Journal of ELECTRICAL ENGINEERING, VOL. 56, NO. 3-4, 2005, 100–105

REAL–TIME PERFORMANCE MEASURES OF LOW DELAY PERCEPTUAL AUDIO CODING

Eros Pasero — Alfonso Montuori *

The selection of the wavelet packet best basis of a perceptual real-time audio encoder has been a challenging task for researchers. Due to the subjective nature of the auditory masking phenomenon, fully objective performance measures (entropy versus SNR, delay, *etc*) are not sufficient for the selection of optimal representations. In our work we present a DSP application by means of which we can measure, for each hybrid wavelet packet basis resulting from a predetermined library of wavelet filters, the bitrate of a perceptually transparent coding of the audio input and the overall filter bank delay. The novelty is that the transparency of the compression is established with a subjective evaluation. The tool user can customize both the filter bank parameters and the psychoacoustic model parameters by means of a MATLAB® graphical interface and listen in real-time to the reconstructed signal. The audio codec runs in real-time on a TI TMS320C6711™ DSK.

K e y w o r d s: Psychoacoustics, perceptual audio coding, wavelets, subband coding, real-time audio coding, DSP

1 INTRODUCTION

The energy spectrum of 48 kHz sampled audio signals is uneven and time varying. Linear prediction based algorithms which have been used effectively in narrowband speech coders to achieve high quality speech at low bit rates are inefficient for wideband audio coding because they attempt to match all frequencies equally well. These considerations provide motivation for using adaptive subband coding. The basic idea is to filter the signal into subbands and to treat each subband differently on the basis of the energy distribution of the current speech frame. Moreover, a subband subdivision is mandatory if we want to take advantage of the masking properties of the human auditory system. By studying the limitations of auditory perception, particularly how it reduces the information rate of the received signal through time and frequency masking constraints, improvements can be made in the efficiency of audio coding schemes without impacting the quality of the reconstructed signal.

It has been shown that wavelets can approximate time varying non-stationary signals in a better way than the Fourier transform representing the signal on both time and frequency domains [1]. Hence they can easily detect local features in a signal. Furthermore, wavelet decomposition allows analyzing a signal at different resolution levels. Due to these properties wavelet filters concentrate speech information into a few neighbouring coefficients, which is essential for a low bit rate representation. In fact, in this way we can advantageously encode them using an entropy coding system.

The discrete wavelet transform (DWT) is usually implemented using an octave-band tree structure. This is accomplished by dividing each sequence into a component containing its approximated version (low-frequency part) and a component with the residual details (high-frequency part) and then iterating this procedure at each stage only on the low-pass branch of the tree. The main drawback of the octave-band tree structure is that it does not provide a good approximation of the critical subband decomposition of the human auditory system. We can obtain a better approximation with the discrete wavelet packet transform (DWPT) [3]. In the wavelet packet transform one can also split the high-frequency part in a low and high frequency parts. By iterating this procedure we can get different representations of the signal that have different time and frequency resolutions.

Several studies and algorithms have been proposed which address the issue of designing the best wavelet packet filter banks for perceptual audio compression [4, 5]. These algorithms are mainly based on fully objective efficiency measures (bitrate, delay, *etc*) that are calculated for a large set of different tree-structured wavelet packet configurations with a coding scheme obeying to the constraints of a predetermined psychoacoustic model. The entropy of the coded signal, *ie* , the lowest rate at which transparent coding is possible is referred to as perceptual entropy [6]. Quite often the predetermined psychoacoustic model is not equally optimized for all the different tree-structured wavelet packet configurations, first of all because the tested filter banks have different time localization and frequency selectivity. In other words this approach does not guarantee the perceptual transparency of the compression for all the tested configurations.

We propose a tool to compare the performances of different filter bank configurations which have proven to produce perceptually transparent coding. Our first aim is

* Politecnico di Torino, Dipartimento di Elettronica, C.so Duca degli Abruzzi, 24, 10100 Torino, Italy,
e-mail: {eros.pasero, alfonso.montuori}@polito.it

ISSN 1335-3632

the validation of optimal filter bank configurations which achieve transparency with minimum bitrate maintaining the filter bank delay below a fixed threshold. By means of a GUI (Graphical User Interface) the user can customize the codec by changing both the encoding parameters (the decomposition tree, the filters for each stage of the decomposition tree, *etc*) and the psychoacoustic model parameters establishing in real-time the transparency of the compression with a subjective evaluation. Once we get a perceptually transparent compression, it makes sense to observe the objective performances of the coder: estimated bitrate and overall filter bank delay.

