FIXED-POINT ACCURACY EVALUATION IN THE CONTEXT OF CONDITIONAL STRUCTURES

Jean-Charles Naud, Quentin Meunier, Daniel Menard, Olivier Sentieys

INRIA, IRISA, University of Rennes 1
6 rue de Kerampont, 22300, Lannion, France
jean-charles.naud@irisa.fr

ABSTRACT

The automation of fixed-point conversion requires generic methods to study accuracy degradation. Accuracy evaluation is often based on simulation approaches, at the cost of an important execution time. This paper proposes a new approach using fast analytical noise power propagation considering conditional structures. These structures are generated from programming language statements such as *if-then-else* or *switch*. The proposed model takes two key points into account in fixed-point design: first, an alternative processing of noise depending on the condition; second, decision errors generated by quantization noise affecting the condition. This method is integrated in the fixed-point conversion process and uses path probabilities of execution alternatives. This work extends existing analytical approaches for fixed-point conversion. Experiments of our analytical method show that it has a fairly accurate noise power estimation compared to the real accuracy degradation.

1. INTRODUCTION

The complexity of image and signal processing application, implemented in embedded systems, increases. These systems have usually limited resources. To reduce costs such as execution time, area, and power consumption, fixed-point arithmetic is widely used. Fixed-point arithmetic uses smaller data representation formats and less complex operators than floating point. The main drawback is the application performance degradation due to quantizations. This degradation needs to be studied and controlled to keep the system functional.

The fixed-point conversion requires between 25% and 50% of development time if done manually [1]. The trend shows an augmentation of development time due to algorithm complexity. In addition, the algorithm complexity increases the probability of error during manual conversion. To solve it, a higher abstraction level is required. Consequently, tools targeting an automatic fixed-point conversion have been proposed [2, 3].

One of the most sensitive step of fixed-point conversion corresponds to the numerical accuracy evaluation. Indeed, in the word-length optimization process, the numerical accuracy is evaluated many times. Thus, the evaluation method must be fast to ensure a reasonable optimization time and generic to support any system. Simulation based approaches are totally generic. However, their main drawback is the execution time required to get a good estimation of accuracy degradation. Simulation time increases with the complexity of the algorithm and quickly becomes impractical.

Analytical approach [4] using mathematical expressions helps decreasing the time taken for accuracy evaluation. Therefore, it is important to use them as much as possible even though they are not totally generic. In existing approaches based on perturbation theory and using propagation of the noise source moments [5], the conditional structures are not taken into account.

In this paper, an analytical method which considers conditional structures is presented and aims at extending our fixed-point conversion tool (ID.Fix). This tool converts a source C code with floating-point data types into a C code using fixed-point data types like those proposed by Mentor Graphics (ac_fixed) or in *System C*. Conditional structures are generated by *if-then-else* or *switch* statements that direct data flow in the algorithm. The difficulty lies in the propagation of the noise through these structures. Graphs have been chosen to implement our model and to evaluate it. In addition, this method should be integrated with existing one [6] for noise transmission through arithmetic operators. Our proposed approach is composed of two parts corresponding to the processing of noise through alternative paths and the management of the quantization noise occurring in the conditional value. This quantization noise generates decision errors that influence the accuracy of the application.

This paper is organized as follows. First, existing conversion methods applied to conditional structures are described in Section 2. Section 3 explains our analytic method to study noise power through conditional structures. This method is applied on a real context in Section 4. Finally, Section 5 concludes the paper.

2. RELATED WORK

Fixed-point arithmetic uses data words with a fixed number of bits to code both integer and fractional parts. The conversion process optimizes the Fractional part Word-Length (FWL) and the Integer part Word-Length (IWL) in order to reduce costs. First, to determine the IWL, the dynamic range of the data is evaluated such as overflows are avoided. Then, FWL minimization needs to study accuracy degradation to satisfy accuracy constraint keeping the system functional. This degradation is due to quantization noises resulting from the fixed-point format.

Automation of fixed-point conversion requires to be tool centric and to use generic methods to study the accuracy degradation. The goal is to build a tool with minimum restrictions for the user. Conversion tools generally such as in [2, 3] have been proposed in the literature.

Methods used to study the degradation are clearly categorized into either simulation based or analytic approaches. Many automatic conversion tools use fixed-point simulation

because this method is simple and generic and hence does not require restrictive hypotheses on algorithms [2, 3]. Nevertheless, to obtain accurate estimations, a great number of samples is necessary. Moreover, the optimization often requires design-space exploration for high number of combination of FWLs. Consequently, the optimization time can grow exponentially.

