ARCHIVES OF ACOUSTICS
31, 2, 155–165 (2006)

COMPUTING OF MASKING THRESHOLDS FOR AUDIO CODERS BASED ON A QUATERNIONIC 4-BAND WAVELET PACKET TRANSFORM

M. PARFIENIUK, J. BASZUN, A. A. PETROVSKY

Białystok University of Technology
Wiejska 45A, 15-351 Białystok, Poland
e-mail: marekpk@ii.pb.bialystok.pl
jb@ii.pb.bialystok.pl, palex@ii.pb.bialystok.pl

(*received April 30, 2005; accepted October 25, 2005*)

A wavelet packet based on 4-band building blocks was used to implement an auditory model for 44.1 kHz and 16 kHz sampling frequency. The underlying paraunitary filter bank is implemented using a quaternionic lattice structurally insensitive to the quantization of its coefficients. Both the linear phase and orthogonality are possible for 4-band wavelets, so a better perceptual quality can be expected and an increased compression ratio for the coders based on the proposed solution in comparison to standard 2-band wavelet packets or a warped DFT transform. These features and a low computational complexity predestinate this approach to be a tempting alternative to widely known solutions.

Key words: wavelet packet transform, quaternion, masking threshold.

1. Introduction

In modern digital audio processing an important role is played by psychoacoustic used in signal coding and speech enhancement to keep the signal distortions unperceivable to the listener. Every psychoacoustically motivated system is based on a model mimicking the behavior of a part of the human auditory system, usually the inner ear, in analyzing sound in nonequal so called critical bands (CBs). The other part of the model is applied to determine the masking threshold, i.e. the power level that shows which signal components will be inaudible in the signal. This threshold controls the quantization in signal coding or weighting in noise removal.

The masking threshold depends on the signal power spreading in time and frequency, so the first and most important step in its calculation is the CB power analysis. This analysis with other computations (spreading, tonality estimation, normalization, absolute threshold comparison) finally leads to the perceptual threshold. There are a few competitive approaches in the CB power analysis. The first method was inspired by JOHNSTON [1] and is based on computing FFT of a windowed signal and then combines the coefficients into groups corresponding to CBs of hearing. The attainment of a reasonable spectral resolution in the narrowest bands requires the application of a rather long time window, therefore the method is limited because of a poor temporal resolution that is insufficient for dealing with such phenomena as pre-masking [2].

A class of solutions free of this drawback exploits a nonuniform filter bank to decompose the signal. The main shortcoming of such an approach is the general complexity, especially when a good approximation of CBs is of interest. Some promising alternative to nonuniform banks is a warped transform [3] with frequency samples allocated in accordance with the perceptual scale. Such solution constitutes the FFT, however at the cost of increased complexity.

Another approach exploits wavelet transforms. The common property of all the wavelet transforms applied to sounds is their adjustment to the characteristics of the human auditory system. Namely, the related subband decomposition approximates the critical bands of hearing [4]. Such an approach allows one to exploit optimally the two following nuisances of audio processing. Firstly, the statistical redundancies of the source are eliminated due to the decorrelating effect of the wavelet transform. On the other hand, the perceptual masking phenomena allow to treat certain signal components as irrelevant ones.

There are known different wavelet-based coding schemes for audio. At the beginnings, static time-invariant decomposition trees were used and the only considered adaptation has involved the wavelet coefficients and filter lengths [5]. During the recent years, several inherently different algorithms based on the dynamic reconfiguration of the wavelet tree from frame to frame have been used [6, 7], what gives time-frequency tiling close to the optima in the perceptual entropy sense. In both approaches, only two-band wavelets are used, although generalized M-band wavelets are also known [8].

The main concept is to approximate the critical bands of hearing with multi-band filter banks being simultaneously orthogonal and linear-phase called 4-band wavelet packet transforms [9]. This allows to use a symmetric extension of the process frames in isolation without the necessity of the overlap-add technique [10]. This, together with an improved coding gain, is supposed to give a significant bit-rate reduction in coders based on the wavelet transform presented. The implementation of this bank as a quaternionic 4-channel lattice has the unique advantage of structural orthogonality (losslessness) regardless of coefficient quantization. A similar result was obtained previously only for 2-channel lattice filter bank structures, which has undoubtedly influenced their popularity as superiority over the M-channel systems. This is so because it greatly simplifies the fixed-point hardware implementation of the wavelet packed tree based on such a filter bank structure.