TMS320 DSPs have proven to be efficient for real-time compression of audio signals [7–10]. The codec algorithm was first written in C and then optimized on the TMS320C6000 DSP platform. The graphical interface was written in MATLAB. The codec communicates in real-time with the graphical interface by means of a Real Time Data Exchange $\left(\text{RTDX}^{\text{TM}}\right)$ link [11].

2 FILTER BANK SELECTION

2.1 Subband Decomposition

The continuous variation in the frequency response of the human ear can roughly be described as a band pass filter bank consisting of overlapping band pass filters. If we want to make use of the spectral masking properties of the human ear, the most natural subband decomposition is a tree structure chosen so that the resulting subbands mimic the critical bands of the human auditory system (*Bark frequency Scale*) [12].

However, the critical band decomposition is not guaranteed to be an optimal decomposition in the sense defined above. If, on the one hand, the choice of the closest possible match to the critical band decomposition guarantees the maximum exploitation of the masking model, on the other hand the truncation of the tree structure at some level before reaching the critical band decomposition brings down the overall filter delay and the bitrate overhead for the transmission of side information.

In order to compare the performances of different decompositions, the GUI allows the user to select an arbitrary dyadic tree-structured subband decomposition (see Fig. 1).

The psychoacoustic model calculates the masking thresholds for each critical band (see the next section). In order to achieve a perceptually transparent compression, the coding algorithm assumes that for each subband selected by the user the noise threshold to be considered is the minimum of all constituent critical bands (including those critical bands which are not totally included in the user subband, but partially overlap it). The transform domain coefficients were quantized by means of a 6-bit uniform mid-tread quantizer so that the quantization error did not exceed the noise threshold.

2.2 Filter Selection

The GUI allows also the user to select a filter for each level of the transform within a preset wavelet packet filter library (see Fig. 1). The DWPT generalization which includes the choice of an optimal wavelet transform at each level of the decomposition is referred to as Hybrid Wavelet Packet Transform [13]. At each stage of the decomposition the choice of the filters of the transform as well as their length influence the separation of the subband signals, the time resolution, the filter bank delay and the compression performances.

The compression algorithm is optimized in a basis that produces few high amplitude coefficients and many small coefficients. A high amplitude wavelet coefficient occurs when the wavelet has a support that overlaps a transient or a sharp attack. The number of high amplitude coefficients created by a transient depends on the length of the high pass analysis filter impulse response, which should thus be as short as possible. Furthermore, also synthesis filters' impulse response must be compact in order to prevent excessive time spreading of quantization errors during synthesis (perceptible, for example, as the so-called "pre-echo" problem) [14]. Over smooth regions, the wavelet coefficients are small at high frequencies if the wavelet has enough vanishing moments. However the support size of the wavelet increases proportionally to the number of vanishing moments. The choice of an optimal wavelet is therefore a trade-off between the number of vanishing moments and its support size [1]. The advisability of a trade-off is strengthened by the fact that if, on the one hand, the filter bank must be frequency selective in order to make a correct use of the masking properties in different subbands and avoiding unmasked aliased quantization noise in side lobes of the subband filter [15], on the other hand analysis and synthesis filters' lengths must be controlled so as to keep the overall delay acceptable. In order to reduce the filter bank delay a commonly used approach is to reduce gradually the filter lengths through the cell structure, even though assigning to the first stages of the decomposition the filters with the longer impulse response is not always the best choice [4].

The wavelet packet library was constructed in a previous fully objective and computationally expensive selection step (in which we compared the bit rates achieved by the error spectrum lying under the spectral domain masking curve of a predetermined psychoacoustic model), being careful to include in the library the best filters with different frequency selectivity, time resolution and smoothness properties. In the library were included also biorthogonal filter banks, very popular for image compression applications, which, besides producing non-phase- distorting filters, have the advantage that can be designed so that the characteristics of the analysis filter bank be different from synthesis filter bank.

