An adaptive wavelet-based approach for perceptual low bit rate audio coding attending to entropy-type criteria

N. RUIZ REYES[1], M. ROSA ZURERA[2], F. LOPEZ FERRERAS[2], D. MARTINEZ MUÑOZ[1]

[1] Departamento de Electrónica
Universidad de Jaén
Escuela Universitaria Politénica.
C/ Alfonso X el Sabio, 28
23700 Linares - Jaén (SPAIN)

Departamento de Teoría de la Señal y Com.
Universidad de Alcalá
Escuela Politécnica
Ctra. Madrid-Barcelona, km. 33,6
28871 Alcalá de Henares - Madrid (SPAIN)

Abstract: - This paper outlines an adaptive wavelet-based perceptual audio coding scheme attending to various entropy-type criteria. Its performance using some different wavelet families and various filter lengths and decomposition depths has also been investigated. An optimal choice of these parameters is accomplished in order to evaluate both quality and bit rate of compressed signals for four different entropy-type criteria and four representative samples of audio material. The proposed coding scheme performs a periodized wavelet packet transform for each audio frame leading to a decomposition tree which is adapted to the characteristics of the audio frame attending to some entropy criterion. After time-frequency mapping, a thresholding to zero step is carried out to take advantage of entropy coding methods. Next, an uniform quantifier controlled by a psychoacoustic model taking advantage of the masking effect in human hearing is used. Finally, statistical redundancies of audio signals are reduced by using Huffman and run length coding. Experimental results indicate that the proposed approach can achieve almost transparent coding of monophonic CD quality audio signals at bit rates of approximately 64 kb/s (1.45 bit/sample). In addition, the use of the periodized wavelet transform leads to lower coding delay than other similar methods in the literature. The performance of our method is compared to some non-adaptive wavelet-based methods and to MPEG standard in terms of compression versus quality performance. IMACS/IEEE CSCC'99 Proceedings, Pages:3441-3445

Key-Words: - wavelet-based audio coding, psychoacoustic model, entropy-type criteria, MPEG

1 Introduction

Coding of high-fidelity audio signals has become a key technology in the development of audio systems. In many applications, such as the design of cost-effective multimedia systems and high quality audio transmission and storage, the goal is to achieve transparent (or nearly transparent) coding of hi-fidelity audio signals at the lowest possible bit rates.

Most audio coding algorithms rely on: 1) removal of statistical redundancies in the audio signal, and 2) exploitation of masking properties of the human auditory system to "hide" distortions. Traditional subband and transform coding techniques provide a convenient framework for coding based on these two principles. They also indicate that almost perceptually transparent coding of monophonic CD quality signals can be achieved approximately at bit rates of 96 kbps. Several of these techniques have contributed to the development of the ISO-MPEG [1] audio coding standard. More wavelet-based recent works include the adaptive wavelet selection method combined with dynamic dictionary coding [2], and the pitch-synchronous wavelet transform [3], which claim to achieve similar quality at bit rates of 64 kbps with fs=44.1 kHz and 21 kbps with fs=8 kHz, respectively. The disadvantage of these two methods is the long coding delay. This factor is very important for real-time coding applications.

In comparison to the above techniques, our approach, based on a adaptive wavelet packet transform (controlled by some entropy criterion) combined with hard thresholding, uniform quantization and entropy noiseless Huffman coding, claims perceptually transparent coding at similar bit rates but with shorter delay. The main music related applications of our audio coding scheme are: storage and editing of digital audio on small computers (home studio), computer-based multimedia, digital audio broadcasting (DAB), transmission via narrow-band ISDN for reporting links, and tele- or videoconferencing.