2. Quaterionic based 4-band wavelet packet transform

M-band wavelets are a direct generalization of the dyadic ones and can be obtained by arranging filter banks into tree-structures as well. However in this case, there are, instead two, M filters with the impulse responses $h_i(l)$, $l = 0 \ldots L-1$. The scaling function is the solution of the following two-scale difference equation

$$\psi_0 = \sqrt{M} \sum_{l=0}^{L-1} h_0(l)\, \psi_0(Mt - l) \tag{1}$$

involving only the lowpass filter. Knowing the above scaling function, one can define the wavelets:

$$\psi_i = \sqrt{M} \sum_{l=0}^{L-1} h_i(l)\, \psi_0(Mt - l), \qquad i = 1, \ldots, M - 1. \tag{2}$$

The translates and dilates of these $M - 1$ wavelets can be used to represent an arbitrary square integrable function. In the most essential M-band wavelet transform, only the lowpass band serves as the input for the next level of the multiresolution analysis. Another alternative is the wavelet packet transform imposing no restrictions on the tree structure. Moreover, its adaptive version allows a dynamic modification of the wavelet tree to adjust the corresponding tiling of the time-frequency plane to the signal nonstationarities.

M-band wavelet transforms have several advantages over their dyadic counterpart. The first one is a better energy compaction (higher coding gain). Secondly, the 2-channel orthogonal paraunitary filter banks (PUFB) cannot have a linear phase. This restriction does not exist for $M > 2$.

The lattice factorization for the polyphase transfer matrix $\mathbf{E}(z)$ of a M-channel linear phase paraunitary filter bank was derived in [8], and has the following form

$$\begin{aligned} \mathbf{E}(z) &= \mathbf{G}_{N-1}(z)\mathbf{G}_{N-2}(z)\cdots\mathbf{G}_1(z)\mathbf{E}_0, \\ \mathbf{E}_0 &= \frac{1}{\sqrt{2}}\mathbf{\Phi}_0\mathbf{W}\,\mathrm{diag}\left\{\mathbf{I},\, \mathbf{J}\right\}, \qquad \mathbf{W} = \begin{bmatrix} \mathbf{I} & \mathbf{I} \\ \mathbf{I} & -\mathbf{I} \end{bmatrix}, \\ \mathbf{G}_i(z) &= \frac{1}{2}\mathbf{\Phi}_i\mathbf{W}\mathbf{\Lambda}(z)\mathbf{W}, \qquad \mathbf{\Lambda} = \mathrm{diag}\left\{\mathbf{I},\, z^{-1}\mathbf{I}\right\}. \end{aligned} \tag{3}$$

In the 4-channel case, it corresponds to the lattice shown in Fig. 1.

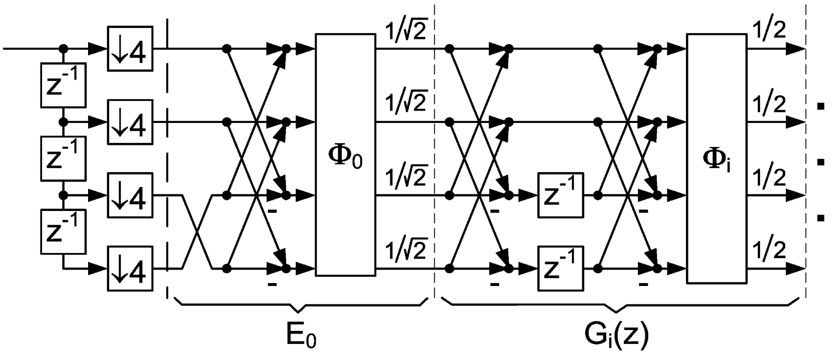

Fig. 1. Lattice structure for linear phase PUFB.

