STRUCTURALLY ORTHOGONAL FINITE PRECISION IMPLEMENTATION OF THE EIGHT POINT DCT

Marek Parfieniuk and Alexander Petrovsky

Bialystok Technical University
Faculty of Computer Science
ul. Wiejska 45A, 15-351 Bialystok, Poland
Email: marekpk@ii.pb.bialystok.pl, palex@it.org.by

ABSTRACT

This paper presents a novel approach to the finite precision implementation of the eight point discrete cosine transform (DCT). Two multiplierless computational schemes of the plane rotation block are constructed to obtain the transform approximations maintaining orthogonality regardless of their coefficient quantization. This is the main difference with respect to the solutions based on lifting schemes being developed in recent years, characterized by inherent biorthogonality. In our technique, structural orthogonality comes at the cost of a complexity increase. To keep it at a moderate level, we implement rotations effectively using the denormalized lattice and the three-value coordinate rotation digital computer (CORDIC) algorithm with double μ-rotations.

1. INTRODUCTION

In recent years, multiplierless approximations of the discrete cosine transform (DCT) of type II attract a considerable attention. This variant of the DCT is of great significance for natural image processing, as it possesses decorrelation abilities close to those of the optimal Karhunen-Loève transform. Although many DCT algorithms have appeared since its invention, there is still the need to look for new design compromises better suited to particular functional expectations and implementational limitations. For example, in mobile devices such as digital cameras or cell phones, hardware complexity as well as power consumption has to be minimized. So, even fixed-point arithmetic in the full extent is treated as a luxury, to say nothing of the floating-point one. The use of binary shifts and additions is only allowed.

The binDCT seems to be the most notable result in this field [1]. This is the name of the approach derived from known lattice factorizations for the DCT by replacing plane rotations with lifting schemes [2]. It offers good performance and design flexibility at extreme simplicity. However, the transforms obtained this way are no longer orthogonal as biorthogonality is an inherent property of ladder structures. This is not a trouble in most applications, but sometimes orthogonal DCT approximations with clear relation between errors in the signal and transform domains may be preferred.

In this paper, we propose a simple modification of the known Loeffler's factorization of the DCT matrix, making it structurally orthogonal regardless of coefficient quantization. The related computational scheme still consists of plane rotations whose number increases, however. Fortunately, multiplierless implementations of the resulting lattice are possible, leading to only a moderate efficiency loss with respect to the binDCT. The first considered approach is based on the lattice denormalization, whereas the second utilizes the coordinate rotation digital computer (CORDIC) algorithm.

Notations: Matrices are denoted by upper-case bold-faced characters. $\mathbf{I}_M$ denotes the $M \times M$ identity matrix.

This work was supported by the Polish Ministry of Education and Science in years 2005–2006 (grant No. 3 T11F 014 29).

2. STRUCTURALLY LOSSLESS LATTICE FOR DCT

Let us look at the Loeffler's factorization of the eight point DCT shown in Fig. 1. The structure is mainly composed from multiplierless butterflies. The only nontrivial operations are the three Givens rotations indicated with the dashed lines. This stage of the schema corresponds to a block diagonal matrix composed from one identity and three rotation matrices. The quantization of the rotation coefficients makes its column norms unequal and hence violates the perfect reconstruction property and the losslessness (orthogonality) of the transform involving finite precision rotations [3].

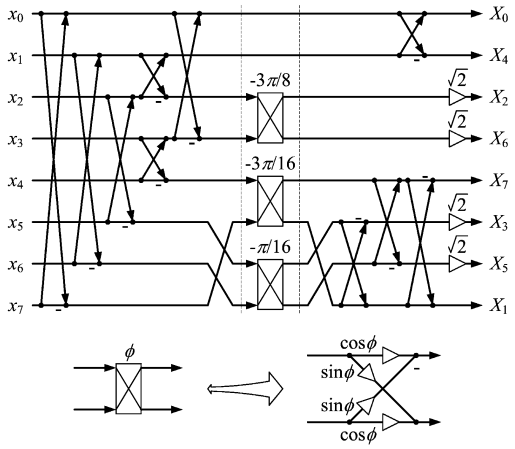

Fig. 1. Signal flow graph of Loeffler's factorization [2] of the eight point DCT.

