QUANTIZATION MODE OPPORTUNITIES IN FIXED-POINT SYSTEM DESIGN

D. Menard †, D. Novo‡, R. Rocher †, F. Catthoor‡, O. Sentieys †

† IRISA/INRIA, University of Rennes
6 rue de Kerampont
F-22300 Lannion
daniel.menard@irisa.fr

‡ IMEC vzw
Kapeldreef 75
B-3001 Leuven
novo@imec.be

ABSTRACT

Finding the optimal tradeoff in terms of area, delay and energy consumption which satisfies a given DSP functionality is the main objective of hardware and embedded software designers. Signal bit-widths importantly impact these metrics. Signals with less bits also require operators with smaller area, shorter critical path and lower energy consumption. In some applications, these minimal signal bit-widths can vary significantly depending on the quantization mode. As a result, a rounding-based implementation may require smaller minimal bit-widths than a truncation-based one and potentially lead to cheaper system implementations. The optimal quantization mode combination (QMC) can reduce significantly the implementation cost compared to a traditional implementation based on the truncation mode. This has been demonstrated on different representative kernels. For example, in the case of a LMS filter, the optimal QMC can reduce up to 46% of the area of an implementation based on truncation.

1. INTRODUCTION

Finding the optimal trade-off in terms of area, delay and energy consumption which satisfies a given DSP functionality is the main objective of hardware and embedded software designers. Signal's bit-widths importantly impact these metrics. Signals with less bits also require operators with smaller area, shorter critical path and inferior energy consumption. Software implementations can also benefit from reduced bit-widths as long as processors include some hardware support, such as sub-word parallel data-paths. In that particular case, the smaller the sub-word the more parallel operations can be perform simultaneously.

Fixed-point arithmetic typically requires less bits and simpler operators than floating point arithmetic, resulting in cheaper implementations. Unfortunately, fixed-point arithmetic introduces an unavoidable quantization error which degrades floating-point performances and needs to be carefully controlled.

Importantly, this quantization error depends on the quantization mode. The latter determines how an w bits signal is accommodated into $w-k$ bits by removing its k Less Significant Bits (LSB). Traditional methods are either truncation or rounding. Truncation simply discards the k LSB. Alternatively, rounding considers the highest of the k LSB to decide on the increment-by-1 of the resulting $w-k$ signal. At the operator level, rounding implies more hardware than truncation. For example, an $w \times w$-bit multiplier with w-bit output requires an extra w-bit adder in case that rounding is implemented. This comes with an increase in area, delay and energy consumption. However, rounding introduces a smaller quantization error than truncation.

An efficient fixed-point implementation contains the minimal signal bit-widths that satisfy a user-defined quantization error constraint. In some applications, these minimal signal bit-widths can vary significantly depending on the quantization mode. As a result, a rounding-based implementation may require smaller minimal bit-widths than a truncation-based one. Thus, despite the fact that rounding operators are more expensive, they may enable a cheaper implementation at the system level: they may need smaller minimum bit-widths to provide the same precision than the one offered by operators based on truncation. To the best knowledge of the authors, this trade-off has not been explicitly studied in previous works.

In this paper, the opportunities offered by quantization modes to optimize the design of fixed-point systems are analyzed. Representative kernels of relevant applications are implemented onto an FPGA considering different combinations for the quantization modes. The optimal combination of the quantization modes can reduce significantly the implementation cost compared to a traditional implementation based on the truncation mode. In our experiments, the optimal combination saves up to 46% of the area required by traditional implementation. The rest of the paper is organized as follows. In Section 2, the fixed-point conversion process is explained. The computing accuracy according to different quantization modes is analyzed in Section 3. In Section 4, the cost function is presented for the three quantization modes. The experiment results are presented in Section 5. Finally, Section 6 draws conclusions.

2. FIXED-POINT CONVERSION

The floating-point to fixed-point conversion process is made up of two main steps corresponding to the determination of the integer part word-length and the optimization of the fractional part word-length.