Daubechies filters [2] showed the best performances between orthogonal filters, whereas Villasenor filters [16] were the best between biorthogonal filters. We included

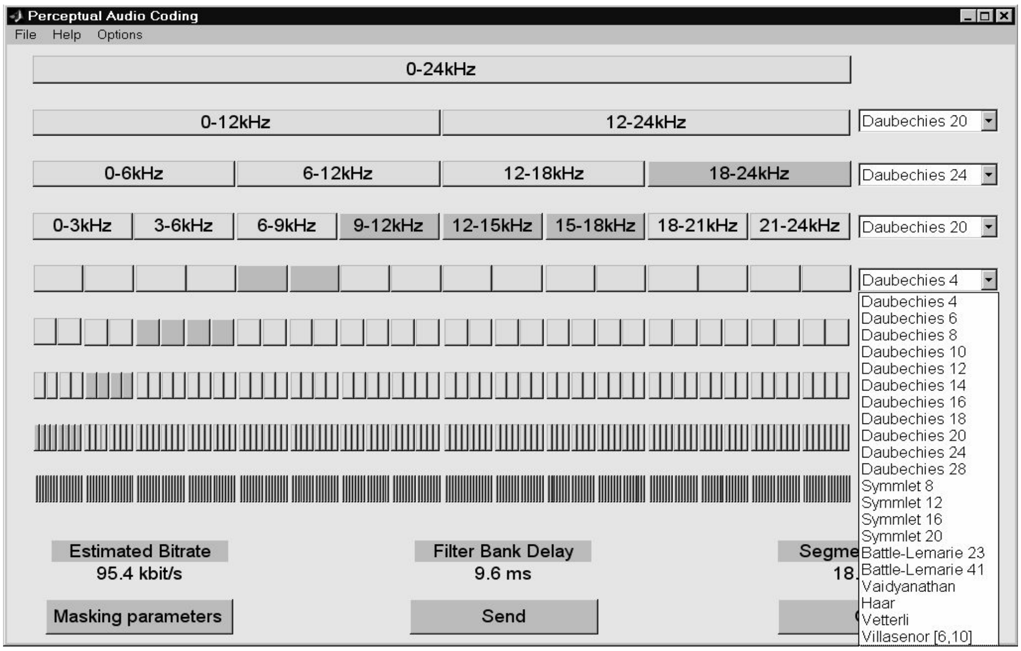

Fig. 1. Subband decomposition and filter bank selection

among the others in the library, the Vaidyanathan, Vetterli and Cohen-Daubechies-Faveau filters [1].

3 PSYCHOACOUSTIC MODEL OPTIMIZATION

The samples of the time-scale representation must be quantized to reduce the amount of data sent to the transmission channel. The allocation of the bits to the subbands considered the perceptual noise masking characteristics of the human ear.

The hybrid wavelet packet transform was used as frequency analyzer for the psychoacoustic model in place of the Fourier transform in order to avoid the redundant computation of frequency analysis on psychoacoustic model and filter bank [17].

The calculus of the masking thresholds involved several steps [18]:

a) First we used the wavelet packet transform with the filters selected at each level by the user to calculate the signal energy in each critical band (*discrete Bark spectrum*). The critical band boundaries and the relative Bark values were set as in ISO/MPEG psychoacoustic model II [19].

b) The masking effect is not band-limited to within the boundaries of a single critical band, but a masker centred within one critical band has some predictable effect on detection thresholds in other critical bands. The spread of masking was modelled by the rounded modified function [19, 20]:

$$S_{dB}(z, z_k) = a + \frac{l+r}{2}\left[1.05(z - z_k) + 0.47\right] - \frac{l-r}{2}\left[f + \left(1.05(z - z_k) + 0.47\right)^2\right]^{\frac{1}{2}} + x \quad \text{with}$$
$$x = \begin{cases} 8\big((z - z_k)b - 0.5\big)^2 - 2\big((z - z_k)b - 0.5\big) \\ \qquad \text{if } 0.5 < (z - z_k)b < 25\,, \\ 0 \quad \text{otherwise.} \end{cases} \tag{1}$$