The solution is to derive an alternative structure for the rotation stage, with components maintaining orthogonality regardless of their quantization. Knowing the elementary identity $\mathbf{R}(\phi)\mathbf{R}(\psi) = \mathbf{R}(\phi + \psi)$ for the rotation matrix

$$\mathbf{R}(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \quad (1)$$

Table 1. The angles used in the discussed factorizations

Angle	Value	$\sin(\cdot)$	$\cos(\cdot)$	$\tan(\cdot)$	$\cot(\cdot)$
α	$-3/16\pi$	-0.5556	0.8315	-0.6682	-1.4966
β	$-3/32\pi$	-0.2903	0.9569	-0.3033	-3.2966
γ	$-1/32\pi$	-0.0980	0.9952	-0.0985	-10.1532
$\Sigma_{\alpha\beta\gamma}$	$-5/16\pi$	-0.8315	0.5556	-1.4966	-0.6682
$\Sigma_{\alpha\beta}$	$-9/32\pi$	-0.7730	0.6344	-1.2185	-0.8207
$\Sigma_{\alpha\gamma}$	$-7/32\pi$	-0.6344	0.7730	-0.8207	-1.2185
$\Sigma_{\beta\gamma}$	$-1/8\pi$	-0.3827	0.9239	-0.4142	-2.4142

we can factorize the rotation stage in the following way

$$\begin{aligned}
&\operatorname{diag}\{\mathbf{I}_2, \mathbf{R}(2\alpha), \mathbf{R}(2\beta), \mathbf{R}(2\gamma)\} = \\
&\quad \operatorname{diag}\{\mathbf{R}(-\alpha), \mathbf{R}(\alpha), \mathbf{R}(-\alpha), \mathbf{R}(-\alpha)\} \\
&\quad \operatorname{diag}\{\mathbf{R}(\alpha), \mathbf{R}(\alpha), \mathbf{R}(\alpha), \mathbf{R}(\alpha)\} \\
&\quad \operatorname{diag}\{\mathbf{R}(\beta), \mathbf{R}(\beta), \mathbf{R}(\beta), \mathbf{R}(\beta)\} \\
&\quad \operatorname{diag}\{\mathbf{R}(-\beta), \mathbf{R}(-\beta), \mathbf{R}(\beta), \mathbf{R}(-\beta)\} \\
&\quad \operatorname{diag}\{\mathbf{R}(\gamma), \mathbf{R}(\gamma), \mathbf{R}(\gamma), \mathbf{R}(\gamma)\} \\
&\quad \operatorname{diag}\{\mathbf{R}(-\gamma), \mathbf{R}(-\gamma), \mathbf{R}(-\gamma), \mathbf{R}(\gamma)\}
\end{aligned} \tag{2}$$

where α, β and γ are as in Table 1.

Each of the block diagonal matrices at the right side of (2) preserves orthogonality and constant column norm regardless of its representation precision, so its cascade is also orthogonal. However, the number of the rotations in this simple solution is huge, as it equals 24. Fortunately, because the order of a rotation product is insignificant, we can easily derive the four subsequent factorizations:

$$\begin{aligned}
&\operatorname{diag}\{\mathbf{I}_2, \mathbf{R}(2\alpha), \mathbf{R}(2\beta), \mathbf{R}(2\gamma)\} = \\
&\quad \operatorname{diag}\{\mathbf{R}(-\alpha), \mathbf{R}(\alpha), \mathbf{R}(-\alpha), \mathbf{R}(-\alpha)\} \\
&\quad \operatorname{diag}\{\mathbf{R}(\Sigma_{\alpha\beta\gamma}), \mathbf{R}(\Sigma_{\alpha\beta\gamma}), \mathbf{R}(\Sigma_{\alpha\beta\gamma}), \mathbf{R}(\Sigma_{\alpha\beta\gamma})\} \\
&\quad \operatorname{diag}\{\mathbf{R}(-\beta), \mathbf{R}(-\beta), \mathbf{R}(\beta), \mathbf{R}(-\beta)\} \\
&\quad \operatorname{diag}\{\mathbf{R}(-\gamma), \mathbf{R}(-\gamma), \mathbf{R}(-\gamma), \mathbf{R}(\gamma)\}
\end{aligned} \tag{3}$$

$$\begin{aligned}
&\operatorname{diag}\{\mathbf{I}_2, \mathbf{R}(2\alpha), \mathbf{R}(2\beta), \mathbf{R}(2\gamma)\} = \\
&\quad \operatorname{diag}\{\mathbf{R}(\beta), \mathbf{R}(-\beta), \mathbf{R}(\beta), \mathbf{R}(\beta)\} \\
&\quad \operatorname{diag}\{\mathbf{R}(-\Sigma_{\alpha\beta}), \mathbf{R}(\Sigma_{\alpha\beta}), \mathbf{R}(\Sigma_{\alpha\beta}), \mathbf{R}(-\Sigma_{\alpha\beta})\} \\
&\quad \operatorname{diag}\{\mathbf{R}(\Sigma_{\alpha\gamma}), \mathbf{R}(\Sigma_{\alpha\gamma}), \mathbf{R}(-\Sigma_{\alpha\gamma}), \mathbf{R}(\Sigma_{\alpha\gamma})\}) \\
&\quad \operatorname{diag}\{\mathbf{R}(-\gamma), \mathbf{R}(-\gamma), \mathbf{R}(\gamma), \mathbf{R}(\gamma)\}
\end{aligned} \tag{4}$$

$$\begin{aligned}
&\operatorname{diag}\{\mathbf{I}_2, \mathbf{R}(2\alpha), \mathbf{R}(2\beta), \mathbf{R}(2\gamma)\} = \\
&\quad \operatorname{diag}\{\mathbf{R}(\alpha), \mathbf{R}(\alpha), \mathbf{R}(-\alpha), \mathbf{R}(\alpha)\} \\
&\quad \operatorname{diag}\{\mathbf{R}(-\Sigma_{\alpha\beta}), \mathbf{R}(\Sigma_{\alpha\beta}), \mathbf{R}(\Sigma_{\alpha\beta}), \mathbf{R}(-\Sigma_{\alpha\beta})\} \\
&\quad \operatorname{diag}\{\mathbf{R}(\Sigma_{\beta\gamma}), \mathbf{R}(-\Sigma_{\beta\gamma}), \mathbf{R}(\Sigma_{\beta\gamma}), \mathbf{R}(\Sigma_{\beta\gamma})\} \\
&\quad \operatorname{diag}\{\mathbf{R}(-\gamma), \mathbf{R}(\gamma), \mathbf{R}(-\gamma), \mathbf{R}(\gamma)\}
\end{aligned} \tag{5}$$

$$\begin{aligned}
&\operatorname{diag}\{\mathbf{I}_2, \mathbf{R}(2\alpha), \mathbf{R}(2\beta), \mathbf{R}(2\gamma)\} = \\
&\quad \operatorname{diag}\{\mathbf{R}(\alpha), \mathbf{R}(\alpha), \mathbf{R}(\alpha), \mathbf{R}(-\alpha)\} \\
&\quad \operatorname{diag}\{\mathbf{R}(-\Sigma_{\alpha\gamma}), \mathbf{R}(\Sigma_{\alpha\gamma}), \mathbf{R}(-\Sigma_{\alpha\gamma}), \mathbf{R}(\Sigma_{\alpha\gamma})\} \\
&\quad \operatorname{diag}\{\mathbf{R}(\Sigma_{\beta\gamma}), \mathbf{R}(-\Sigma_{\beta\gamma}), \mathbf{R}(\Sigma_{\beta\gamma}), \mathbf{R}(\Sigma_{\beta\gamma})\} \\
&\quad \operatorname{diag}\{\mathbf{R}(-\beta), \mathbf{R}(\beta), \mathbf{R}(\beta), \mathbf{R}(-\beta)\}
\end{aligned} \tag{6}$$