The first step of the fixed-point conversion process corresponds to the data dynamic range evaluation. These results are used to determine the integer part word-length which avoids overflows. For linear systems, an analytical approach such as the $L1$ or Chebycheff norms can be used. For non-linear and non-recursive systems, the interval arithmetic can be considered [4]. For the other systems, an estimation based on simulation [5] of representative inputs is required.

The second step of the floating-point to fixed-point conversion process determines the fractional part word-length of each data format. The number of bits for the fractional part modifies the computing accuracy. So, this step must be carried out with an accuracy constraint. It corresponds to an optimization process under constraints. The optimization is an iterative process which minimizes an implementation cost

 ISSN 2076-1465

$C(\mathbf{w})$ under an accuracy constraint $SQNR_{min}$ where $\mathbf{w}$ is the vector containing the data word-lengths of all variables.

$$\min(C(\mathbf{w})) \text{ with } SQNR(\mathbf{w}) \geq SQNR_{min} \quad (1)$$

The optimization process returns the fixed point configuration of minimal cost $C_{min}(\mathbf{w})$. For software implementation, the cost can be defined by the execution time and/or energy consumption of the application whereas for a hardware implementation, architecture area can also be included. The optimization process requires to evaluate the architecture cost $C(\mathbf{w})$ and the computation accuracy $SQNR(\mathbf{w})$ defined through the Signal to Quantization Noise Ratio (SQNR) metric at the output of the system. This metric corresponds to the ratio between the signal power and the quantization noise power due to finite precision. In this work, the computation accuracy is evaluated analytically [8]. At each iteration of the optimization process, the computation accuracy and the architecture cost is determined by the analytical method introduced in the next sections. A heuristic algorithm based on *Min + b bits* procedure [1] is used to significantly reduced the number of iterations required by the optimization process.

3. ACCURACY & QUANTIZATION MODES

The computing accuracy according to different quantization modes is analyzed in this section. The first and second order moments of the quantization noise are detailed, and the analytical model to evaluate the computing accuracy is presented.

3.1 Quantization noise statistics

Let x, be a fixed-point variable with a word-length of w bits. The word-length of the fractional part is equal to $n+k$ bits as presented in figure 1. The quantization process $Q()$ leads to the variable x_Q with a word-length of $w-k$ bits. Let X_Q be the set containing all the values which can be represented in the format after quantization. Let q, be the quantization step associated with the data x_Q. This term corresponds to the difference between two consecutive values of X_Q and is equal to the weight of the least significant bit b_{-n}. The term $\Delta = 2^{-(n+k)}$ is the quantization step associated with the data x before quantization. For continuous amplitude data, Δ goes to zero.

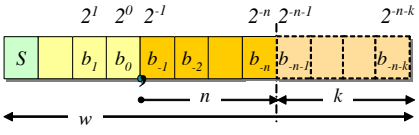

Figure 1: Fixed-point data specification

Let e_q, be the quantization error defined as difference between the data x and x_Q. The quantization process can be modeled as the sum of the data x and the quantization error e_q which is considered to be a uniformly distributed white noise [10]. This quantization noise is uncorrelated with the signal and other noise sources. A discrete noise model is used [3]. According to the type of quantization, the noise distribution will differ. In this work, three different quantization modes are considered: truncation, conventional rounding and convergent rounding.

3.1.1 Truncation

In the case of truncation, the data x is always rounded towards the lower value available in the set X_Q :

$$x_Q = \lfloor x \cdot q^{-1} \rfloor \cdot q = \quad kq \quad \forall x \in [k \cdot q; (k+1)q[\quad (2)$$

with $\lfloor \cdot \rfloor$, the floor function defined as $\lfloor x \rfloor = \max(n \in \mathbb{Z} | n \leq x)$. The value x_Q after quantization is always lower or equal to the value x before quantization. Thus, the truncation adds a bias on the quantized signal and the output quantization error will have a non zero mean.