Analytical approach [7, 6, 5] uses mathematical expressions for noise power evaluation in order to accelerate the conversion process. The noise expression is propagated through the operators in the algorithm to extract the degradation. In the case of arithmetic operators, accuracy methods are already implemented. This paper introduces a new methodology to study accuracy degradation in conditional structures with an analytical approach to extend existing approaches.

Conditional structures can be generated by *if-then-else* or *switch* statements in the C language. The conditional structures direct data to different paths depending on a condition value. These structures produce two kinds of error in fixed-point. The first deals with quantization noises through alternative paths. The second is relative to noises affecting the conditional value generating paths error and are called decision errors.

Shi and Brodersen [8] has put in evidence three cases in which conditional structures can be found. The first assumes that no decision error can happen or else they can be ignored. The second and the third consider decision errors. The second case assumes that the noise quantization is independent from the signal and that the conditional structure corresponds to a continuous function (e.g. the absolute operator). This case is part of regular perturbation theory [7] which can treat quantization errors as a noise. Such errors are categorized as *weak* decision errors. The third case corresponds to a conditional structure which is not a continuous function. This decision error is categorized as *strong* and does not satisfy perturbation theory.

Other works such as mixed approaches [9] use both simulation and analytical techniques. The goal is to use a simulation approach when operators or structures cannot be solved by analytic methods. This approach is considered acceptable in simulation time if there is a reasonable number of unsupported operators or structures. However, in the case of systems with many complex structures or operators (e.g. conditional structures), this method is limited by the simulation time as explained previously.

Finally, work presented in [10] solves analytically the noise evaluation in conditional structures. However, this method uses the worst case, *i.e.* only the path regenerating the noise with the greatest power is considered. In the case of fixed-point conversion, worst case approach is too pessimistic and a more accurate approach is required.

3. ACCURACY EVALUATION

3.1 Approach based on perturbation theory

In [6] and [5], an approach based on perturbation theory is proposed. This approach considers the noise power at the algorithm output as the accuracy degradation. This approach is valid with algorithms modeled with linear time-varying functions. Its objective is to determine the expression of the output noise power as a function of the quantization noises and the system input noise. Under some hypothesis, the input noise can be modeled as a quantization noise source. In this case, the system can be modeled as shown in Figure 1.

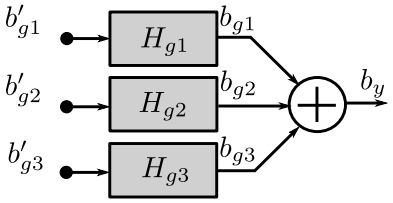


Figure 1: Noise Model

b'_{gi} corresponds to a quantization noise source and b_{gi} to the effect of b'_{gi} in the algorithm output. H_{gi} defines the time-varying system between b'_{gi} and b_{gi}. The term $h_{gi}(n)$ corresponds to the system impulse response of H_{gi} and can change in time. The noise b_y is

$$b_y = \sum_{i=1}^{N_g} b'_{gi}(n) * h_{gi}(n). \tag{1}$$

The noise power P_{b_y} corresponding to the second-order moment of the noise b_y is on

$$P_{b_y} = \sum_{i=1}^{N_g} a_i . \sigma^2_{b'_{gi}} + \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} A_{ij} . \mu_{b'_{gi}} . \mu_{b'_{gj}} \tag{2}$$

with $\mu_{b'_{gi}}$ and $\sigma^2_{b'_{gi}}$, the mean and variance of b'_{gi} and with

$$\begin{aligned} a_i &= \sum_{n=0}^{\infty} \mathrm{E}\left[h^2_{gi}(n)\right], \\ A_{ij} &= \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \mathrm{E}\left[h_{gi}(n) h_{gj}(m)\right]. \end{aligned}$$

In Equation, a_i and A_{ij} are constant because they depend on the algorithm semantic. Therefore, the output noise power P_{b_y} has only $\mu_{b'_{gi}}$ and $\sigma^2_{b'_{gi}}$ as variables and therefore depend on the data word-length.

3.2 Conditional Structures

To take conditional structures into account, the previous model is modified. The first modification step is the modeling of conditional structures. Previous work in [6] to automate the accuracy evaluation, uses directed graphs such as the Signal Flow Graph (SFG) to model the algorithm. The latter allows modelling data, operators and delays, but no control structures. To model conditional structures, phi nodes (φ) are introduced, as shown in the example in Figure 2, to model the convergence of data coming from different alternatives of the conditional structures.

