder achieves near perceptually lossless quality at or below 80 kb/s for most audio sources. Real-time encoding/decoding is possible by using only a fraction of a Pentium or faster CPU.

1 Introduction

Reduction in bit rate requirement for high quality audio has been an attractive proposition in applications such as multimedia, efficient disk storage, and digital broadcasting. A number of audio compression algorithms including [1, 2, 3, 4, 5, 6] were developed. Among them, the most notable is arguably the the ISO/MPEG standard [1, 2], which is discrete cosine transform (DCT) based and provides high quality audio at about 64kb/s. [3] and [4] drive down the bit rate even further by using signal adaptive filter banks. All these methods use psychoacoustic models to mask distortion incurred by quantization.

While bit rate reduction has been the focus of these algorithms, the following issues in real-time multimedia applications did not get enough attention:

- The bit rate should scale to any desired level to accomodate many practical communication channels, such as ISDN links and modems.
- The complexity of an audio coder should be low such that real-time encoding/decoding on most personal computers, without additional hardware, is possible.
- Most industrial standard sampling rates (including 44.1 kHz, 32 kHz, 22 kHz, 16 kHz, 11 kHz, and 8 kHz) should be supported.

In this paper, we present an audio coder that satisfies the above requirements. The wavelet transform [7, 8] has been identified as an effective tool for data compression. To provide high quality audio at low bit rate in our coder, statistical redundancy in audio data is removed by decomposing the data using a wavelet packet (WP) transform. Auditory masking effects are used to minimize quantization noise. The tree structures for WP decomposition were chosen in a way to closely mimic the critical bands in a psychoacoustic model. The tree structures for audio sources of different sampling frequencies (including 44.1 kHz, 32 kHz, 22 kHz, 16 kHz, 11 kHz and 8kHz) are illustrated. An efficient bit rate control scheme was designed such that the audio coder operates at virtually any desired bit rate level. The algorithm is optimized, and real-time encoding/decoding is achieved using a fraction of a Pentium's CPU power.

2 Coder Structure

Assume that we have a stream of audio data $x(n)$ that is serial-to-parallel buffered. The buffered audio data is passed through a WP transform. The transformed data is then quantized and entropy encoded. The output bit rate from the entropy coder is monitored by a bit rate control scheme, which in turn controls the quantizer and entropy coder to adjust the output bit rate to the desired level. The whole encoder block dia-

gram is shown in Figure 1. The decoder is the reverse flow of the encoder and is not included.

The encoder in Figure 1 can be divided into four functional blocks: WP decomposition of audio data into subbands using a psychoacoustic model, uniform quantization of wavelet coefficients, bit rate control, and entropy coding (run-length coding and Huffman coding). Quantization and entropy coding of coefficients are well known and will not be described in this paper. The WP decomposition using a psychoacoustic model and the bit rate control scheme will be elaborated in more details in the following.

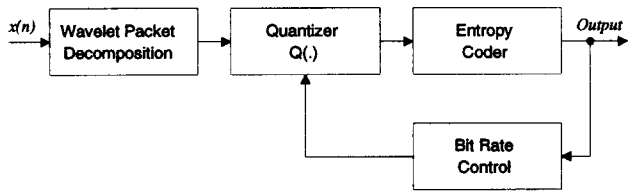

Figure 1: Encoder Block Diagram

2.1 Wavelet Packet Decomposition Using Psychoacoustic Modeling

Psychoacoustic auditory masking is a phenomenon whereby a weak (noise) signal is made inaudible by a simultaneously occurring stronger (audio) signal. Most progress in audio compression in recent years can be attributed to successful application of psychoacoustics to signal compression. In a psychoacoustic model, the whole signal spectrum bandwidth is divided into a number of critical bands. The maximum signal energy and the masking threshold in each band is calculated. The ratio of the two is taken to determine the number of bits required for perceptually lossless quantization. This process is described in [9, 10].

In order to achieve the critical band resolution, most audio coders use a Fast Fourier Transform (FFT) on the input buffered data. Since an N point FFT requires $N \log_2 N$ multiplications, the whole process becomes computationally intensive as N becomes large.

Instead of transforming data using FFT and calculating signal energy and hearing thresholds in the Fourier domain, some wavelet packet transform based audio coders [11] directly calculate the signal energy and hearing thresholds in the wavelet domain. Since an N point FFT is dropped, the complexity is greatly reduced. The WP decomposition tree structure is designed to closely mimic the critical bands in a psychoacoustic model. The wavelet filters are usually chosen with many taps and narrow transition bands so that aliasing between bands is minimized.

In our coder, a WP decomposition tree structure is developed for 32kHz, 16kHz, and 8kHz sampling frequencies and is shown in Figure 2. The lower and higher cut-off frequencies of each band are also illustrated. This tree structure approximately achieves the frequency resolution of the critical bands described in [9, 10].