The Probability Density Function (PDF) of the quantization noise e_q is given by expression 8 with δ being the Kronecker symbol.

$$p_{e_q}(x) = \frac{1}{2^k} \sum_{j=0}^{2^k-1} \delta(x - j \cdot \Delta) \quad (3)$$

3.1.2 Conventional rounding

To improve the precision after the quantization, the rounding quantization mode can be used. The latter significantly decreases the bias associated with the truncation. This quantization mode rounds the value x to the nearest value available in the set X_Q as defined below :

$$x_Q = \left\lfloor \left(x + \frac{1}{2} q \right) \cdot q^{-1} \right\rfloor \cdot q \quad (4)$$

The quantization value x_Q can also be expressed with the following equation:

$$x_Q = \begin{cases} kq & \forall x \in [k \cdot q; (k+\frac{1}{2})q[\\ (k+1)q & \forall x \in [(k+\frac{1}{2})q; (k+1)q] \end{cases} \quad (5)$$

The midpoint $q_{1/2} = (k+\frac{1}{2})q$ between kq and $(k+1)q$ is always rounded up to the higher value $(k+1)q$. Thus, the distribution of the quantization error is not exactly symmetrical and a small bias is still present. For this quantization mode, the PDF is given by the following relation

$$p_{e_q}(x) = \frac{1}{2^k} \sum_{j=-2^{k-1}}^{2^{k-1}-1} \delta(x - j.\Delta) \quad (6)$$

3.1.3 Convergent rounding

To reduce the small bias associated with the conventional rounding, the convergent rounding can be used. To obtain a symmetrical quantization error, the specific value $q_{1/2}$ must be rounded-up to $(k+1)q$ and rounded-down to kq with the same probability.

The probabilities that a particular bit is 0 or 1 are assumed to be identical and thus the rounding direction can depend on the bit b_{-n} value.

$$x_Q = \begin{cases} kq & \forall\, x \in [k.q; (k+\frac{1}{2})q[\\ (k+1)q & \forall\, x \in](k+\frac{1}{2})q; (k+1)q] \\ kq & \forall\, x = q_{1/2} \quad \text{and} \quad b_{-n} = 0 \\ (k+1)q & \forall\, x = q_{1/2} \quad \text{and} \quad b_{-n} = 1 \end{cases} \quad (7)$$

For the convergent rounding, the PDF is equal to

$$p_{e_q}(x) = \frac{1}{2^k}\sum_{j=-2^{k-1}+1}^{2^{k-1}-1}\delta(x-j.\Delta) + \frac{1}{2^{k+1}}\Big(\delta(x-2^{k-1}\Delta)+\delta(x+2^{k-1}\Delta)\Big) \quad (8)$$

Quantization mode	Truncation	Conventional rounding	Convergent rounding
Mean	$\frac{q}{2}(1-2^{-k})$	$-\frac{q}{2}(2^{-k})$	0
Variance	$\frac{q^2}{12}(1-2^{-2k})$	$\frac{q^2}{12}(1-2^{-2k})$	$\frac{q^2}{12}(1+2^{-2k+1})$

Table 1: Noise statistical parameters

3.2 Noise power

The mean μ_{e_q} and the variance $\sigma^2_{e_q}$ of the quantization error e_q are computed from the PDF. For the three quantization modes, the results are presented in Table 1. The quantization noises $e_q(n)$ at time n propagate through the different operators in the system and modify the computing accuracy generating an output noise $e_y(n)$. Each contribution $e'_{q_i}(n)$ of a quantization noise $e_{q_i}(n)$ comes from its propagation through the system which is characterized by its impulse response $h_i(k)$ which is time-varying in case of non-linear systems. As explained in [8] the output noise $e_y(n)$ is the sum of all the N_e noise source contributions. Given that the signal and noise terms are assumed to be uncorrelated, the output noise power P_b is obtained with expressions 9 and 10 [8].

$$P_b = \sum_{i=1}^{Ne}\sum_{j=1}^{Ne}\mu_{e_{q_i}}\mu_{e_{q_j}}G_{ij} + \sum_{i=1}^{Ne}\sigma^2_{e_{q_i}}G'_i \quad (9)$$

$$G_{ij} = \sum_{k=0}^{n\to\infty}\sum_{m=0}^{n\to\infty}E[h_i(k)h_j(m)] \quad G'_i = \sum_{k=0}^{n\to\infty}E[h_i^2(k)] \quad (10)$$

4. COST FUNCTION

Despite the proposed quantization mode exploration can be applied to different sorts of implementations, in this work hardware implementation is considered and the results are given only for area obtained with spatial implementations. In this case an operator is instantiated for each operation. For other implementations or other optimization goals only cost functions need to be modified. As an example, [2] proposes time execution and energy consumption cost functions for software implementations. In the next subsections, the cost function is presented for the three quantization modes.