The authors have proposed a factorization alternative to that shown in [8] using the following two matrices

$$\mathbf{M}^{+}(Q) = \begin{bmatrix} q_1 & -q_2 & -q_3 & -q_4 \\ q_2 & q_1 & -q_4 & q_3 \\ q_3 & q_4 & q_1 & -q_2 \\ q_4 & -q_3 & q_2 & q_1 \end{bmatrix}, \tag{4}$$

$$\mathbf{M}^{-}(Q) = \begin{bmatrix} q_1 & -q_2 & -q_3 & -q_4 \\ q_2 & q_1 & q_4 & -q_3 \\ q_3 & -q_4 & q_1 & q_2 \\ q_4 & q_3 & -q_2 & q_1 \end{bmatrix} \tag{5}$$

related to the matrix-vector notation for non-commutative quaternion multiplication. The matrix (4) is written if Q is the left multiplicand, and (5) otherwise. The factorization (3) becomes quaternionic after assuming

$$\mathbf{\Phi}_i = \mathbf{M}^{-}(P_i), \qquad i = 1 \dots N-1 \tag{6}$$

in all its stages except $\mathbf{E}_0$ which must be represented as

$$\mathbf{\Phi}_0 = \mathbf{M}^{-}(P_0)\,\mathbf{M}^{+}(Q_0). \tag{7}$$

The proof can be found in [11]. As the synthesis filter bank is obtained simply by reversing the transformations in (3), the above results apply to it as well.

In our experiments, we have used the filter bank whose characteristics are shown in Fig. 2. Both its ideal and quantized lattice coefficients are shown in Table 1 (those being zero by definition are omitted). It was designed in order to achieve a maximal stopband attenuation. The factorization (3) has 7 $\mathbf{G}_i(z)$ stages, so the filters have a length of 32. The stopband achieved was 27 dB, and the coding gain for the AR(1) signal model was 8.44 dB. Such a filter bank can be the basis of the psychoacoustically motivated wavelet packet tree.

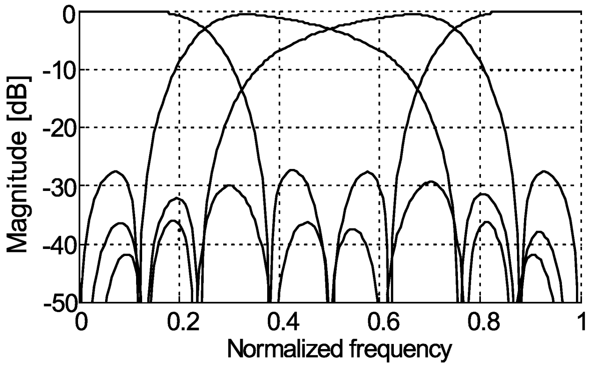

Fig. 2. Frequency response characteristics of the considered filter bank.

Table 1. Quaternionic lattice coefficients.

Coefficient	Precise value	Quantized value	CSD expansion
$Q_{0,1}$	-0.7062053	-0.7500000	$-2^{-0}+2^{-2}$
$Q_{0,2}$	0.7080071	0.7500000	$+2^{-0}-2^{-2}$
$P_{0,1}$	0.1604117	0.1562500	$+2^{-3}+2^{-5}$
$P_{0,2}$	0.9870502	0.9843750	$+2^{-0}-2^{-6}$
$P_{1,1}$	-0.7985673	-0.7500000	$-2^{-0}+2^{-2}$
$P_{1,2}$	-0.6019055	-0.6250000	$-2^{-1}-2^{-3}$
$P_{2,1}$	-0.8506589	-0.8750000	$-2^{-0}+2^{-3}$
$P_{2,2}$	0.5257179	0.5312500	$+2^{-1}+2^{-5}$
$P_{3,1}$	0.0441147	0.0468750	$+2^{-4}-2^{-6}$
$P_{3,2}$	-0.9990265	-1.0000000	-2^{-0}
$P_{4,1}$	-0.9388232	-0.9375000	$-2^{-0}+2^{-4}$
$P_{4,2}$	-0.3443996	-0.3750000	$-2^{-1}+2^{-3}$
$P_{5,1}$	0.8957751	0.8750000	$+2^{-0}-2^{-3}$
$P_{5,2}$	0.4445076	0.4375000	$+2^{-1}-2^{-4}$
$P_{6,1}$	-0.1761618	-0.1875000	$-2^{-2}+2^{-4}$
$P_{6,2}$	0.9843612	0.9843750	$+2^{-0}-2^{-6}$
$P_{7,1}$	0.8211270	0.8750000	$+2^{-0}-2^{-3}$
$P_{7,2}$	0.5707455	0.5625000	$+2^{-1}+2^{-4}$