In this expression, z_k is the bark value of the signal being spread, z is the bark value of the band being spread into, l and r, multiplied by b, are approximately the left slope and the right slope in decibels per Bark, f is the peak flatness and a is a compensation factor needed to satisfy the condition $S_{dB}(0) = 0$. We let the user adjust the value of the parameters r, l and f, which were initially set to 25 dB/Bark, -10 dB/Bark and 1, respectively, as in [19]. The spreading function (1), converted to a linear scale, was convolved with the discrete Bark spectrum to account for the spread of masking (*spread critical band spectrum*). Due to the non-normalized nature of the spreading function, the spread critical band spectrum was renormalized as in [19].

c) As tonal maskers and noise maskers generate different masking patterns, several approaches have been proposed to calculate a tonality index for the signal. The AAC Auditory Masking Model considers the predictability of individual frequency components across time, in terms of magnitude and phase tracking properties (*unpredictability measure*) [21]. Due to the limited frequency selectivity of the DWPT, the spectral flatness measure (SFM) [22] was used in place of the

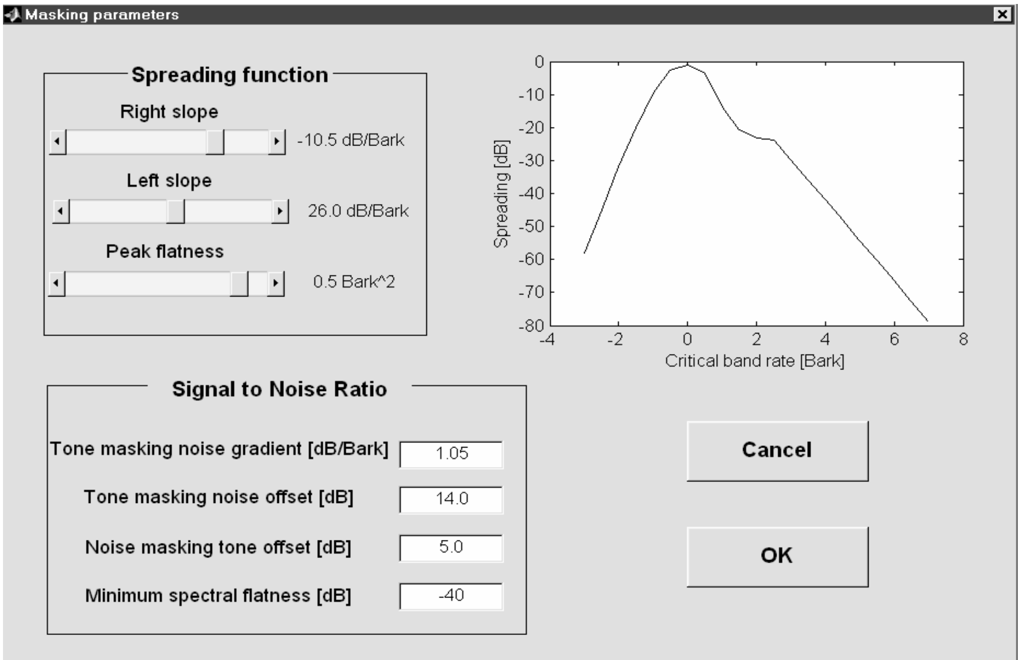

Fig. 2. Psychoacoustic model optimization

Table 1. A filter selection for transparent compression of 48 kHz audio samples at 96 kb/s with an overall filter bank delay less then 10 ms.

Level	Filter	Length
1	Daubechies	20
2	Daubechies	28
3	Daubechies	24
4	Villasenor	6, 10
5	Villasenor	6, 10
6	Daubechies	6
7	Haar	2

unpredictability measure. The SFM is defined by the ratio of the geometric mean to the arithmetic mean of the power spectral density components in each critical band. It has the property that it is bounded by zero and one. Values close to one will occur if the spectrum is *noiselike*, whereas values close to zero will occur if the there are dominant sinusoidal components in the spectrum. The SFM was converted in decibel and used to calculate a tonality coefficient α as follows:

$$\alpha = \min\left(\frac{SFM_{dB}}{SFM_{dB\,\min}}, 1\right) \tag{2}$$

where $SFM_{dB\,\min}$ is the SFM value in dB used to estimate that a signal is entirely *tonelike*, which was initially set to -60 dB.