4.1 Truncation cost function

Truncation rounding is widely used because of its cheapest implementation. The k LSB of x are discarded and no supplementary operation is required. Let γ_i be the term defining the kind of operation of o_i. The cost c of each operation o_i depends on the kind of operation γ_i and the operand word-lengths $\mathbf{w}(i)$. This cost is obtained from a library of synthesized operators. The implementation cost is estimated from the cost of each operation o_i. For the truncation, the global implementation cost C_T is expressed with expression 11 where i is the index on the arithmetic operations.

$$C_T(\mathbf{w}) = C_0(\mathbf{w}) = \sum_i c_{\gamma_i}(\mathbf{w}(i)) \quad (11)$$

4.2 Conventional rounding cost function

The conventional rounding can be directly implemented from equation (4) or by using the technique presented in [6]. In this case, the conventional rounding is obtained by the addition of x and the value $b_{-n-1}.2^{-n}$ and then the result is truncated on $w-k$ bits. This implementation requires an adder of $w-k$ bits.

Let $\mathbf{w}'$ be the vector of the data word-length before each quantization operation. Let $\mathbf{k}$ be the vector of the number of bits eliminated for each quantization operation. These two vectors are computed from the vector $\mathbf{w}$. The conventional rounding requires a supplementary addition operation for each quantization operation. For the conventional rounding, the global implementation cost C_R is expressed with expression 12 where j is the index on the quantization operations.

$$C_R(\mathbf{w}) = C_0(\mathbf{w}) + \sum_j c_{\text{ADD}}(\mathbf{w}'(j)-\mathbf{k}(j)) \quad (12)$$

4.3 Convergent rounding cost function

The specific value $q_{1/2}$ has to be detected to modify the computation in this case. For this specific value, the addition of the data x with the value 2^{-n-1} has to be done only if the bit b_{-n} is equal to one.

The alternative to this conditional addition is to add the value $b_{-n-1}.2^{-n}$ in every case. Then, for the specific value $q_{1/2}$, the least significant bit of the data x_Q (b_{-n}) is forced to 0 to obtain an even value. This last operation does not modify the result when b_{-n} is equal to 1 and discard the previous addition operation if b_{-n} is equal to 0.

Our implementation of this quantization is based on this last approach. The convergent rounding requires a supplementary addition operation and an operation (DTC) to detect the value 2^{-n-1} and then to force bit b_{-n} to zero. For the convergent rounding (CR), the global implementation cost C_{CR} is expressed with the following expression

$$C_{CR}(\mathbf{w}) = C_0(\mathbf{w}) + \sum_j c_{ADD}(\mathbf{w}'(j)-\mathbf{k}(j)) + c_{DTC}(\mathbf{k}(j)) \quad (13)$$

5. EXPERIMENTS

Different experiments on representative signal processing kernels have been conducted to compare the results obtained with the different quantization mode combinations (QMC). First, the example of the IIR filter is presented and then the results obtained for different signal processing kernels are given. For each quantization operation, the three quantization modes are tested. For a given QMC and accuracy constraint ($SQNR_{min}$), the fixed-point conversion is achieved and the implementation cost is optimized. The implementation cost C_{c_x} obtained for QMC c_x is compared with the cost C_T of a traditional implementation based only on truncation. To analyze the improvement of this QMC c_x, the relative QMC

gain $\phi_{c_x}^T$ of the cost c_x compared to the cost C_T is computed with expression 14. Let $\phi_{c_{optim}}^T$ be the relative QMC gain obtained with the combination c_{optim} which leads to the minimal implementation cost.