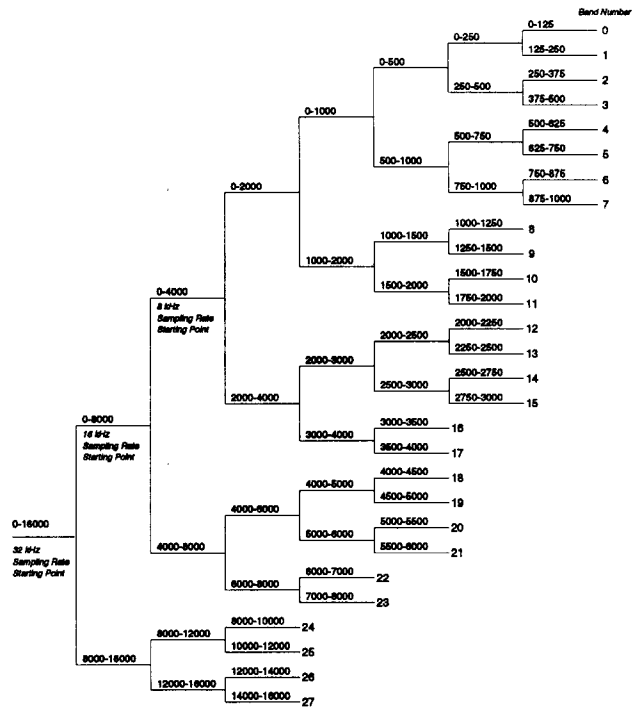

Figure 2: WP Tree Structure for 32kHz, 16kHz, 8kHz Sampling Rates

The 29 band tree structure proposed in [4] is used in our coder for 44.1kHz sampled audio. The tree structure for audio sources of other bandwidth are obtained by modifying the 29 band tree structure. For example, for 22kHz sampled audio, the higher four bands representing frequency range between 11kHz and 22kHz are truncated while the lower 25 bands are preserved. Similarly, for 11kHz sampled audio, frequency bands above 5.5kHz are dropped. This results in a tree structure with 20 subbands for 11kHz sampled audio.

Although aliasing effects between neighboring bands can be reduced by using filters with narrow transition bands, such effects will inevitably exist since any practical filters have to be of finite length. Furthermore, either symmetric extension of data (for symmetric filters) or circular convolution is used for filtering, thus creating additional frequency distortions on certain data points, especially on those at subband boundaries. All these factors contribute to the incorrect calculation of signal power and masking threshold

in a psychoacoustic model. However, our experiments showed that even with the presence of all those factors, better performance is achieved by using a psychoacoustic model.

2.2 Bit Rate Control

In cases where the output bit rate is higher than the desired level, a control scheme is used to reduce the bit rate. Although this results in audible distortion, it is necessary due to the limited bandwidth offered by many practical communication channels.

Assume that G_i is the maximum absolute value and M_i is the masking threshold in the i^{th} band, respectively. M_i is calculated using the model provided by [4]. The minimum number of symbols S_i required for perceptually lossless quantization of wavelet coefficients is given by

$$S_i = \left\lceil \frac{G_i}{M_i} \right\rceil , \qquad (1)$$

where $\lceil . \rceil$ returns the smallest integer greater than or equal to the operand. The actual assigned quantization bits to the i^{th} band is scaled by a factor of μ ($0 \leq \mu \leq 1$):

$$Q_i = \lceil \log_2(\mu S_i + U(\mu S_i - 1)) \rceil , \qquad (2)$$

where $U(.)$ is the unit step function. One extra symbol is added in (2) for run-length encoding symbol if $\mu S_i > 1$. The i^{th} band does not need to be quantized and transmitted if $\mu S_i \leq 1$. when $\mu = 1$, the quantization will be perceptually lossless. If $\mu < 1$, the bit rate is reduced causing audible quantization noise to be present.

The value of μ is adjusted according to the difference of the actual output bit rate and the desired bit rate. Assuming μ_i is the current value of μ and μ_{i+1} is the next value of μ, μ is adjusted according to

$$\mu_{i+1} = \mu_i + \frac{b_{\text{desired}} - b_{\text{used}}}{b_{\text{desired}}} \alpha , \qquad (3)$$

where $\mu_0 = 1$ (or some other predetermined value), b_{desired} is the desired number of bits for a given frame, b_{used} is the actual number of bits that were used, α is a factor that controls the convergence of μ. A big value of α makes μ converges faster, however, oscillation may occur. A small value of α insures smooth convergence of μ, but it may take more iterations.

While it is impossible to retain perceptually lossless quality at low bit rates, the importance of each subband may be different. For example, lower frequency bands are usually more important to the perceptual quality of the reconstructed audio data than the higher frequency bands. For this reason, a weighting scheme is designed. Specifically, if B is the total number of subbands in the WP decomposition and Q_i is the number of quantization bits in the i^{th} band, as determined from a psychoacoustic model,

$$Q_i = \begin{cases} Q_i & \text{if } f(\mu)B > i \\ 0 & \text{otherwise} \end{cases} ,$$

where $f(.)$ is a function of μ. In other words, if there are not enough bits for all the coefficients, the coefficients at higher frequency bands are simply discarded.