3. Psychoacoustic model implementation

The wavelet packet tree for the 4-band wavelet decomposition of an acoustical band for sampling frequency 44.1 kHz and 16 kHz is shown in Fig. 3. Table 2 shows the

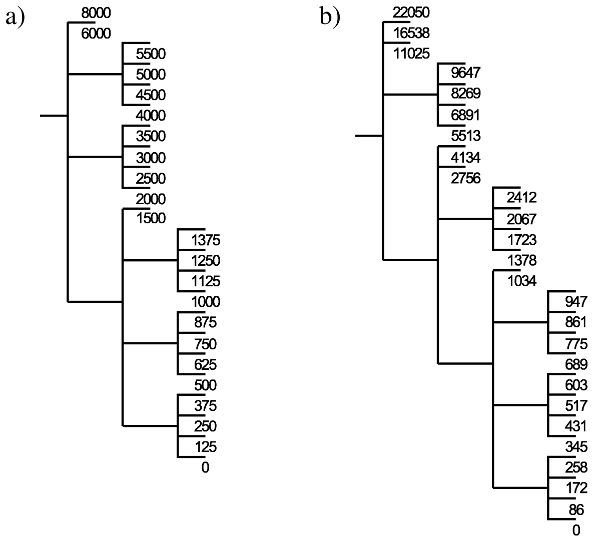

Fig. 3. Wavelet packet tree for a) 16 kHz and b) 44.1 kHz sampling. Frequencies are in Hz.

list of critical band centers and edge frequencies corresponding to this decomposition. The time frequency tiling map of this decompositions is shown in Fig. 4. The analyzed signal is divided in 512 point frames in both cases which gives two samples per frame for the lowest bands in the case of the 44.1 kHz sampling rate. This decomposition trees are the base for the psychoacoustic Bark scale and the psychoacoustical implementation model.

Table 2. Critical band centers and edge frequencies.

Bands	Subbands parameters [Hz]					
z	Lower		Center		Higher	
	16 kHz	44 kHz	16 kHz	44 kHz	16 kHz	44 kHz
1	0	0	63	43	125	86
2	125	86	188	129	250	172
3	250	172	313	215	375	258
4	375	258	438	301	500	345
5	500	345	563	388	625	431
6	625	431	688	474	750	517
7	750	517	813	560	875	603
8	875	603	938	646	1000	689
9	1000	689	1063	732	1125	775
10	1125	775	1188	818	1250	861
11	1250	861	1313	904	1375	947
12	1375	947	1438	991	1500	1034
13	1500	1034	1750	1206	2000	1378
14	2000	1378	2250	1550	2500	1723
15	2500	1723	2750	1895	3000	2067
16	3000	2067	3250	2239	3500	2412
17	3500	2412	3750	2584	4000	2756
18	4000	2756	4250	3445	4500	4134
19	4500	4134	4750	4823	5000	5513
20	5000	5513	5250	6202	5500	6891
21	5500	6891	5750	7580	6000	8269
22	6000	8269	7000	8958	8000	9647
23		9647		10336		11025
24		11025		13781		16538
25		16538		19294		22050

An auditory model is a procedure that tries to imitate the human hearing mechanism. It uses the knowledge of such areas as biophysics and psychoacoustics. The most important phenomena that occur in the hearing process for this model is a simultaneous frequency masking. The auditory model processes the audio information to get information on the final masking threshold. This information is used to compute the bit rate for signal coding.