$$\phi_{c_x}^T = \frac{C_T - C_{c_x}}{C_T} \tag{14}$$

Let H be a first order infinite impulse response (IIR) filter having x as input and y as output. The expression of the output $y(n)$ is equal to $y(n) = x(n) - a.y(n-1)$ with $|a| < 1$. One noise source $e_g(n)$ due to the quantization of the addition output is considered. Let Γ be a term equal to 0 for the convergent rounding mode and 1 for the truncation. The expression of the power P_{e_y} of the output quantization noise $e_y(n)$ is as follows

$$P_{e_y} = \frac{q^2}{4}\left[\frac{1}{3}\left(\frac{1}{1-a^2}\right) + \Gamma.\left(\frac{1}{1-a}\right)^2\right] \tag{15}$$

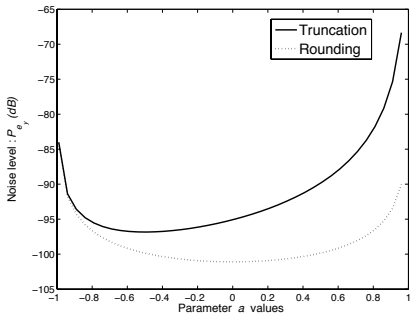

Figure 2: Output quantization noise level for different values of the parameters a

The noise level (P_{e_y}) is presented in figure 2 for different values of the parameter a in the case of rounding and truncation quantization mode. The results show that the difference of the noise level between the two quantization modes depends of the parameters a value. For a in the range $]-1;0[$, the impulse response $h(n)$ oscillates between positive and negative values. The sum of the impulse response terms leads to a small value in the range $[\frac{1}{2};1]$. Thus, the difference between the two quantization modes is small. For a in the range $]0;1[$, the impulse response is always positive and the sum of the impulse response terms can lead to a huge value and tends to infinity when a tends to 1. The difference between the two quantization increases when a tends to 1. To obtain the same noise power, 1 supplementary bit is required for the truncation for a in the range $[-0.78;0.4]$, 2 supplementary bits for a in the range $]0.4;0.82]$ and 3 bits for a greater than 0.82.

A second order IIR filter has been analyzed. As for the first order filter, the noise characteristics depends of the location of the filter poles. The case of two complex conjugate poles (ρ and ρ^*) have been considered. The gain between the noise source and the output depends of the impulse response of the transfer function between the source and the output. The impulse response depends of the pole modulus and the pole argument (θ) as follows

$$h_{b_y}(n) = |\rho|^n \frac{\sin((n+1)\theta)}{\sin(\theta)} \tag{16}$$

The relative QMC gain is presented in figure 3 for the different pole location inside the unity circle. When θ is low, the oscillation frequency of the impulse response envelope is low. Consequently, the sum of the impulse response terms can become huge when θ tends to 1 and $|\rho|$ tends to 1. The gain associated with the mean depends on the distance between the poles and the point $(1,0)$. On the results, the frontier between the regions are based on circles having this point as center. The maximal relative QMC gain is around 35%.

The relative QMC gain $\phi_{c_{optim}}^T$ have been measured for different DSP kernels and for different accuracy constraints between 30 dB and 90 dB. For each kernel, the mean value and the maximal value are reported in Table 2. For the different kernels, the maximal value of $\phi_{c_{optim}}^T$ is between 0.1% and 46.3% and the mean value for the different accuracy constraints is between 0% and 35.6%. The results show the opportunities offered by quantization modes to optimize the design of fixed-point systems. The optimal combination of the quantization modes can reduce significantly the implementation cost compared to a traditional implementation based on the truncation mode. The results show that the gain is kernel dependent and even for a given kernel varies depending on parameters such as the coeffcient values for filter, the number of taps for the LMS or the number of points for the FFT.

