COMPARING SCALAR AND Z_n LATTICE BASED ENCODING OF WAVELET COEFFICIENTS IN SINUSOIDAL PLUS WAVELET PACKET CODING

Márk Fék[†‡], *Annamária R. Várkonyi-Kóczy*[†], *and Jean-Marc Boucher*[‡]

e-mail: fek@alpha.ttt.bme.hu, koczy@mit.bme.hu, JM.Boucher@enst-bretagne.fr

[†]Budapest University of Technology and Economics, Dpt. of Measurement and Information Systems
Budapest, Magyar tudósok körútja 2. H-1117, Hungary

[‡]ENST de Bretagne, Dpt. Signal et Communications
Technopôle de Brest Iroise, BP 832, 29285 Brest CEDEX, France

ABSTRACT

We have recently proposed a combined sinusoidal and *Wavelet Packet Transform* (WPT) codec for joint speech and audio coding. In this paper, we compare the performance of scalar and Z_n lattice based encoding of the WPT coefficients. The quantization of the WPT coefficients are determined by a psychoacoustic masking model, and it is identical in both cases. The audio quality remains the same for the scalar and the Z_n lattice based encoding, as only the encoding of the quantized WPT coefficients is changed. The mean bit rate of the coder (depending on the encoded signal) was reduced from 62-32 kbps to 54-30kbps by applying the Z_n lattice based coding. Demonstration sound files are available at *www-sc.enst-bretagne.fr/~fek/demo/*.

1 INTRODUCTION

Some applications require the encoding of both speech and generic audio inputs. One example is Internet radio broadcast, where the successions of commenator speech and music-recordings is transmitted. Another application is the digital archiving of already existing mixed speech and audio recordings, such as musical tales for children.

Speech coding algorithms using a speech specific source model fail to encode music with good quality. Uniform transform based audio coding algorithms use long transform blocks to encode stationary parts of the signal, while the transient parts are encoded by using short blocks. Applying long blocks to encode the rapidly varying speech signal leads to artifacts known as pre-echos. The overuse of short blocks increases the required bit rate considerably.

One solution is to use separate speech and audio codecs for the different types of input. In [1], a speech/music discriminator is used to select the specific encoding for a given input segment. This method does not provide a perfect solution as erroneous decisions lead to coding artifacts on misclassified segments.

In [2], we have proposed a combined sinusoidal and *Wavelet Paket Transform* (WPT) algorithm to encode speech and audio signals. The input is band-limited to 50-7000 Hz and sampled at 16 kHz using 16 bits per sample. The sinusoidal modeling extracts the stable sinusoidal components of the signal. The residual is obtained by extracting the re-synthesized sinusoids from the input, and is processed by a WPT simulating the critical bands of the *Human Auditory System*.

We have achieved mean bit rates between 62-32 kbps by applying uniform scalar quantization and Huffman coding on the WPT coefficients. In order to reduce the bit rate further, we have replaced the scalar quantization and the encoding by a *Lattice Vector Quantization* (LVQ) scheme. The LVQ uses a geometrically structured codebook which eliminates the need of quantizer training and codebook storage, and more importantly it provides a fast codevector search by algebric means. We have used the (scaled) Z_n lattice to quantize and encode the WPT coefficients. The Z_n lattice contains every integer coordinate points of the n-dimensional space. Altough it is not the optimal lattice in sense of mean square error, the quantization and encoding procedure is less complex than for other lattices. It also facilitates the comparison with scalar quantization. The quantized WPT coefficients have exactly the same values for both the scalar and the Z_n lattice quantizers, therefore the audio quality of the compressed signal is the same in both cases.

The paper is organized as follows: Section 2 presents the overall codec architecture. Section 3 describes the masking model based quantization of the WPT coefficients. Section 4 describes the scalar encoding of the quantized WPT coefficients, while section 5 explains the Z_n lattice based encoding of the coefficients. Section 6 presents the bit rate results for the different encodings. Section 7 summarizes the results and presents further perspectives.

2 CODEC ARCHITECTURE

Figure 1 shows the overall structure of the algorithm. The input is band-limited to 50-7000 Hz and sampled at 16 kHz using 16 bits per sample. The encoding and decoding works on a frame-by-frame basis.