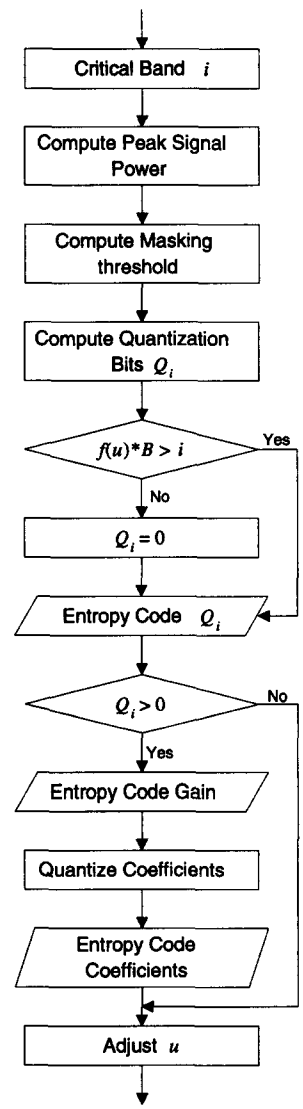

Figure 3: Bit Rate Control Scheme

The bit rate control scheme described in this section is also illustrated in Figure 3.

3 Results

The proposed audio coder was developed and optimized for the Pentium. It was tested on a variety of 16 bit audio sources sampled at different frequencies. The bit rates (kb/s) for near perceptually lossless transmission of the quantized wavelet coefficients and the corresponding CPU usage for real-time encoding/decoding on a Pentium-75 PC for three audio sources ("The Jack" by AC/DC, "Eruption" by Van Halen, Beethoven's Moonlight Sonata performed by Van Cliburn) are given in the following table:

Sampling Freq. (kHz)	Bit Rate AC / DC	Bit Rate V.H.	Bit Rate Moonlight	CPU (%) Enc./Dec.
44.1	55	65	80	59/54
32	45	60	70	43/40
22	42	58	70	31/29
16	35	50	60	23/22
11	35	50	60	16/15
8	25	30	35	12/11

These results show that this coder achieves near perceptually lossless quality for the tested audio sources at or below 80 kb/s, while real-time encoding/decoding is achieved using only a fraction of a Pentium-75's CPU power. The bit rate can be scaled down to any desired level to accomodate the bandwidth requirement of a practical communications system. Although the audio is lossy at lower bit rate, the quality is generally considered as good at about half the bit rate at which perceptually lossless compression is achieved. For example, 16 bit 44.1 kHz sampled audio generally sounds good at about 35 kb/s. Compared with other audio coders we have ever seen, our coder performs at least as good, if not better.

4 Conclusions

A low complexity high quality WP based audio coder is developed in this paper. Compared with many other coders that only address 44.1 kHz wideband CD audio, most of the industrial sampling frequencies are supported by our coder. Perceptually lossless audio quality is achieved at about 80 kb/s for 16 bit 44.1 kHz CD audio, and at lower bit rates for other lower sampling frequencies. An efficient bit rate control scheme is designed such that the bit rate could be scaled down to any desired level. Good audio quality is maintained at about half the bit rate where lossless quality is achieved. The proposed audio coder is of low complexity and real-time encoding/decoding is achieved with a fraction of a Pentium's CPU power. It satisfies real-time multimedia audio application requirements.

References

[1] K. B. Brandenburg and G. Stoll, "ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio", *J. Audio Eng. Soc.*, **42**(10):780–792, October 1994.

[2] D. Pan, "A tutorial on mpeg/audio compression", *IEEE Multimedia*, pages 60–74, 1995.

[3] J. Princen and J. Johnston, "Audio coding using signal adaptive filterbanks", *Proc. ICASSP*, pages 3071–3074, 1995.

[4] D. Sinha and A. Tewfik, "Low bit rate transparent audio compression using adapted wavelets", *IEEE Trans. on Signal Processing*, **41**(12):3463–3479, December 1993.

[5] M. Purat and P. Noll, "Audio coding with dynamic wavelet packet decomposition based on frequency-varying modulated lapped transforms", *Proc. ICASSP*, pages 1021–1024, 1996.

[6] S. Boland and M. Deriche, "Audio coding using the wavelet packet transform and a combined scalar-vector quantization", *Proc. ICASSP*, pages 1041–1044, 1996.

[7] G. Strang and T. Nguyen, *Wavelet and Filter Banks*, Wellesley-Cambridge Press, 1996.

[8] C. K. Chui, *An Introduction to Wavelets*, Academic Press, Inc., 1992.

[9] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception", *Proc. of the IEEE*, **81**(10):1385–1421, October 1993.

[10] J. Johnston, "Transform coding of audio signals using perceptual noise criteria", *IEEE Journal on Selected Areas in Communications*, **6**(2):314–323, February 1988.

[11] M. Black and M. Zeytinoglu, "Computationally efficient wavelet packet coding of wide-band stereo signals", *Proc. ICASSP*, pages 3075–3078, 1995.