based on the angles being the sums of α, β, and γ, shown in Table 1 too. These constructions require only 16 rotations each but still possess structural losslessness. They seem to be tightly connected to the factorizations of 8×8 block diagonal orthogonal matrices reported in [4, 5], based on quaternionic approach. However, for the sake of brevity, we omit this relation in our discussion. Much more interesting for us is how to implement particular rotations and their sequences in an efficient manner, using only shifts and additions.

3. "DENORMALIZED" IMPLEMENTATION

3.1. Idea

This classical approach to efficient lattice implementations was proposed in [6]. It was termed "denormalized" as it consists in the extraction of the factors from the branches of the rotation butterfly, explained in Fig. 2. The selection of its particular variant is aimed at the reduction of the dynamic range of the coefficient remaining inside the lattice. Given a rotation sequence, we can merge all extracted factors into a single one (referred to as ζ in the further discussion) placed on the output. Thus, the number of the required multiplications decreases about two times.

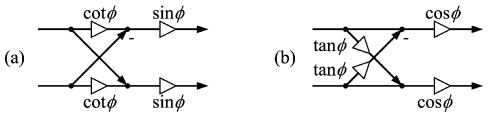

Fig. 2. Denormalized lattices for Givens rotation.

The structure for the rotation stage obtained this way starting with (4) is shown in Fig. 3. $\zeta = \cos\beta \sin\Sigma_{\alpha\beta} \cos\Sigma_{\alpha\gamma} \cos\gamma$ in this case, and it can be quantized without orthogonality violation.

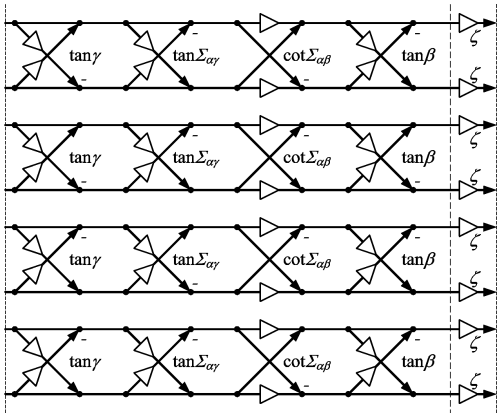

Fig. 3. Structurally lossless rotation stage.

3.2. Design example

Table 2 contains exemplary coefficients for the lattice in Fig. 3. They are represented using the "canonic-signed-digit" (CSD) code to facilitate an efficient implementation. Two DCT approximations differing in $\tan\beta$ are considered. For its value $-9/32$, the complete rotation stage requires 40 shifts and 56 additions. The resulting transform is characterized by the coding gain 8.8037 dB (calculated for the AR(1) input model; 8.8259 dB for the original DCT), and its magnitude responses are shown in Fig. 4. It is evident that the DC leakage to 5th band is nonzero though the attenuation 50 dB seems to be sufficient to neglect its effect. The reduction of the leakage depends on the approximation of the identity

$$\mathbf{R}(\beta)\mathbf{R}(-\Sigma_{\alpha\beta})\mathbf{R}(\Sigma_{\alpha\gamma})\mathbf{R}(-\gamma) = c\mathbf{I}_2 \qquad c = \text{const.} \tag{7}$$

Table 2. The multiplierless lattice coefficients

Function	Rational approx.	CSD expansion	Number of shifts	Number of adds
$\tan\beta$	$-9/32$	$-2^{-2}-2^{-5}$	2	2
$\tan\beta$	$-71/256$	$-2^{-2}-2^{-5}+2^{-8}$	3	3
$\cot\Sigma_{\alpha\beta}$	$-7/8$	$-1+2^{-3}$	1	2
$\tan\Sigma_{\alpha\gamma}$	$-3/4$	$-1+2^{-2}$	1	2
$\tan\gamma$	$-1/16$	-2^{-4}	1	1