In this example, the φ node merges the two versions of the data y coming from the *Then* and *Else* parts. The condition is $c > K$, with c being a variable and K a constant. Moreover, Directed Acyclic Graphs (DAG) impose a unique assignment for each variable to simplify FWL optimization. Therefore, y_1 and y_2 are introduced as predecessors of y.

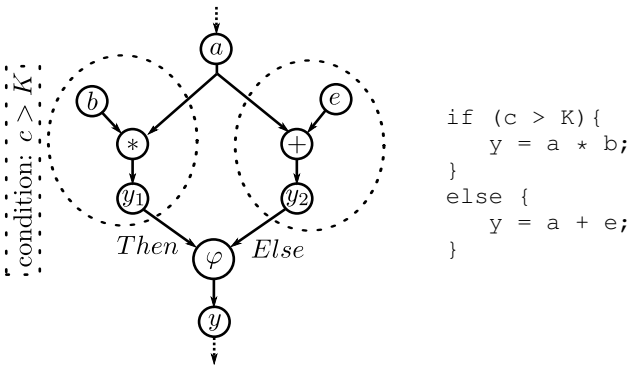
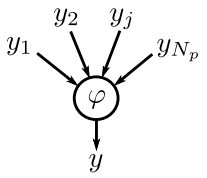


Figure 2: Conditional Structure with phi node (φ)

3.2.1 Phi Node (φ) Semantic

The operation of the φ node is similar to population mixing as described in Figure 3. The φ node has N_p inputs represented by y_i and one output noted y. y has the value of y_j if $c \in E_j$. In the example of Figure 2, E_1 corresponds to $c > K$ and E_2 to $c \leq K$. Therefore, a probability α_i associated with each alternative is determined and these probabilities satisfy Equation 3.



such that $y = y_j$
if $c \in E_j,\ j \in \{1,..,N_p\}$

Figure 3: φ Node as Population Mixing

$$P(Y = Y_j) = \alpha_j \text{ with } j \in \{1,...,N_p\},\ \sum_{j=1}^{N_p} \alpha_j = 1. \quad (3)$$

The goal of this model is to propagate the noise power *i.e.* the two first statistical moments of the noise through the algorithm. In this aim, the probability density function for mixed population is used at the output of the φ node as is Equation 4.

$$\int_{-\infty}^{y} f_Y(y)dy = \sum_{j=1}^{N_p} \alpha_j \int_{-\infty}^{y_j} f_{Y_j}(y_j)dy_j. \quad (4)$$

Mean (Equation 5) and variance (Equation 6) can be extracted out of Equation 4 if no decision errors happen or can be ignored.

$$E[Y] = \sum_{j=1}^{N_p} \mu_{y_j}\alpha_j \quad (5)$$

$$V[Y] = \sum_{j=1}^{N_p} (\sigma_{y_j}^2 + \mu_{y_j}^2)\alpha_j - (\sum_{j=1}^{N_p} \mu_{y_j}\alpha_j)^2 \quad (6)$$

3.3 Noise Propagation Model without Decision Errors

In case of conditional structures, the previous propagation model is adapted to obtain the new propagation model shown in Figure 4. This model divided into three parts corresponding to the consideration of noise sources inside conditional structures (Section 3.3.1), the extension to the generic model for several conditional structures (Section 3.3.2), and the combination of all noise sources (Section 3.3.3).

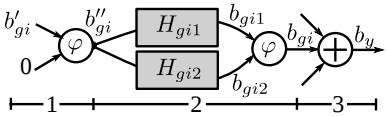


Figure 4: Noise Propagation Model with Conditional Structures

3.3.1 Noise Source Inside Conditional Structures

The first part of the transmission model allows us to consider quantization noise sources located inside one or several nested conditional structures. In this case, the noise source b'_{gi} is generated only when the associated conditional structure alternative is executed. In other words, b'_{gi} influences sometimes the system output depending on occurrence probability (α_{occ_i}). Therefore, in the proposed model, b''_{gi} becomes different from b'_{gi} because it represents b'_{gi} considering α_{occ_i}. The transmission of $\mu_{b'_{gi}}$ and $\sigma^2_{b'_{gi}}$ is obtained w.r.t. Equations 7 and 8.

$$\mu_{b''_{gi}} = \alpha_{occ_i}.\mu_{b'_{gi}} \quad (7)$$

$$\sigma^2_{b''_{gi}} = \sigma^2_{b'_{gi}}.\alpha_{occ_i} + \alpha_{occ_i}(1-\alpha_{occ}).\mu_{b'_{gi}} \quad (8)$$

In these equations, when b'_{gi} is not inside a conditional structure, $\alpha_{occ_i} = 1$ and implies $b''_{gi} = b'_{gi}$.