2 The proposed audio coding scheme

The audio coding scheme we outline here can be seen in figure 1 (encoder) and figure 2 (decoder), and it consists of the following stages:

• First of all, the input audio signal is divided into overlapping windowed frames of 2048 samples each one. This task is accomplished by an input buffer. Other frame lengths are also possible.
• Afterwards, each audio frame is decomposed in M subbands (being M a variable number) using an adaptive technique based on some entropy criterion. This method adaptively matches the decomposition tree to a given signal.
• In order to maximize the coding gain, the number of non-zero wavelet coefficients is reduced by using a hard thresholding stage. The thresholding level is computed by the following expression:

thr_lev = Cte*mean (abs (detail coeff. level 1) (1)

• A masking threshold is estimated for each audio frame to determine the inaudible quantization noise that can be added in each subband. The used psychoacoustic model assumes that the masking is an additive process. In this stage, we also estimate other parameters, such as the peak value in each subband.
• Previous estimates allow us to compute the step size necessary to quantify the wavelet coefficients within each subband without any noticeable noise. In our scheme, the encoder consists of two steps: uniform quantization and entropy noiseless coding.
• Input audio signal is represented by their coded wavelet coefficients. These coefficients are multiplexed with the side information, which includes an index to represent the selected best basis, giving rise to the audio bit stream transmitted to the decoder.

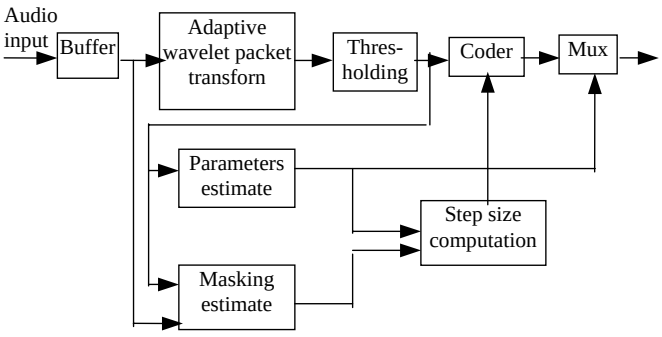

Fig 1: Encoder structure.

• The decoder receives the audio bit stream and, by demultiplexing, extracts the side information and the coded wavelet coefficients. From the side information, convenient step sizes to decode wavelet coefficients are obtained.
• The inverse wavelet packet transform recovers the output signal from the decoded wavelet coefficients.

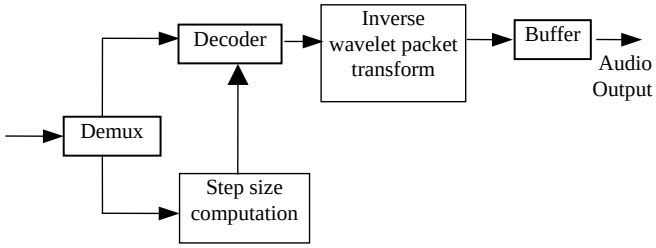

Fig. 2. Decoder structure

2.1 The adaptive wavelet packet transform

In discrete-time signal processing, the wavelet transform can be implemented using only two filters (high-pass and low-pass filters), that must satisfy certain orthogonality conditions to constitute a perfect reconstruction filter bank [4]. The one-step transform shown in figure 3 can be iterated over the low-pass signal (or approximation signal) to obtain higher resolution in the frequency domain, resulting the so called “discrete wavelet decomposition”.

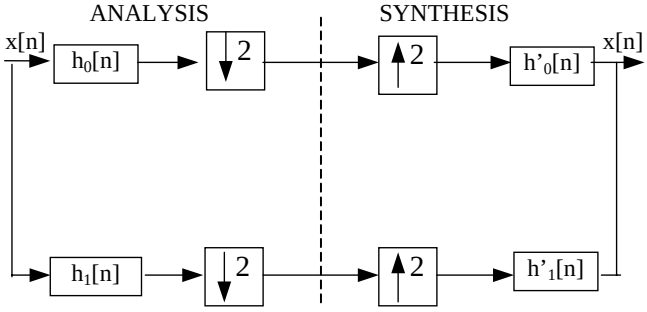

Fig. 3. Filterbank to implement one-step wavelet transform

To adapt the wavelet transform to the analysis made by the human hear, it can be used a “wavelet packet decomposition”, that is a generalization of wavelet decomposition and offers a richer signal analysis. The wavelet packet decomposition is free to continue in upper and lower bands, as it is required, without losing the orthogonality and perfect reconstruction features.

This type of decomposition is represented by a binary tree in which one has the freedom to stop or continue the decomposition at any node. Several choices for a basis are thus possible, and it is interesting to find an optimal decomposition with respect to a convenient criterion, computable by an efficient algorithm [5]. We are looking for a minimum of the criterion. Classical entropy-based criteria are well suited for efficient searching of binary-tree structures and describe information-related properties for an accurate representation of a given signal.