Application	$\max\left(\phi_{c_{optim}}^T\right) - (\%)$	$mean\left(\phi_{c_{optim}}^T\right) - (\%)$
FIR 16	8.9	5.5
IIR 2 ($\rho = 0.99$ $\theta = 0.003$)	36.6	25.2
IIR 2 ($\rho = 0.5$ $\theta = 3$)	0.1	0.01
FFT 128	30.8	23.9
FFT 1024	25.1	21.5
Voltera filter	7.7	4.7
LMS 32	46.3	35.6
LMS 128	44.2	29.9
APA	45	35.2
Sphere decoder	25.5	18.3

Table 2: Maximal and mean value of the optimal relative QMC gain $\phi_{c_{optim}}^T$ for different kernels

In the particular case of the IIR filter, the optimal QMC depends on the location of the filter poles. The case of a second order IIR filter (IIR 2) with two complex conjugate poles (ρ and ρ^*) has been considered. The gain G_{ij} associated to the mean between the noise source and the output depends on the impulse response h_i obtained from the pole modulus and argument (θ). When θ is low, the oscillation frequency of the impulse response envelope is low. Consequently, the sum of the impulse response terms can become huge when θ tends to 0 and $|\rho|$ tends to 1. The mean relative QMC gain is around 25.2% for the case of $\rho = 0.99$ and $\theta = 0.003$ and is null for the case of $\rho = 0.5$ and $\theta = 3$. In this last case, the truncation solution leads to a cheaper implementation. In conclusion, for low-pass filter, the rounding modes will provide better results and for high-pass filter, the truncation mode will be better.

Figure 4 shows the relative QMC gain $\phi_{c_x}^T$ for a 128 sized LMS algorithm obtained for six different QMC c_x and different accuracy constraints $SQNR_{min}$. Each QMC c_x is defined by a triplet (Q_1, Q_2, Q_3) defining the quantization mode Q for the input data, for the coefficients and for the filter output. The terms T, R and CR stand for Truncation, conventional Rounding and Convergent Rounding, respectively. The relative QMC gain depends on the accuracy constraint. For the

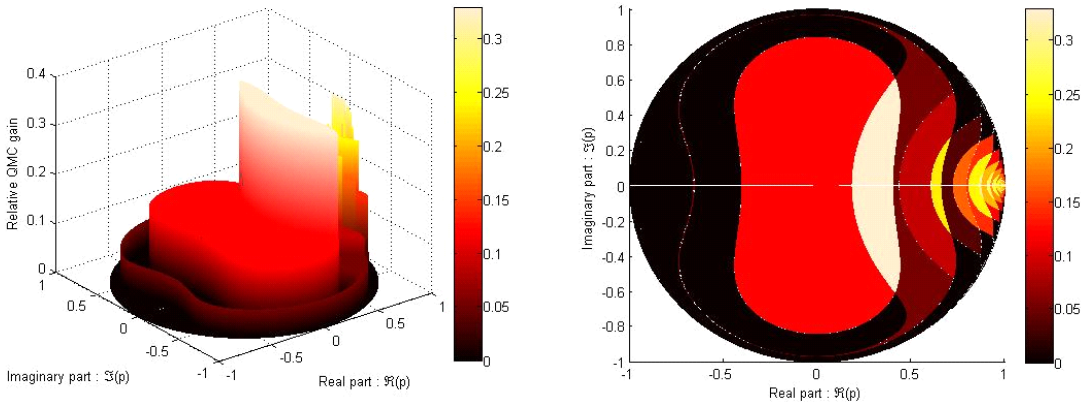

Figure 3: Relative QMC gain for different values of the poles ρ

optimal QMC, the gain varies from 44% to 22% for an accuracy constraint varying from 30 dB to 90 dB. The optimal QMC is obtained with the classical rounding for the input, the convergent rounding for the coefficients and the truncation for the filter output in all the SQNR. As shown in the output noise power expression, presented in [7], the noise due to coefficient quantization is the dominant source of noise. Moreover, the amplification gain G_{ij} associated with the mean of the coefficient quantization noise is very high due to the recursion inside this application. Thus, only the convergent rounding remove the effect of the mean of the coefficient quantization noise. Thus, specific LMS instructions, like in the C5000 DSP, incorporate automatically the rounding mode for coefficient update computation [9]. For the filter part, the effect of the mean of the output quantization noise is limited. Thus, the truncation leads to the best result because it reduces the implementation cost.