d) The averaged value of the tonality index was used to weight geometrically the tone-masking-noise component and the noise- masking-tone component (*masking indices*) for each critical band I to calculate the signal to noise ratio in decibels [12, 21, 23]. The noise threshold estimate was then formed for each critical band I by dividing the value of the renormalized spread critical band spectrum by the signal to noise ratio (converted to a linear scale).

e) The amount of energy needed in a pure tone such that it can be detected by a listener in a noiseless environment is the *absolute threshold of hearing* (ATH). The maximum between the Bark threshold and absolute threshold of hearing was taken as final masking threshold, taking care to select the minimum value of AHT inside each critical band.

The masking thresholds were updated every 10.67 ms for adjacent frames of 512 samples. The computation was performed on overlapping analysis windows. The length of the analysis window was depending on the delay of the hybrid analysis filter bank selected by the user.

The GUI provide the user an interactive window which allows him to customize the spreading function parameters (right slope, left slope and peak flatness), the masking indices and the $SFM_{dB\,\min}$ value, while listening in real-time the effect of his changes (see Fig. 2).

4 REAL TIME IMPLEMENTATION

We implemented our tool on the TI TMS320C6711 DSK, equipped with the TMDX326040A Audio Daughter Card for sampling up to 48 kHz. The DSK was connected to a PC through its parallel interface.

The input signal was sampled and digitalized by means of the 16-bit A/D converter. In order to transfer continuously the digitalized voice from the A/D serial port to the CPU memory without loading the CPU, we performed this operation by programming the Enhanced Direct Memory Access controller (EDMA). Although the CPU and the EDMA controller function independently of one another, when both are performing simultaneous data accesses it is necessary to properly schedule and configure them in order to minimize conflict and waiting while meeting real-time requirements. To allow the CPU activity to be distanced from the EDMA activity we implemented a ping-pong buffering technique [24]. In ping-pong buffering there are two sets of data buffers for all incoming and outgoing data streams. While the EDMA is transferring data to the ping buffers, the CPU is compressing the data in the pong buffers. When both CPU and EDMA activities complete, they switch.

We used a Real Time Data Exchange (RTDX) 2-way data channel to transfer data between the host computer and the DSP target without interfering with the application. On the host computer a MATLAB script writes on the Host-to-Target channel the filter bank and masking parameters whenever the user updates them and periodically reads on the DSP-to-Host channel the current values of bitrate and overall filter bank delay. Conversely, the code running on the DSP periodically checks the Host-to-Target channel to update the codec with the new parameters set by the user and writes on the DSP-to-Host channel the current values of bitrate and delay.

The bitrate was calculated by summing the perceptual entropy in the wavelet domain [25] and the overhead for the coding of the masking thresholds.

5 MAIN RESULTS

We used a database of several 48 kHz audio samples (viola, clarinet, bass, harpsichord, castanets, piano, guitar, drums, *etc*).

First of all, we observed that the optimal masking model does not keep unchanged if we modify the filter bank configuration. In fact, a trade-off between the frequency selectivity and the time localization does not guarantee the transparency of the coding performed with the procedure described in section 3, which corresponds to the idea of adding the quantization noise independently in non-overlapping subbands.

In particular, if we use the wavelet packet transform as frequency analyzer to calculate the SFM, as in [17], a decrease of the analysis filter bank frequency selectivity leads to an overestimate of the SFM when the spectrum in a particular band is narrowband. We observed that this overestimate could be partially compensated by raising the $SFM_{dB\,\min}$ default value.

Furthermore, a decrease in frequency selectivity should be also compensated with a lowering of the peak flatness parameter f of the spreading function.

In order to preserve frequency selectivity as the number of stages of the DWPT increases, a commonly used approach is to use orthogonal Daubechies filters in each stage of the transform, reducing gradually the filter lengths through the cell structure (with Haar filters in the last stages of the transform). We found that also biorthogonal filters, such as the Cohen-Daubechies-Faveau and Villasenor filters, can be advantageously used in the middle stages of the transform (see table 1).