In this paper, we evaluate four different entropy criteria (many others can also be used) to match the decomposition tree to a given data set (an audio frame). The entropy E must be an additive cost function, such that $E(0)=0$ and $E(s)=\text{sum}(E(s_i))$, where s is the signal and s_i the coefficients of s in a basis. The entropy criteria used here are:

- The (non-normalized) Shannon entropy.

$$E_1(s) = -\text{ sum } (s_i^2 \log (s_i^2)) \quad (2)$$

- The concentration in l^p norm with $1 \le p < 2$.

$$E_2(s) = \text{sum } (|s_i|^p) = ||s||^p \quad (3)$$

- The logarithm of the "energy" entropy.

$$E_3(s) = \text{sum } (\log (s_i^2)) \quad (4)$$

- The threshold entropy.

$$E_4(s) = 1 \text{ if } |s_i| > \varepsilon \text{ and } 0 \text{ elsewhere} \quad (5)$$

Starting with the root node, the best tree for each audio frame is calculated using the following scheme: a node N is split into two nodes N_1 and N_2 if and only if the sum of the entropy of N_1 and N_2 is lower than the entropy of N. This is a local criterion based only on the information available at the node N. For each node splitting, we use the periodized (periodical extension) wavelet transform. In order to avoid the clicks due to the border effect inherent to this kind of transform, the audio data is first divided into overlapping analysis frames. Each frame is then windowed with a trapezoidal window. Experimental tests tell us that only a small overlapping (20 samples) is sufficient to eliminate the clicks.

2.2 The psychoacoustic model

Masking is a psychoacoustic fenomenon that renders low-level signals concentrated in a given frequency region inaudible in the presence of higher signals at neighboring frequencies. The masking model used here is based on the well-known method developed by Johnston [6], but with some notable improvements. It relies on the computation of:

1) Summations of signal energy over frequency regions corresponding to *critical bands* of the auditory human system.
2) A cochlear *spreading function*, which describes the effects of signal energy in one critical band over the masking effects in adjacent bands.
3) A measure of *tonality* for each spectral component calculated.
4) A *masking threshold* is obtained based on the previous frequency domain analysis.

This masking threshold is defined in the Fourier domain. It must be translated to the wavelet domain to determine a perceptual upper bound on the quantization noise power that can be tolerated in each frequency band for each given audio frame. To incorporate the psychoacoustic information to the wavelet domain, we perform a renormalization process [7] (instead of deconvolution, which may cause instability). Another approach to translate the masking threshold condition to the wavelet transform domain is based on a perceptual norm criterion [2], but it has the disadvantage of the need for long and very selective filters to implement the wavelet transform. The adopted approach in our coding scheme avoids this fact and makes possible to translate the psychoacoustic information into the wavelet domain using any kind of wavelets.

3 Experimental results

Now, let us discuss some experimental results that we have obtained with the proposed coding scheme. Throughout the simulations presented in this section, a five level decomposition is always performed, because it allows transparent audio coding with the lowest complexity and delay. The set of audio source material used to check the performance of our coding scheme is listed in table 1.

20 kHz source material sampled at 44.1 kHz, 16 bit PCM	
CODE	INSTRUMENT / STYLE
Vocal	Female vocal pop song
Wind	Wind instruments
Violin	Violin with orchestra
Piano	Solo piano

Table 1. List of audio source material used in the tests

3.1 Objective quality measure

We have chosen the segmental SNR measure to provide an objective measure of the performance of our coding scheme.

This measure is more correlated with the subjective quality measures than a single SNR computed for the whole audio signal.

Table 2 shows the behaviour of the four entropy criteria described in section 2 for the audio material listed in table 1.

	Vocal	Piano	Wind	Violin
Shannon	1,48 / 28,10	1,56 / 23,30	1,40 / 24,17	1,40 / 26,20
Log energy	1,47 / 28.00	1,53 / 22,85	1,40 / 24,18	1,41 / 26,40
Threshold	1,49 / 28,05	1,52 / 22,31	1,49 / 24,90	1,30 / 26,13
Norm (p=1)	1,47 / 28,00	1,58 / 23,34	1,40 / 24,05	1,40 / 26,39

Table 2. Performance of different entropy-type criteria. Bit rate (bit/sample) / segmental SNR (dB) values.