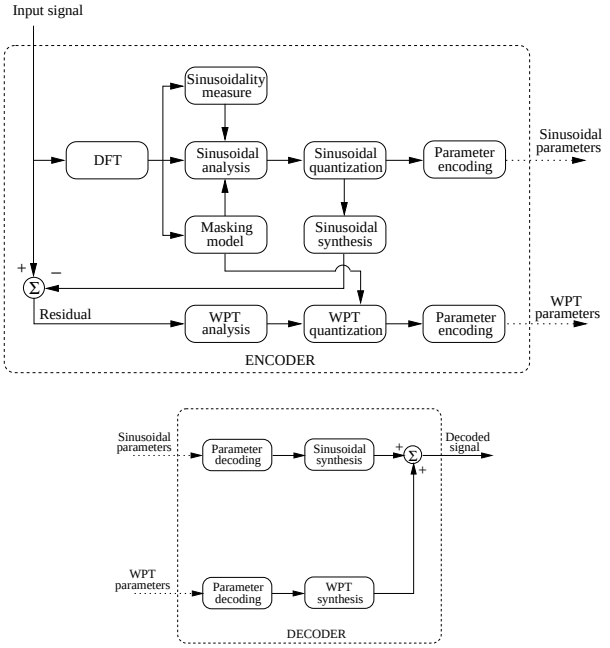

Figure 1: S+WPT encoder and decoder structure.

2.1 Sinusoidal analysis

The encoder carries out a sinusoidal analysis to identify the stable sinusoidal components of the input. The frame size of the sinusoidal analysis is 512 samples with an overlap of 256 samples between two consecutive frames. A masking model is also calculated, based on the MPEG1 psychoacoustic model 2 implementation [3]. The sinusoidal components below the masking threshold are not extracted.

The sinusoidal model works as follows. First, a Kaiser windowed DFT (zero padded to 1024 points) of the current frame is calculated. Next, a *Sinusoidal Similarity Measure* (SSM) [4] is computed as the correlation between a spectral pattern corresponding to the main lobe of the Kaiser window, and the magnitude spectra. A peak in the SSM is considered to represent a valid sinusoid, only if it exceeds a certain threshold. The amplitude, frequency, and phase parameters corresponding to peaks of valid sinusoids are extracted from the magnitude spectra.

The sinusoids are re-synthesized using the parameters extracted from two consecutive analysis frames. The re-synthesis follows the trajectory matching and synthesis procedures described in [5]. Sinusoids not associated with other sinusoids found in the preceding or following frame, are considered to originate from noise components, hence they are eliminated. Peaks having an amplitude below the masking threshold are also eliminated.

2.1.1 Sinusoidal parameter quantization and coding

The sinusoidal amplitudes are quantized on a logarithmic-scale using 6 bits, the frequencies on a *Bark*-scale using 10 bits, and the phases on a linear-scale using 5 bits. The encoding of the quantized sinusoidal parameters is described in [2].

2.2 Residual processing

The residual is formed by subtracting the re-synthesized sinusoids from the original signal. We apply a WPT to process the residual. The WPT realizes a non-uniform filter bank simulating the critical (Bark-) band model of the *Human Auditory System*. Only the first 21 Bark-bands lying in the input frequency range are considered. To avoid coding artifacts, such as pre-echos, on speech input, we use the decomposition described in [6], as it was designed specifically for speech inputs. The WPT is implemented by cascading *Quadrature Mirror Filters* in a tree structure. The *Daubechies* filter of length 10 is used as the prototype filter.

To eliminate the perceptual redundancy, the masking thresholds are used again to quantize the WPT coefficients. The frame size of the WPT is 256 samples. There is no overlap between consecutive frames, but the effective filter lengths stretch beyond the block size. For both the sinusoidal and the WPT analyses, the same psychoacoustic model is used with a frame size of 512 samples and overlap of 256 samples providing a masking threshold for every 256 samples.

The quantized masking levels are encoded and sent to the decoder. The quantized WPT coefficients are either encoded by simple scalar quantization or as vectors on the Z_n lattice. In [2], we have applied *Perceptual Noise Substitution* (PNS) to encode noisy sub-bands. As the noise detection we applied was not reliable enough, this part has been turned off in the following experiments.

The decoder decodes the sinusoidal and residual bitstreams, then it re-synthesizes the two signal components. The re-synthesized sinusoidal and residual components are added together to form the decoded signal.

3 RESIDUAL QUANTIZATION

We calculate the masking thresholds using the MPEG1 psychoacoustic model 2 [3]. The masking thresholds are calculated with a resolution of 512 frequency points, as required by the sinusoidal analysis. They must therefore be converted to give values for each critical band. The masking threshold T_i in a critical band is determined as the minimum masking threshold value in that critical band. It indicates the maximum allowed noise energy that can be introduced in the sub-band without making an audible distortion.