At more accurate $\tan\beta \approx -71/256$, with eights operations more and significantly enlarged wordlength, the attenuation at the DC frequency in the 5th band response is 70 dB and the transform coding gain increases to 8.8057 dB. It should be noted that the normalization factor ζ has the value 0.5774 being unexpectedly very close to $1/\sqrt{3}$. It is normally coalesced with that from the inverse transform to obtain a rational number, and incorporated into the quantization process what leads to the so-called scaled DCT [7, 2].

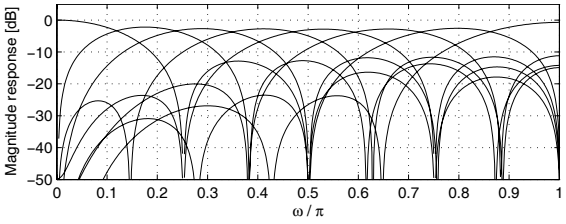

Fig. 4. Magnitude response of DCT approximation based on coefficients in Table 2 ($\tan\beta \approx -9/32$).

4. CORDIC IMPLEMENTATION

4.1. CORDIC algorithm

The CORDIC is a robust algorithm for computations based on orthogonal rotations [8, 7, 9]. Its essential variant consists in the decomposition of a given rotation in terms of the elementary rotations

$$\mathbf{R}(\alpha) \approx \prod_{k=0}^{W} \frac{1}{K_k} \begin{bmatrix} 1 & -\sigma_k 2^{-k} \\ \sigma_k 2^{-k} & 1 \end{bmatrix} \tag{8}$$

for the wordlength W, with the scale factor

$$K_k = \sqrt{1+\sigma_k^2 2^{-2k}} \tag{9}$$

and $\sigma_k^2 \in \{\pm 1\}$. This scheme is easy to implement in hardware due to its simplicity and regular layout. However, there are two disadvantages. Firstly, sometimes it is more efficient and accurate to omit certain rotations, e.g. to allow $\sigma_k^2 \in \{-1, 0, 1\}$ [8]. Secondly, an irrational scale factor makes the normalization difficult under finite precision. To cope with this problem, the CORDIC based on double μ-rotations

$$\mathbf{R}(\alpha) \approx \prod_{k=0}^{W} \frac{1}{K_k^2} \begin{bmatrix} 1 & -\sigma_k 2^{-k} \\ \sigma_k 2^{-k} & 1 \end{bmatrix}^2 \tag{10}$$

has been cosidered in [9], together with the efficient scaling procedure

$$\frac{1}{K_k^2} = (1-2^{-2k}) \prod_{s=1}^{\log_2\lceil W/2k\rceil} (1+2^{-2^{s+1}k}) \tag{11}$$

Table 3. The nonzero CORDIC parameters

Angle	σ_0	σ_1	σ_2	σ_3	σ_4	σ_5
α		-1		-1		
β			-1		-1	
γ					-1	-1

based on shifts and additions.

We aim to obtain CORDIC implementation of the DCT, exploiting the above mentioned improvements. The factorization (2) of the rotation stage allows the direct utilization of double μ-rotations.

4.2. Design example

Assuming the approximation of the rotations summarized in Table 3, we have the structure depicted in Fig. 5. As each pair of μ-rotations requires different scaling and a certain error in (11) is unavoidable, we decided to equalize the norm in each path from the input to the output, to have only one scaling factor $\zeta = 1/(K_1^2K_2^2K_3^2K_4^2K_5^2)$. The error related to such a scaling affects whole structure uniformly and hence does not touch its orthogonality.