From table 2, we can deduce that all the entropy criteria evaluated have a similar behaviour, and therefore in the rest of the simulations we have used Shannon criteria. Figure 4 shows the performance of our coding scheme with the number of filter coefficients for different wavelet families. The wavelet families considered for comparison are: minimum-phase Daubechies wavelets (DAUB), orthogonal Coiflet wavelets (COIF) and biorthogonal spline wavelets (BIOR). These wavelet families are asymmetric, near symmetric and symmetric, respectively. They all are compactly supported.

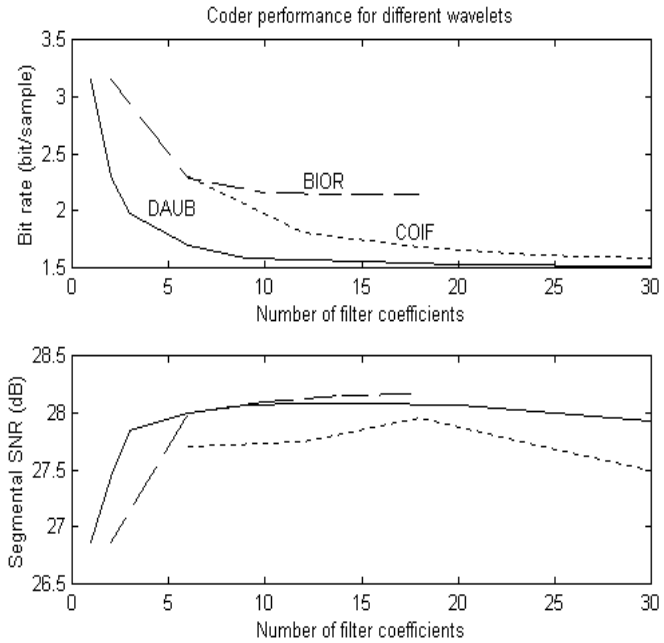

Fig. 4. Bit rate and segmental SNR versus filter length for different wavelet families. Audio source code: Vocal.

From figure 4, we see that as the number of filter coefficients increases the coder performance improves. However, this improvement appears to reach an asymptote for filter lengths greater than 20 coefficients. For higher filter lengths, the subjective quality of the encoded signals decreases. So, we choose this value as an optimal filter length. If we compare the behaviour of the three evaluated wavelet families, it can be deduced that DAUB and COIF wavelets provide similar results for the optimal filter length, and they both make better than BIOR wavelets. Therefore, we are interested in orthogonal wavelets, which agrees with other works in the literature [8].

Finally, in table 3 we compare our coding approach to some others wavelet-based. The DAUB family has been chosen for the experiments here performed after an evaluation of the performance of various wavelet families. All the methods considered for comparison make use of the same psychoacoustic model and are based on:

A) Non-periodized critical band WPT.
B) Periodized critical band WPT.
C) Periodized critical band WPT (with overlapping)
D) Periodized adaptive WPT (with overlapping)
E) Periodized adaptive WPT (with overlapping, thresholding and entropy coding).

	Vocal	Piano	Wind	Violin
A	1,72 / 25,8	1,65 / 23,06	1,47 / 24,24	1,54 / 27,00
B	2,69 / 25,26	1,70 / 22,00	2,33 / 23,23	2,48 / 25,70
C	1,99 / 24,71	1,89 / 22,00	1,44 / 23,23	1,51 / 25,60
D	2,20 / 28,57	2,15 / 24,08	1,65 / 24,34	1,69 / 26,25
E	1,48 / 28,10	1,56 / 23,30	1,40 / 24,17	1,40 / 26,21

Table 3. Performance of different wavelet-based coding methods. Bit rate (bit/sample) / segmental SNR (dB).

It can be seen that the method who provides better results for the whole audio material is our coding method (here coded as E). Besides, it must be noted the substantial improvement due to overlapping. It is also interesting to emphasize the influence of the adaptive WP decomposition for the good behaviour of our coder.