The WPT quantization follows the method described in [6]. The quantization noise of a uniform quantizer can be modeled as white noise. If the input is uniformly distributed over a quantization region, then the quantization noise energy σ_q^2 can be expressed as:

$$\sigma_q^2 = \frac{\delta^2}{12}, \tag{1}$$

where δ denotes the quantization step of the uniform quantizer. By equating the quantization noise σ_q^2 to the masking threshold T_i, we can derive the quantization step δ_i in the ith sub-band:

$$\delta_i = \sqrt{12T_i}. \quad (2)$$

3.1 Quantization and encoding of the quantization steps

The 21 quantization steps δ_i have to be transmitted to the decoder. The quantization steps are quantized on a logarithmic scale using 3 bits. The number of levels l_{ij} required to quantize the jth WPT coefficient c_{ij} in the ith sub-band is determined by dividing the WPT coefficient by the respective quantized quantization step $\hat{\delta}_i$:

$$l_{ij} = \frac{c_{ij}}{\hat{\delta}_i}. \quad (3)$$

The decoder reconstructs the 256 WPT coefficients by multiplying the quantization levels l_{ij} and the respective quantization steps $\hat{\delta}_i$.

The quantization steps are differentially encoded. The first value is encoded on 3 bits. The symbols (i.e. the differences) are encoded using prefix codes. We use shorter codewords if there are fewer than three symbols. Two bits are used to indicate the number of symbols (1, 2, 3, or more).

4 SCALAR ENCODING OF THE WPT COEFFICIENTS

The scalar encoding encodes the quantized WPT coefficients one by one. The quantization levels representing the WPT coefficients are encoded using Huffman codes. We use shorter codewords if there are less than 8 symbols. The number of symbols (1, 2, ..., 7, or more) are encoded on 3 bits. Occasionally, large parts of the signal are below the masking threshold, hence the quantization levels contain long runs of zero symbols. We apply run-length coding to encode the zero symbols, if it requires fewer bits than the separate encoding. An additional bit indicates whether run-length coding was applied in the given frame.

5 Z_n LATTICE BASED QUANTIZATION AND ENCODING OF THE WPT COEFFICIENTS

Sub-band number	1-8	9-14	15-17	18-21
Number of coeffs.	4	8	16	32
Lattice dimension	4	8	16	16

Table 1: Number of coefficients and VQ dimensions in the different WPT sub-bands.

Table 1 shows the number of coefficients per sub-band in a WPT frame. The masking threshold and the distribution of the coefficients are the same for all coefficients within a sub-band, but may vary among sub-bands. Therefore we quantize the coefficients in each sub-band separately. To limit the complexity of the implementation, we use maximum 16 dimensional lattices.

The quantization of a vector on the Z_{n_i} lattice is equivalent to uniform scalar quantizations in each of the n_i dimensions. The quantization levels l_{ij} are calculated using (3). They determine the lattice point $\mathbf{l}_i \in Z_{n_i}$ to which the coefficients $c_{i1}, c_{i2}, \ldots, c_{in}$ are quantized:

$$\mathbf{l}_i = [l_{i1}, l_{i2}, \ldots, l_{in_i}]. \quad (4)$$

5.1 Z_n lattice based encoding of the WPT coefficients

We suppose that the distribution of WPT coefficients in a sub-band follows a Laplacian distribution. We have estimated the real distributions using the method of [7]. We have found that the distributions of the WPT coefficients are more peaky than the Laplacian distribution, especially in the higher sub-bands. The encoding method described below does not exploit this propriety, thus it is sub-optimal in rate-distortion sense. However, it provides a low-complexity solution.

Supposing Laplacian distribution, the vectors having a constant l_1 norm define a hyper-pyramid of constant probability density [8]. Using this propriety, we can partition the points of constant l_1 norm of the Z_n lattice into shells containing points of equal probability.

To encode the lattice points, we use a simple form of entropy coding suggested by [9]. The first part of a product code, the l_1 norm of the lattice point, defines the shell which contains the lattice point. The second part of the product code identifies the point within the shell. We use a Huffman code to encode the l_1 norm. The position within the ith shell is encoded using codewords of length $\log_2(m_i)$, where m_i is the number of lattice points on the ith shell.

To find the index of a lattice point within a shell, we use the algorithm described in [10] for the case of Laplacian sources.

6 RESULTS

We have compared the performance of the scalar and the Z_n lattice based encoding methods using different speech and music samples. The same sinusoidal extraction and coding method was used in the two cases. As the quantization of the WPT coefficients is identical, and only the encoding of the quantized WPT coefficients is different, the audio quality is the same for both cases. Sound files demonstrating the compressed audio quality are avaliable at *www-sc.enst-bretagne.fr/~fek/demo/*.

The third row in Table 1 shows the vector dimensions used for the LVQ. As we have limited the maximum VQ dimension to 16, the VQs in the last four sub-bands are applied two times in a frame to encode all the 32 coefficients in these bands. We have measured the distribu-

tion of the l_1 norm length of the vectors within different sub-bands. The distributions within the sub-bands $1-8$, $9-14$ and $15-21$ were approximately identical. Therefore we have used only one Huffman table for each of the three sub-band groups. Figure 2 shows the histogram of the $l1$ norms in the different sub-bands for the Carmen test signal.