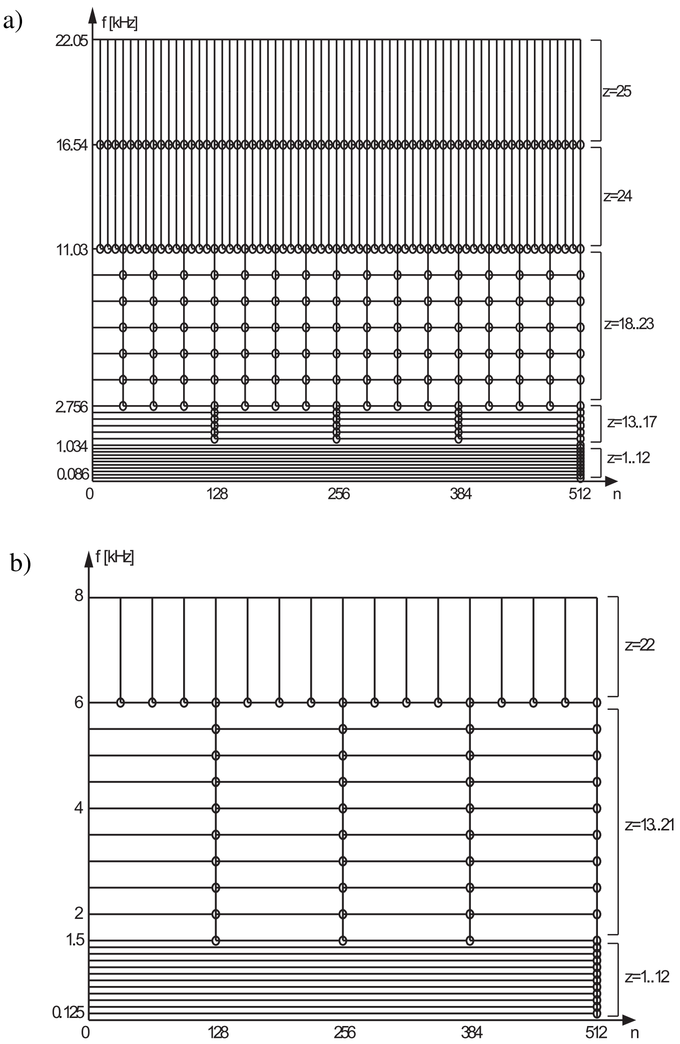

Fig. 4. Time-frequency tiling of the transform for: a) 44.1 and b) 16 kHz sampling.

Many studies have shown the nonuniform temporal and spectral resolutions of the human ear. Frequency components of sound are divided into critical bands whose centers and bandwidths have been measured. The center frequency locations of these subbands, i.e. the critical band rate can be approximated by the expression

$$z(f) = 13 \cdot \arctan(0.00076 \cdot f) + 3.5 \cdot \arctan\left((f/7500)^2\right), \tag{8}$$

where f is the frequency in hertz [4]. This relation is shown in Fig. 5 for 44.1 and 16 kHz sampling rates, respectively. The critical bandwidth can be expressed by

$$\text{CBW}(f) = 25 + 75 \cdot \left(1 + 1.4 \cdot (f/1000)^2\right)^{0.69} \ \text{[Hz]}, \tag{9}$$

where f is the center frequency in hertz [4]. This dependence is plotted in Fig. 6 for 44.1 and 16 kHz sampling rate, respectively, and compared to the other 2-band solutions in [12] and [6].

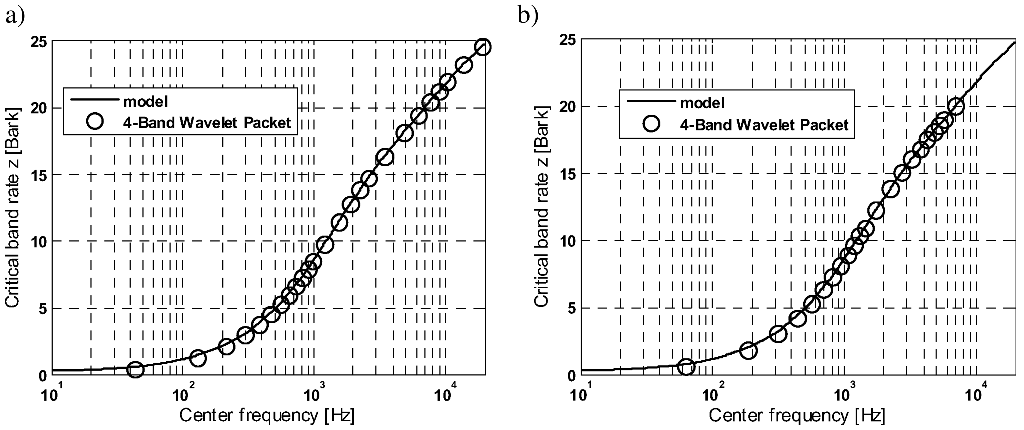

Fig. 5. Critical band rate as a function of the center frequency for: a) 44.1 and b) 16 kHz sampling.