Comparing our approach to the one based on non-periodized WPT (here coded as A), we find our approach to be superior in two important aspects: coding delay and subjective quality.

The results shown in table 3 reveal that our coder is better adapted to signals with sharp attacks than to nearly steady signals. The best results are obtained for vocal source and the worst for piano source.

3.2 Subjective quality measure

We have carried out an informal subjective test with a reduced group (20 persons of our research group). Two types of listening tests have been performed at a binary rate of about 64 kbps:

a) Test of transparency
b) Comparison with MPEG layer 2 and 3.

The results confirm that the proposed coder achieves nearly transparent coding with all the audio sources evaluated. Also, its quality is better than the MPEG layer 2 one, where a distortion such as a filtering effect has been observed at this binary rate, and similar than the MPEG layer 3 one.

The quality of the piano signal encoded with the proposed coding approach was not as good as that of the other audio pieces. The piano sample contains long segments of nearly steady sinusoids. The wavelet-based coder does not seem to handle steady sinusoids as well as other kinds of signals. It needs to be further optimised for such signals.

4 Conclusion

We have just presented a novel audio coding method based on adaptive optimal wavelet decomposition. Our studies indicate that optimization of the wavelet decomposition to match the audio data clearly results in a significant reduction in the bit rate requirement for the same audio quality.

The main advantages of the proposed audio coder are: low bit rate requirement for nearly transparent audio compression, lower complexity and delay compared to other wavelet-based coders [3], and high flexibility to be used in diverse applications, like multimedia communications.

Several improvements in the proposed method are possible in terms of reducing its computational complexity and its bit rate requirements. The most promising approach to bit rate reduction undoubtedly involves vector quantization of groups of wavelet coefficients [9][10]. This will be one of the focus of our future studies.

An interesting way to reduce the computational complexity of our approach would be the development of a masking model directly in the wavelet domain. This will be another research issue to work on. Also, it would be insteresting to check the performance of our scheme with different entropy coding methods [10][11].

Other issues to work on are scalable coding, stereo and multichannel coding, optimal choice of the wavelet family for each audio frame, evaluation of different psychoacoustic models and more extensive subjective quality evaluation.

References:

[1] ISO/IEC 11172-3, "Information technology - Coding of moving pictures and associated audio for digital storage media at up to 1.5 Mbit/s" - (Part 3, Audio), 1991.

[2] D. Sinha, A. H. Tewfik, "Low Bit Rate Transparent Audio Compression Using Adapted Wavelets", *IEEE Trans. on Signal Processing*, Vol. 41, No. 12, 1993, pp. 3463-3479.

[3] G. Evangelista, "Pitch Synchronous Wavelet Representations of Speech and Music Signals", *IEEE Trans. on Signal Processing*, vol. 41, No. 12, 1993, pp. 3313-3330.

[4] M. Vetterly, "Wavelets and Filter Banks: Theory and Design", *IEEE Trans. on Signal Processing*, Vol 4, No. 9, 1992.

[5] R. R. Coifman, M. V. Wickerhauser , "Entropy–Based Algorithms for Best Basis Selection", *IEEE Trans. on Information Theory*, Vol. 38, No. 2, 1992, pp. 713-718.

[6] J. D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria", *IEEE Journal on Selected Areas on Communications*, Vol. 6, No. 2, 1988, pp. 314-323.

[7] M. Rosa, "Optimization of Signal Processing Algorithms Applied to Audio Compression", Ph.D. Thesis, Alcala University, 1998.

[8] C. W. Kok, T. Q. Nguyen, "Multirate Filter Banks and Transform Coding Gain", *IEEE Nordic Signal Processing Sympossium*, 1996.

[9] W. Y. Chan, A. Gersho, "High Fidelity Audio Transform Coding with Vector Quantization", *Proc. ICASSP*, 1990, pp. 1109-1112.

[10] A. Gersho, R. M. Gray, "*Vector Quantization and Signal Compression*", Kluwer Academic Publisher, 1992.

[11] I. H. Witten, R. M. Neal, and J. G. Cleary, "Arithmetic coding for data compression", *Commun. ACM*, Vol. 30, 1987, pp. 520-540.