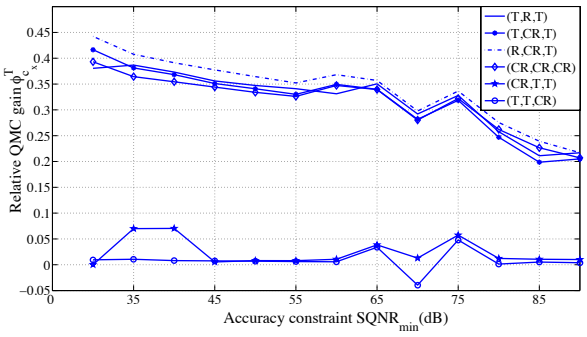

Figure 4: Relative QMC gain $\phi_{c_x}^T$ for different combinations c_x for a 128 sized LMS

6. CONCLUSION

A method to evaluate the impact of different quantization modes on the area of parallel DSP implementations was presented. The contribution of the different quantization modes on the output noise power depends on the DSP functionality. Quantization modes which imply a more expensive operator can lead to cheaper system implementations. This has been demonstrated on different representative kernels with different results. An optimal combination of the quantization modes can reduce significantly the implementation cost compared to a classical approach with truncation mode. Our experiments show that the implementation cost can be reduced up to 46%. The results underline the opportunities offered by an optimal combination of the quantization mode. They motivate the need for techniques, as the one described here, that explore the different quantization alternatives to select the optimal fixed-point implementation. The optimal combination of the quantization mode depends on the accuracy constraint and the kernel parameters (size, coefficients values, ...) and a general rule can not be established. Thus for each application with its specific parameters, the different quantization mode combinations have to be tested.

REFERENCES

[1] M.-A. Cantin, Y. Savaria, and P. Lavoie. A comparison of automatic word length optimization procedures. *Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on*, 2:II–612–II–615 vol.2, 2002.

[2] F. Catthoor, J. I. Gomez, S. Himpe, Z. Ma, P. Marchal, D. P. Scarpazza, C.Wong, and P. Yang. *Systematic methodology for real-time cost-effective mapping of dynamic concurrent task-based systems on heterogeneous platforms*, chapter 8. Springer Verlag, 2007.

[3] G. Constantinides, P. Cheung, and W. Luk. Truncation Noise in Fixed-Point SFGs. *IEE Electronics Letters*, 35(23):2012–2014, November 1999.

[4] R. Kearfott. Interval Computations: Introduction, Uses, and Resources. *Euromath Bulletin*, 2(1):95–112, 1996.

[5] S. Kim, K. Kum, and S. Wonyong. Fixed-Point Optimization Utility for C and C++ Based Digital Signal Processing Programs. *IEEE Transactions on Circuits and Systems II*, 45(11):1455–1464, November 1998.

[6] P. Lapsley, J. Bier, A. Shoham, and E. A. Lee. *DSP Processor Fundamentals: Architectures and Features*. Berkeley Design Technology, Inc, Fremont, CA, 1996.

[7] R. Rocher, D. Menard, P. Scalart, and O. Sentieys. Accuracy Evaluation of Fixed-point LMS algorithm. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2004)*, pages 237–240, Montreal, Canada, May 2004.

[8] R. Rocher, D. Menard, P. Scalart, and O. Sentieys. Analytical accuracy evaluation of Fixed-Point Systems. In *12th European Signal Processing Conference (EUSIPCO 2007)*, Poznan, Poland, September 2007.

[9] Texas Instruments. *TMS320C54X Dsp Cpu And Peripherals Reference Set Volume I*. Texas Instruments, Dallas, January 1999.

[10] B. Widrow, I. Kollár, and M.-C. Liu. Statistical Theory of Quantization. *IEEE Transactions on Instrumentation and Measurement*, 45(2):353–361, April 1996.