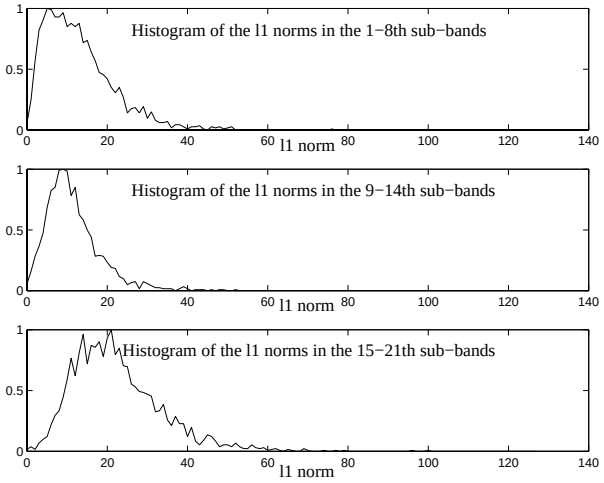

Figure 2: Histogram of the l_1 norms in different sub-bands for the Carmen test signal.

Coded signal	Scalar bit rate	LVQ bit rate
Ger. fem. speech	32.4/(4.7) kbps	32.0/(4.7) kbps
Eng. male speech	31.3/(1) kbps	30.1/(1) kbps
Carmen	55.6/(1.6) kbps	54.1/(1.6) kbps
Castanets	62.6/(0.4) kbps	53.8/(0.4) kbps
Singing	47.6/(4) kbps	45.2/(4) kbps
Rock	46.6/(2.6) kbps	46.1/(2.6) kbps
Bagpipe	56.5/(7.7) kbps	52.8/(7.7) kbps

Table 2: Mean bit rates (total/sinusoidal) for the scalar and Z_n lattice based encodings.

Table 2 shows the mean bit rates for the two encoding procedures. The most significant reduction was obtained for the castanets signal, of which bit rate was the maximum among the test signals.The maximum mean bit rate was reduced from 62.6 kbps to 54.1 kbps.

7 CONCLUSION AND PERSPECTIVES

In this paper, we have compared the performance of scalar and Z_n lattice based encoding of WPT coefficients in a combined sinusoidal and WPT model based coder for speech and music signals.

The Z_n lattice based encoding reduced the bit rate compared to scalar encoding. The quality of the encoded signal remained the same, as the quantization method was not changed.

Further bit rate reduction is possible by using a denser lattice than the Z_n latice. It is also possible to take advantage of the fact, that the distribution of the WPT coefficients is more peaky than the Laplacian distribution. However, it is an open question how to build a reasonable complexity encoder to exploit this propriety.

REFERENCES

[1] L. Tancerel, S. Ragot, V. T. Ruoppila, and R. Lefebvre, "Combined speech and audio coding by discrimination," in *IEEE Workshop on Speech Coding*, 2000, pp. 154–156.

[2] M. Fék, A.R. Várkonyi-Kóczy, and J-M. Boucher, "Joint speech and audio coding combining sinusoidal modeling and wavelet packets," in *Eurospeech 2001 - Scandinavia*, Aalborg, Denmark, September 3-7 2001, vol. 4, pp. 2311–2315.

[3] ISO/IEC IS11172-3, *Information technology - Coding of moving pictures and associated audio for digital storage media up to about 1.5 Mbit/s - Part 3 : Audio*, 1993, International Standard.

[4] X. Rodet, "Musical sound signals analysis/synthesis: Sinusoidal+residual and elementary waveform models," in *Proceedings of the IEEE Time-Frequency and Time-Scale Workshop (TFTS'97)*, Coventry, UK, 27th-29th August 1997.

[5] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-34, no. 4, pp. 744–754, August 1986.

[6] B. Carnero and A. Drygajlo, "Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms," *IEEE Transactions on Signal Processing*, vol. 47, no. 6, pp. 1622–1635, June 1999.

[7] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 1, pp. 52–56, February 1995.

[8] T. R. Fischer, "A pyramid vector quantizer," *IEEE Transactions on Information Theory*, vol. IT-32, pp. 568–583, July 1986.

[9] A. Woolf and G. Rogers, "Lattice vector quantization of image wavelet coefficient vectors using a simplified form of entropy coding," in *ICASSP-94*, 1994, vol. V, pp. V/269–V/272.

[10] J-M. Moureaux M. Antonini, P. Loyer, "Solving lattice codebook enumeration problem for generalized Gaussian sources," in *IEEE ISIT 2000*, Sorento, Italy, June 25-30 2000, p. 204.