The norm equalization is performed by placing the multipliers by K_k^2 in parallel to the corresponding rotation combination. In fact such a multiplier is implemented with only one shift and addition. Thus, the entire considered rotation stage requires 52 these operations. ζ can be realized using (11) — some factors in the product of such expressions can be combined to simplify the circuit.

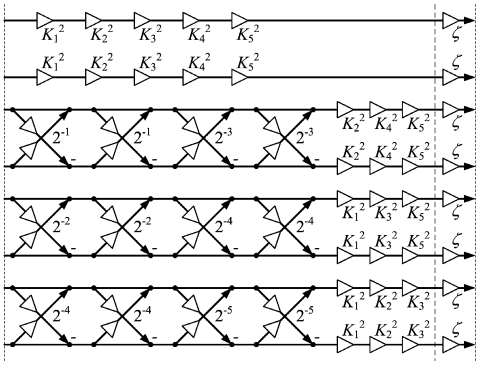

Fig. 5. The CORDIC-based rotation stage in Fig. 1.

The coding gain of such an approximation of the DCT reaches 8.8217 dB. The DC leakage is eliminated structurally. The corresponding frequency responses are shown in Fig. 6.

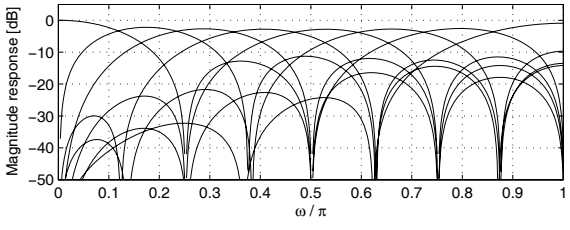

Fig. 6. Magnitude response of DCT approximation based on the CORDIC algorithm.

It should be noted, that the CORDIC-based approach to the DCT implementation is not a new idea [7]. However, neither utilization of double μ-rotations nor orthogonality seems to be considered.

5. FIXED POINT SIMULATIONS

To obtain more insights into the properties of the considered DCT algorithms, we modeled them using MATLAB/Simulink to simulate finite precision arithmetic. Then we evaluated the loss of the selectivity and the orthogonality of the corresponding filter banks. The channel magnitude response is calculated as the square root of the ratio between the power spectra of the subband and the input. In turn, the power spectra are estimated by averaging the periodograms of nonoverlapping and windowed random signal segments (Welch method [10]). The channel magnitude responses obtained this way for 8-bit wordlength are shown in Fig. 7. The sums of their squares are depicted in Fig. 8. For an orthogonal system, such a sum (equivalent to the distortion transfer function) is constant independently of frequency — the responses are power complementary [3].

For the binDCT of type L3 [2], the orthogonality loss is evident, though it is characterized by a weak influence of rounding errors. The denormalized lattice is also robust in this aspect, but its responses are exactly power complementary. The CORDIC approach seems to be the worst option — due to high roundoff error and slight orthogonality loss. This is the result of a careless selection of wordlength, disallowing the precise multiplication by K_5^2, and can be eliminated by the enhancement of the used binary representation.

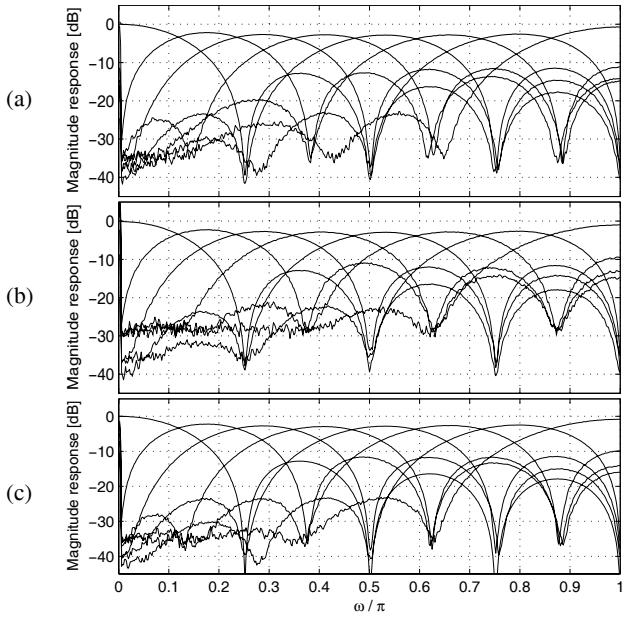

Fig. 7. Magnitude responses obtained using lattice (a), CORDIC (b) and binDCT-L3 (c).