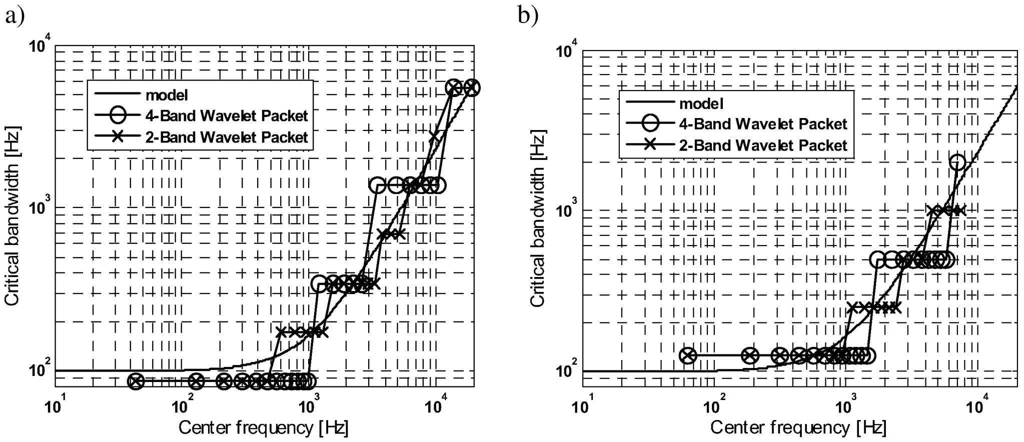

Fig. 6. Critical bandwidth as a function of the center frequency for 44.1 and 16 kHz sampling: 4- (circles) vs. 2-band (x-marks) solutions [6, 12].

4. Masking threshold calculation

The absolute threshold of hearing (ATH) is the average sound pressure level below which the human ear does not detect any stimulus, and is frequency dependent. ATH can be approximated by the expression

$$\mathrm{ATH_{SPL}}(f) = 3.64 \cdot \left(\frac{f}{1000}\right)^{-0.8} - 6.5 \cdot e^{-0.6\cdot\left(\frac{f}{1000}-3.3\right)^2} + 10^{-3} \cdot \left(\frac{f}{1000}\right)^4, \tag{10}$$

where f is the frequency in hertz [4].

The final masking threshold in each subband z is calculated on the basis of the components (wavelet coefficients) found in this subband [4]. As a matter of fact, all the calculations are based on the energy spectrum calculated in each subband according to the equation:

$$A(z) = 10 \log_{10} \left(\frac{1}{N} \sum_{i=i_a}^{i_b} \frac{w_i^2}{10} \right), \tag{11}$$

where i_a and i_b are the coefficient indices of the first and the last transform coefficients w_i within a given critical band z as listed in Table 2, N is the number of coefficients. The absolute thresholds are set such that the signal of 4 kHz with the peak magnitude of 10 is