The best solutions were obtained by stop doing the low frequency subband decomposition at some level before the critical band subdivision one. This choice, in fact, besides to keep the overall delay acceptable, reduces advantageously the bitrate overhead for the transmission of side information, at expense of a partial exploitation of the psychoacoustic model. In particular the use of Daubechies, Villasenor and Haar filters (see table 1) guaranteed a transparent compression at 96 kbit/s and an overall filter bank delay (analysis and synthesis filters delay) less then 10 ms.

6 CONCLUSIONS AND FUTURE WORK

An interactive interface for performance measures of perceptual audio coding has been proposed and a real time implementation in a mixed MATLAB/TMS320C6000 environment has been tested.

The interface has permitted to demonstrate that some simplifying hypotheses, often used for the selection of the wavelet-packet best basis, are not always satisfied. In particular, when the wavelet packet transform is used as spectral analyzer, a "tuning" of the parameters of the psychoacoustic model for each filter bank configuration guarantees meaningfully better performances than the use of a preset masking model.

Even though the interface allowed the identification of a representation suitable for applications requiring low coding delay, such as real-time communication, it finds its more natural application for the validation and the optimization of a set of candidate solutions of a fully objective best basis search algorithm. Moreover, from the educational point of view, this application can support graduate students on wavelets, DSP programming and perceptual coding.

Future work includes the analysis of the performances of a time-varying filter bank, the incorporation of a temporal masking model, and a mean opinion score (MOS) validation procedure.

References

[1] MALLAT, S.: A Wavelet Tour of Signal Processing, Second Edition, Academic Press, 1998.

[2] DAUBECHIES, I. Orthonormal Bases of Compactly Supported Wavelets. II. Variation on a Theme: SIAM J. Math. Anal. **24** No. 2 (1993), 499–519.

[3] WICKERHAUSER, M. V.: INRIA Lectures on Wavelet Packet Algorithms, Problemes Non Lineaires Appliques, Ondellettes and Paquets d'Onde, P.L. Lions Ed., Roquencourt, France, 1991.

[4] PHILIPPE, P.—MOREAU De SAINT-MARTIN, F.—LEVER, M.: Wavelet Packet Filterbanks for Low Time Delay Audio Coding, IEEE Transactions on Speech and Audio Processing **7** No. 3 (1999), 310–322.

[5] SRINIVASAN, P.—JAMIESON, L. H.: High-Quality Audio Compression Using an Adaptive Wavelet Packet Decomposition and Psychoacoustic Modeling, IEEE Transactions on Signal Processing **46** No. 4 (1998), 1085–1093.

[6] JOHNSTON, J. D.: Estimation of Perceptual Entropy Using Noise Masking Criteria, Proceedings of ICASSP 1988, International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. 2524–2527, 11–14 April 1988.

[7] PASERO, E.—MONTUORI, A.: Neural Network Based Arithmetic Coding for Real-Time Audio Transmission on the TMS320C6000 DSP Platform, Proc. of Icassp 2003, International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 761–764, April 2003.

[8] QUAGLIA, D.—MONTUORI, A.—De MARTIN, J. C.—PASERO, E.: Interactive Dsp Educational Platform for Real-Time Subband Audio Coding, Proceedings of Icassp 2002, International Conference On Acoustics, Speech, And Signal Processing, vol. IV, pp. 4136–4139, May 2002.

[9] PASERO, E.—MONTUORI, A.: Wavelet Based Wideband Speech Coding on the TMS320C67 for Real Time Transmission, Proceedings of IEEE MTAC 2001, Multimedia Technology and Applications Conference, Irvine, pp. 208–212, 8–11 Nov. 2001.

[10] FU, X.—ZHANG, Z.: TMS320C6000 DSP Multichannel Vocoder Technology Demonstration Kit Host Side Design, Texas Instruments Appl. Rep. SPRA558B — February 2000.

[11] The MathWorks Inc., MATLAB Link for Code Composer Studio Development Tools, July 2002.

[12] SCHARF, B.: Critical Bands, in Foundations of Modern Auditory Theory, edited by J.V. Tobias, Chapter 5,, Academic Press, NY, 1970.

[13] HEDGES, R. A.: Hybrid Wavelet Packet Analysis, Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis, pp. 221224, 6–9 Oct. 1998.