6. CONCLUSIONS

An alternative approach to the finite precision DCT implementation is presented, focused on a structural orthogonality imposition. Two possible multiplierless lattice realizations are considered to reduce its computational complexity. The forthcoming works comprise a deeper analysis and further optimization of the circuits as well as their practical applications in image and video compression systems.

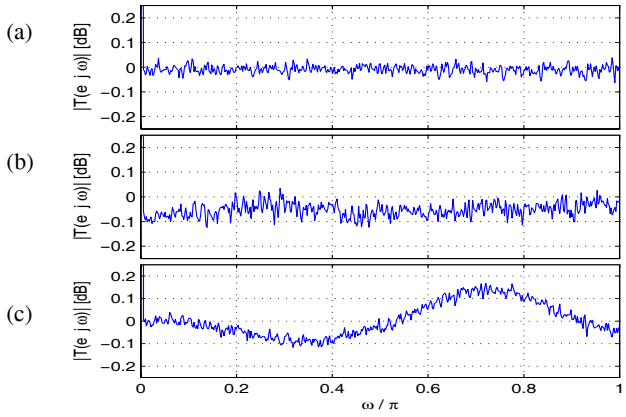

Fig. 8. Sum of squared magnitude responses using lattice (a), CORDIC (b) and binDCT-L3 (c).

7. REFERENCES

[1] T. D. Tran, "The BinDCT: Fast multiplierless approximation of the DCT," *IEEE Signal Processing Lett.*, vol. 7, no. 6, pp. 141–144, 2000.

[2] J. Liang and T. D. Tran, "Fast multiplierless approximations of the DCT with the lifting scheme," *IEEE Trans. Signal Processing*, vol. 49, no. 12, pp. 3032–3044, 2001.

[3] P. P. Vaidyanathan, "On coefficient-quantization and computational roundoff effects in lossless multirate filter banks," *IEEE Trans. Signal Processing*, vol. 39, no. 4, pp. 1006–1008, 1991.

[4] M. Parfieniuk and A. Petrovsky, "Hypercomplex factorizations for 8-channel linear phase paraunitary filter banks," in *Proc. 7th Int. Conf. and Exhibition "Digital Signal Processing and its Applications" (DSPA)*, Moscow, Russia, 16–18 March 2005, pp. 509–513.

[5] M. Parfieniuk and A. Petrovsky, "Quaternionic approach to 8-channel general paraunitary filter bank," in *Proc. 13th European Signal Processing Conf. (EUSIPCO)*, Antalya, Turkey, 4–8 September 2005, CD.

[6] P. P. Vaidyanathan and P.-Q. Hoang, "Lattice structures for optimal design and robust implementation of two-band perfect reconstruction QMF banks," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 1, pp. 81–94, 1988.

[7] S. Yu and E. E. Swartzlander, "A scaled DCT architecture with the CORDIC algorithm," *IEEE Trans. Signal Processing*, vol. 50, no. 1, pp. 160–167, 2002.

[8] A.-Y. Wu and C.-S. Wu, "A unified view for vector rotational CORDIC algorithms and architectures based on angle quantization approach," *IEEE Trans. Circuits Syst. I*, vol. 49, no. 10, pp. 1442–1456, 2002.

[9] P. Rieder, J. Götze, J. A. Nossek, and C. S. Burrus, "Parameterization of orthogonal wavelet transforms and their implementation," *IEEE Trans. Circuits Syst. II*, vol. 45, no. 2, 1998.

[10] S. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, Wiley, Chichester, 2 edition, 2000.