A NEW COST FUNCTION TO SELECT THE WAVELET DECOMPOSITION FOR AUDIO COMPRESSION

N. Ruiz Reyes[1]*, M. Rosa Zurera*[2]*, F. López Ferreras*[2]*, D. Martínez Muñoz*[1]*, and J.M. Villafranca*[2]

[1]Departamento de Electrónica, Universidad de Jaén, Escuela Universitaria Politénica,
23700 Linares, Jaén, SPAIN, e-mail: nicolas@ujaen.es
[2]Dpto. de Teoría de la Señal y Comunicaciones, Universidad de Alcalá, Escuela Politécnica
28871 Alcalá de Henares, Madrid, SPAIN, e-mail: manuel.rosa@uah.es

ABSTRACT

This paper outlines a wavelet-based perceptual audio coding scheme that searches for the wavelet packet decomposition that minimizes a new perceptual entropy-type cost function computed in the wavelet domain. We are interested in decompositions adapted to the nature of audio signals, but attending to the characteristics of the human hearing system. We present results about audio coding with three different decomposition strategies. The first one uses a fixed wavelet tree structure which closely mimics a critical band decomposition, while the second and the third ones use adaptive wavelet trees obtained minimizing different cost functions. The results confirms that perceptual entropy based decomposition is the way you must take if you are looking for maximum compression rates and transparent coding. To achieve a further bit rate reduction ensuring transparent coding, our audio coder performs a filter order optimization, which is possible by using a novel algorithm to translate the psycho-acoustic information from the frequency to the wavelet domain. Experimental results indicate that our coder ensures transparent coding of monophonic CD quality audio signals at bit rates lower than 64 kbps.

1. INTRODUCTION

Coding of CD quality audio signals has become a key technology in the development of current audio systems. CD quality monophonic audio signals are obtained with sampling frequencies of 44.1 kHz and 16 bits PCM coding. So, it is necessary a binary rate of 705.6 kbps for transmission, justifying the research and development of efficient audio coding techniques in order to reduce this high transmission rate. In many applications, such as high quality audio transmission and storage, the goal is to achieve transparent coding of CD quality audio signals at the lowest possible bit rates.

Most audio coding algorithms are based on: 1) removal of statistical redundancies in the audio signal, and 2) masking properties of the human auditory system to “hide” distortions. Traditional subband and transform coding techniques provide a convenient framework for coding based on both principles.

Several of these techniques have contributed to the development of the ISO/MPEG audio coding standards. The first one, called ISO/MPEG-1 [1], supports sampling rates of 32, 44.1 and 48 kHz, and several operation modes with bit rates ranging from 32 to 448 kbps. The last one, the ISO/MPEG-4 standard, is composed of several speech and audio coders supporting bit rates from 2 to 64 kbps per channel. ISO/MPEG-4 includes the already proposed AAC standard, which provides high quality audio coding at bit rates of 64 kbps per channel.

Parallel to the definition of the ISO/MPEG standards, several audio coding algorithms have been proposed that use the wavelet transform as the tool to decompose the signal. The most promising results correspond to adapted wavelet-based coders. Probably, the most cited is the one designed by Sinha and Tewfik [2], a high complexity audio coder that provides low bit rate and high quality audio coding by searching a nearly optimum prototype filter for each audio frame. Unlike it, our approach searches for the wavelet tree structure that gives the minimum bit rate ensuring transparent coding, taking as cost function a perceptual entropy measure defined in the wavelet domain. In [3] a perceptual cost function is also used, but it’s defined in the frequency domain rather in the wavelet one, which does not work properly when short filters are used.

Our coder achieved transparent coding at slightly higher bit rates than those of Sinha and Tewfik’s coder, but with shorter coding delay. Furthermore, it can operate with orthonormal wavelets of any compact support, which allows the filter order optimization already mentioned.

2. AUDIO CODER STRUCTURE

In this section, the audio coder structure is described. It works with monophonic audio signals sampled at 44.1 kHz, but can be easily extended for multi-channel audio signals. Each input sample is PCM coded with 16 bits.

The objective is obtaining a digital representation of the original signal, with the minimum size, preserving as much as possible its perceptual quality. The encoder structure is summarized in figure 1.