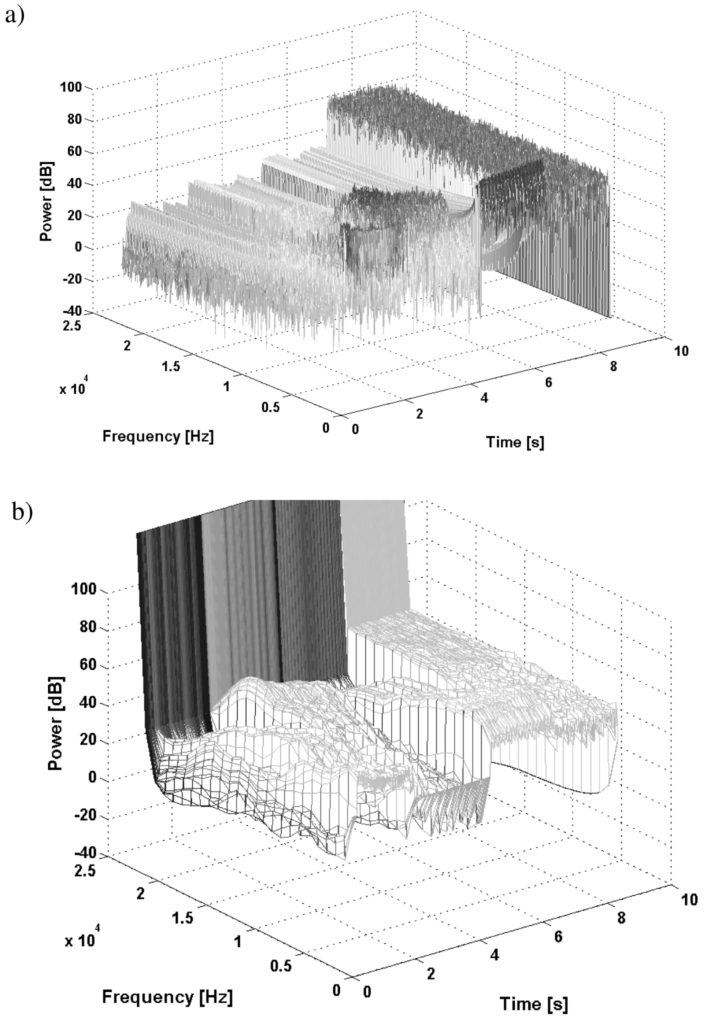

Fig. 7. Power spectrum calculated by FFT (a) and perceptual threshold calculated for speech tone and noise signal (b), 44.1 kHz sampling using a wavelet packet.

at the absolute threshold of hearing. Any subband that has a noise threshold lower than the absolute threshold is changed to the absolute threshold for the corresponding critical band. Because the absolute threshold varies inside the critical band, the mean of the critical band edges is used. Figure 7 shows the power spectrum and the final threshold for absolute threshold conditions computed for a test signal consisting of speech tone and noise.

5. Conclusion

The implementation of psychoacoustics principles which are basic from the point of view of compression and noise reduction systems is shown. The 4-band wavelet packet transforms developed allow to reconcile both linear phase and orthogonality. In this way a better compression ratio and energy compaction can be achieved. This system can perform a lossless compression or perceptual compression of audio signals.

Acknowledgments

This work was supported by the Białystok Technical University under the grant W/WI/2/04.

The paper was presented at the Conference "Structures – Waves – Human Health 2005" hold in Zakopane.

References

[1] JOHNSTON J. D., *Transform coding of audio signals using perceptuals noise criteria*, IEEE Trans. on Select. Areas Commun., **6**, 314–323 (1988).

[2] PAINTER T., SPANIAS A., *Perceptual coding of digital audio*, Proceedings of the IEEE., **88**, 4, 451–513 (2000).

[3] PARFIENIUK M., PETROVSKY A. A., *Warped DFT as the basis for psychoacoustic model*, ICASSP 2004m Montreal, Canada, 17–21 May 2004, on CD.

[4] ZWICKER E., TERHARDT E., *Analytical expressions for critical-band rate and bandwidth as a function of frequency*, JASA, **68**, 1523–1525 (1980).

[5] SINHA D., TEWFIK A., *Low bit-rate transparent audio compression using adapted wavelets*, IEEE Trans. Signal Processing, **41**, 3463–3479 (1993).

[6] PETROVSKY A., KRAHE D., PETROVSKY A. A., *Real-time wavelet packet-based low bit rate audio coding on a dynamic reconfiguration system*, Proc. 114-th AES Convention, Amsterdam, The Netherlands, p. 22, 22–25, March 2003.

[7] ZURERA ROSA M., REYES RUIZ N., CANDEAS VERA P., FERRERAS LÓPEZ F., *Use of the symmetrical extension for improving a time-varying wavelet-packet-based audio coder*, Digital Signal Processing, **13**, 457–469 (2003).

[8] SOMAN A. K., VAIDYANATHAN P. P., NGUYEN T. Q., *Linear phase paraunitary filter banks: theory, factorization and designs*, IEEE Trans. Signal Processing, **41**, 3480–3496 (1996).

[9] PARFIENIUK M., BASZUN J., PETROVSKY A. A., *FPGA friendly quaternionic 4-band wavelet packet transforms for sound processing*, MIXDES 2005, 22–25 June, Kraków 2005.

[10] STRANG G., NGUYEN T., *Wavelets and filter banks*, Wellesley-Cambridge Press, Wellesley, MA, 1996.

[11] PARFIENIUK M., PETROVSKY A., *Linear phase paraunitary filter banks based on quaternionic component*, Proc. Int. Conf. on Signals and Electronic Systems ICSES'04, Poznań, Poland, 13–15 September, pp. 203–206, 2004.

[12] CARNERO B., DRYGAJLO A., *Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms*, IEEE Trans. Signal Processing, **47**, 1622–1635 (1999).