[14] SINHA, D.—JOHNSTON, J. D.: Audio Compression at Low Bit Rates Using a Signal Adaptive Switched Filterbank, Proceedings of Icassp 1996, International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 1053–1056, 7–10 May 1996.

[15] ERNE, M.—MOSCHYTZ, G.—FALLER, C.: Best Wavelet-Packet Bases for Audio Coding Using Perceptual and Rate-Distortion Criteria, Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999, ICASSP '99, vol. 2, pp. 909–912, 15–19 March 1999.

[16] VILLASENOR, J. D.—BELZER, B.—LIAO, J.: Filter Evaluation and Selection in Wavelet Image Compression, IEEE Data Compression Conference, DCC '94. Proc., pp. 351–360, 1994.

[17] CARNERO, B.—DRYGAJLO, A.: Perceptual Speech Coding and Enhancement Using Frame-Synchronized Fast Wavelet Packet Transform Algorithms, IEEE Trans. on Signal Processing **47** No. 6 (1999).

[18] PAINTER, T.—SPANIAS, A.: Perceptual Coding of Digital Audio, Proceedings of the IEEE **88** No. 4 (2000), 451–515.

[19] International Standards Organization, Coding of Moving Picture and Associated Audio for Digital Storage Media at up to About 1.5 Mbit/s — Part 3: Audio, ISO/IEC 11172-3, 1993.

[20] DEUTSCH, W. A.—NOLL, A.—ECKEL, G.: The Perception of Audio Signals Reduced by Overmasking to the Most Prominent Spectral Amplitudes (Peaks), Proceedings of the 92nd Convention of the Audio Engineering Society, Vienna, pp. 129–132, March 1992.

[21] International Standards Organization, Generic Coding of Moving Pictures and Associated Audio Information (Part 7) — Advanced Audio Coding (AAC), ISO/IEC DIS 13818-7, 1996.

[22] JAYANT, N. S.—NOLL, P.: Digital Coding of Waveforms: Principles and Applications to Speech and Video, Englewood Cliffs, NJ: Prentice-Hall, 1984.

[23] HELLMAN, R.: Asymmetry of Masking between Noise and Tone, Percep. Psychphys. **11** (1972), 241–246.

[24] Texas Instruments Inc., TMS320C6000 Peripherals Reference, Lit. Nu. SPRU190D, Febr. 2001.

[25] PETROVSKY, A.: A Dynamic Algorithm Transforms for Reconfigurable Real-Time Audio Coding Processor, Proc. of International Conference on Parallel Computing in Electrical Engineering 2002, PARELEC '02, pp. 422–424, 22–25 Sept. 2002.

Received 16 July 2004

Eros Pasero (Assoc Prof) was born in Italy in 1955. He received his MS degree in electronic engineering from the Politecnico of Torino in 1980. From 1987 to 1990 he was Associate Professor of electronics with the II University of Roma. Since 1991 he has been again with the Politecnico di Torino where he is Associate Professor of electronics. He has been involved in some Italian and European projects (MADESS, Biotecnologie, Esprit Nervess, ANNIE, NEMEFO) dealing with microelectronics and neural networks. He was member of several scientific committees of international conferences (ISCIS, WIRN, FPGA, DSP, ESP, ECS), Vice General Chairman of the IJCNN2000 conference and editorial board of the Journal of Electrical Engineering. Now he is managing the Laboratorio di Neuronica, a laboratory for the study and development of neural networks.

Alfonso Montuori (PhD) was born in Italy in 1964. He received the *summa cum laude* degree in Physics from the University of Naples in 1993. He worked from 1994 to 1995 for CNR (National Research Council) Cybernetics Institute of Arco Felice, Naples, on multimedia applications and diagnostic expert systems. From 1997 to 1998 he co-operated in the design and implementation of an artificial neural network for human motion emulation in virtual environments at Politecnico di Torino, in the context of the European Community ANNIE project. Since 1999 he has worked on artificial neural networks and digital processing of audio signals for Electronics Department of Polytechnic of Turin, where he received the PhD degree in 2002. Since 2002 he has been Research Assistant at Electronics Department of Politecnico di Torino.