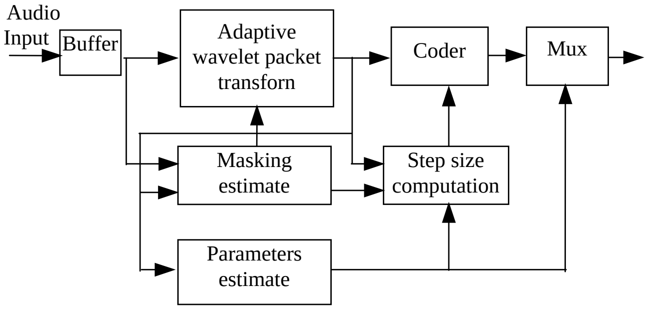

Figure 1. Encoder structure

The main features of the audio coder we propose here are:

1. The input signal is analyzed with a time-varying filter bank that implements a wavelet packet decomposition adapted to the source (audio signals) using a perceptual cost function.
2. Due to the fact that audio signals are not stationary, input signals are segmented in frames of 1024 samples, that can be modeled as windowed samples of a random process.
3. To avoid that the number of wavelet coefficients that characterize an audio frame is higher than the number of samples in the time domain, each frame is interpreted as a periodical signal, implementing a periodized wavelet packet decomposition.
4. Adjacent frames overlap 1/64 of their length, in order to avoid sharp changes in the injected quantization noise power. The overlapping samples of each frame are windowed with the square root of a raised cosine function.
5. Parallel to the decomposition of the input signal, a masking threshold is estimated in the frequency domain for each audio frame. We have used the ISO/MPEG-1 psycho-acoustic model 2.
6. The estimated masking threshold in the frequency domain is not suitable to be applied directly in the wavelet domain, mainly when short filters are used to implement the wavelet transform. We have developed a novel algorithm to translate the psycho-acoustic information from the frequency to the wavelet domain [4].
7. We use block companding with a set of 15 uniform quantizers. The scale factors, which are sent to the receiver as side information together with the bit allocation and the wavelet tree structure information, are coded using 8 bits log PCM.
8. The quantized wavelet coefficients are entropy coded using an adaptive Huffman coding method based on laplacian modeling for subband audio signals [5][6]. The model parameter for each subband audio signal is coded using 5 bits log PCM.
9. At the decoder, the coded wavelet coefficients and the side information are demultiplexed and decoded. Later, each the overlapping samples of the reconstructed frame are windowed again and, after the overlap-add process, the perfect reconstruction property is preserved.

The decoder structure is represented in figure 2.

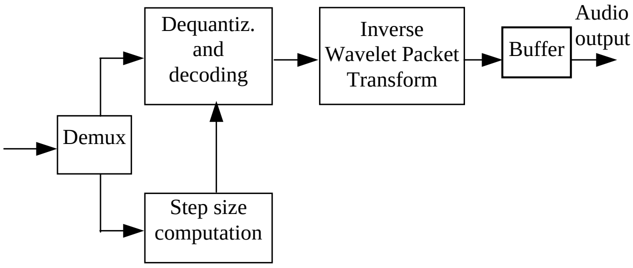

Figure 2. Decoder structure

3. ADAPTIVE WAVELET ANALYSIS

The best advantage of wavelet packet decompositions is the possibility of selecting the optimum binary tree to analyze a given signal. The choice of the optimization criteria will depend on the application where the decomposition is applied. An important fact is the complexity of the best basis search procedure. From this point of view, the algorithm to select the best basis must be as much simple as possible, because a high complexity one is not suitable in practical situations.

Usually, to select the optimum wavelet packet base, a cost function is defined that must be minimized by the search algorithm. Among the cost functions, a very important set is composed of those that are "additive". A cost function $\Im$ is additive over a sequence $\mathbf{x} \in \Re^N$ if the following relations hold:

$$\Im^{add}(0) = 0$$
$$\Im^{add}(\mathbf{x}) = \sum_{i=1}^{N} \Im^{add}(x(i)) \qquad (1)$$

Among the additive cost functions, several classical entropy-type functions must be considered, such as the non-normalized Shannon entropy, the concentration in the l^p norm ($1 \leq p < 2$), the logarithm of the "energy" entropy and the threshold entropy [7].

These entropy-based functions are well suited for efficient searching of binary-tree in classical source coding applications, because they describe information-related properties for an accurate representation of a given signal. However, in perceptual audio coding we are interested in those decompositions that minimize the binary rate allowing a high perceptual quality of the reconstructed signal, instead of a classical measure as the signal to noise ratio.

Therefore, the cost function must have into account psycho-acoustic information in the transform domain (wavelet domain in our particular case).

Here, we use a cost function related to the perceptual entropy, defined by Johnston as the minimum number of bits per sample in order to achieve transparent coding, and try to minimize it. We expect to obtain a wavelet packet decomposition tree, which allows minimizing the binary rate maintaining the perceptual quality of the coded signal.

Given a subband signal with N samples, the proposed cost function is defined as:

$$\Im(\mathbf{x}) = N \cdot \log_2\left(max(SMR)\right) \tag{2}$$

In this relation, SMR represents the signal to mask ratio of that subband signal in the wavelet domain. The signal to mask ratio is estimated in each subband as the ratio between the variance of the subband signal and the variance of quantization noise that can be introduced in that subband to ensure transparent coding. It is an indication of the number of bits needed to represent the input signal ensuring transparent coding.

To obtain the variance of quantization noise that can be introduced in each subband to ensure transparent coding, we propose the following algorithm:

1) We first implement a complete wavelet packet decomposition with a given depth (i.e. a five levels wavelet packet decomposition).
2) Each node of the complete wavelet tree is labeled with a value that comes to represent the masking threshold in the wavelet domain for that node. This node value is obtained by using the following expression [4]:

$$\sigma_i = \underset{\omega \in B_i}{\text{median}} \left(\frac{T(e^{j\omega})}{\sum_{k=1}^{M} \left| F_k(e^{j\omega}) \right|^2} \right) \tag{3}$$

Where:

- B_i is the set of frequency bins inside the i-th sub-band.
- $T(e^{j\omega})$ is the masking threshold estimate in the frequency domain.
- M is the number of subbands in the complete wavelet tree at a given decomposition depth.
- $\left|F_k(e^{j\omega})\right|$ is the equivalent filter frequency response magnitude of the synthesis filter bank k-th branch.
- To obtain a representative value, a median operator is used, instead of a minimum one, which avoid outliers.

The best tree structure for each audio frame is obtained with the following algorithm:

1. We begin at the root node. The initial cost is that of the input audio frame ($\Im(\mathbf{x}_i)$).
2. We decompose the signal using a two bands perfect reconstruction filter bank using orthonormal wavelet filters.
3. We compute the cost of the nodes obtained in the decomposition ($\Im(\mathbf{x}_j)$ and $\Im(\mathbf{x}_k)$).
4. If $\Im(\mathbf{x}_i) \geq \Im(\mathbf{x}_j) + \Im(\mathbf{x}_k)$, then
   - The algorithm is repeated with node *j*.
   - The algorithm is repeated with node *k*.

   else, we come back in the recursion.

Starting with the root node, the best tree for each audio frame is calculated using the following scheme: a node *i* is split into two nodes *j* and *k*, if and only if the sum of the cost function evaluated with subband signal $x_j[n]$ and $x_k[n]$ is lower than that of node *i*.

This algorithm is a growing-based one that leads to sub-optimal solutions. Instead of it, we propose a pruning-based algorithm that gives rise to optimal solutions.

4. EXPERIMENTAL RESULTS

To check the performance of our audio coder, we have obtained some objective and subjective results. Five music samples considered hard to encode have been used, and we have made sure that the set covers a wide variety of signals. It consists of about 15 seconds of drums, guitar, piano, saxophone and pop music.

Special attention has been paid to signals with sharp attacks, like ‘drums’ and ‘guitar’, as these signals are extremely susceptible to the presence of ‘pre-echo’.

4.1 Objective results

Table 1 shows a comparative analysis between three different wavelet-based audio coders: our coder (A), a coder that implements a wavelet packet decomposition close to the critical band one at the inner ear (B), and a coder that searches for the best wavelet tree using the Shannon entropy as cost function (C).

This table represents for every signal and every coding algorithm the minimum binary rate to achieve transparent coding, measured in bits/sample, and the corresponding segmental SNR, measured in dB’s. In the experiments minimum phase Daubechies filters with 24 coefficients are used.

	A	B	C
DRUMS	1,50 / 16,00	1,70 / 15,86	1,78 / 16,02
GUITAR	1,45 / 30,23	1,68 / 27,86	1,73 / 31,00
PIANO	1,38 / 26,88	1,58 / 24,48	1,67 / 28,00
SAXO	1,48 / 23,17	1,59 / 19,96	1,75 / 23,13
POP	1,44 / 21,47	1,70 / 21,37	1,72 / 21,84

Table 1. Comparative analysis of three decomposition strategies

From table 1, it can be observed that the proposed cost function achieves transparent coding with about 0.25 bits per sample less than the Shannon entropy. Other entropy-type cost functions achieve similar results that those of Shannon entropy [8]. In an intermediate position is the coder B, that doesn't implement an adaptive decomposition, but attends to the critical band decomposition at the inner ear.

It must be remarked that the results shown in table 1 are obtained when filter order optimization is implemented. The optimization stage, which is possible because of using our algorithm to translate the psycho-acoustic information from the frequency to the wavelet domain, leads to a bit rate reduction of about 0.15 bit per sample.

4.2 Subjective results

We have performed a test for transparency ("double blind test") at a binary rate of approximately 64 kbps with 20 people selected from our research group, all of them aged from 24 to 35 years. The results when using the above mentioned filters are presented in table 2.

Music Sample	Average probability of original music preferred over encoded one	Comments
Drums	0.56	Transparent
Guitar	0.48	Transparent
Piano	0.47	Transparent
Saxophone	0.54	Transparent
Pop	0.51	Transparent

Table 2. Subjective test results: transparency test

The quality is cataloged as 'transparent' because the average probability is around 0.5 for all the test signals. The results confirm that our coder can be considered as transparent at a binary rate of 64 kbps.

5. SUMMARY

We have presented a new cost function in order to search the best wavelet packet decomposition for audio coding purposes. The results confirms that suitable searches must have into account psycho-acoustic information in order to minimize the resulting binary rate maintaining the perceptual quality as similar as possible to the original one.

Using this cost function to implement an adapted wavelet decomposition, it's possible to accomplish a notable reduction in the bit rate requirement for the same audio quality. Besides, as it has been shown, adaptive wavelet analysis is only interesting for audio coding purposes if psycho-acoustic information is considered in the tree building stage.

Two promising approaches for further bit rate reduction are: 1) vector quantization, 2) model based audio coding. We are now focused on these two issues.

Also, it would be interesting to check the performance of our scheme with other entropy coding methods (i.e arithmetic coding). Other issues to work on are scalable wavelet-based coding, evaluation of different psycho-acoustic models and more extensive subjective quality evaluation.

6. 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), 1992.

[2] D. Sinha, A. H. Tewfik, "Low bit rate transparent audio compression using adapted wavelets", *IEEE Trans. on Signal Processing*, Vol. 41, No. 12, pages 3463-3479, 1993.

[3] P. Srinivasan, L. H. Jamieson, "High-quality compression using an adaptive wavelet packet decomposition and psychoacoustic modeling", *IEEE Trans. on Signal Processing*, Vol. 46, No. 4, pp. 1085-1093, April 1998.

[4] M. Rosa, F. López, P. Jarabo and S. Maldonado, "New method to translate the psycho-acoustic information to the wavelet domain", *EURASIP Conf. DSP for Multimedia Commun. and Services*, Krakow, June 1999.

[5] N. Ruiz, M. Rosa, F. López, P. Jarabo and P. Vera, "On the coding gain of dynamic Huffman coding applied to a wavelet-based audio coder", *IEEE Nordic Signal Processing Symposium*, June 2000.

[6] P. Philippe, F. M. de Saint-Martin and M. Lever, "Wavelet packet filter banks for low time delay audio coding", *IEEE Trans. on Speech and Audio Processing*, Vol. 7, No. 3, pages 310-322, May 1999.

[7] R. R. Coifman and M. V. Wickerhauser, "Entropy–based algorithms for best basis selection", *IEEE Trans. on Information Theory*, Vol. 38, No. 2, pages 713-718, 1992.

[8] N. Ruiz, M. Rosa, F. López and D. Martínez, An adaptive wavelet-based approach for perceptual low bit rate audio coding attending to entropy-type criteria, *Software and Hardware Engineering for the 21th Century*, WSES Press, 1999.