3.3.2 Generic Model for Several Conditional Structures

The second part corresponds to the modeling of general conditional structures. The existing graph transformations poduce the different H_{gi} in the previous noise propagation model. However, with conditional structures, these transformations result in a model which makes it impossible to directly extract the H_{gip} functions (see example Figure 5).

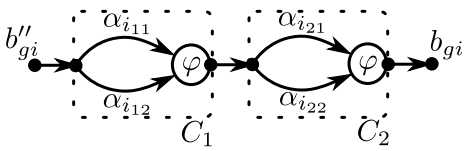


Figure 5: Noise Propagation Model with Conditional Structures

In Figure 5, edges represent the different functions between b''_{gi} and b_{gi}. C_1 and C_2 are conditional structures with two paths. $\alpha_{i_{11}}$ corresponds to the path probability of the

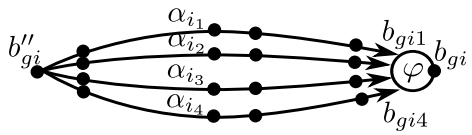


Figure 6: Noise Propagation Model with Conditional Structures

Then path of C_1. To obtain H_{gip} in the case of several structures arranged in serie or parallel, another graph transformation is needed and the result is shown in Figure 6.

This graph transformation dissociates all global paths between b''_{gi} and b_{gip}. In this example, the first global path probability is $\alpha_{i_1} = \alpha_{i_{11}}.\alpha_{i_{21}}$ if conditions C_1 and C_2 are independent. Moreover, this new graph allows to consider the dependence between several conditional structures. Finally, all global paths corresponding to different H_{ip} are extracted and they are used in equation 9 to calculate b_{gip}.

$$b_{gip}(n) = b''_{gi}(n) * h_{gip}(n) \tag{9}$$

All b_{gip} go through the φ node to obtain b_{gi}

3.3.3 Combination of all Noise Sources

The last part is the same as for the previous noise propagation model: the noise output b_y corresponds to the addition of all intermediate noises b_{gi}.

3.4 Noise Power with Conditional Structures

The noise power P_{b_y} at the output of the algorithm can be computed with the knowledge of all path probabilities α_{ip}, occurrence probabilities α_{occ_i} and H_{gip}. In a first time, "profiling" must be used and consists in simulating the algorithm in floating point to determine all probabilities with dependence between conditions. In a second time, H_{gip} are obtained by the graph transformations explained previously and the technique presented in [6]. The noise power is calculated with Equation 10 when no decision error occurs.

$$P_{b_y} = \sum_{i=1}^{Ng} \left[\sigma^2_{b''_{gi}}.a_i + \mu^2_{b''_{gi}}.b_i \right] + \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} A_{ij} \mu_{b''_{gi}} \mu_{b''_{gj}} \tag{10}$$

with

$$a_i = \sum_{i=1}^{N_{ip}} \alpha_{ip} \sum_{n=0}^{\infty} \mathrm{E}\left[h^2_{ip}(n)\right],$$

$$b_i = \sum_{i=1}^{N_{ip}} \alpha_{ip} \left(\sum_{n=0}^{\infty} \mathrm{E}\left[h_{ip}(n)\right] \right)^2 - \left(\sum_{i=1}^{N_{ip}} \alpha_{ip} \sum_{n=0}^{\infty} \mathrm{E}\left[h_{ip}(n)\right] \right)^2,$$

$$A_{ij} = \sum_{p=1}^{N_{ip}} \sum_{q=1}^{N_{ip}} \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \mathrm{E}\left[h_{ip}(n).h_{jq}(m)\right] \alpha_{ip}.\alpha_{jq}.$$

In Equation 10, a_i, b_i and A_{ij} are constant as they depend on h_{gip} and α_{ip}. This introduces a new term $\mu_{b''_{gi}}.b_i$. If no conditional structure is present, $\alpha_{occ_i} = 1$ and the path probability is obtained from Equation 11.

$$\alpha_{ip} = \begin{cases} 1 & \text{if } p = 1 \\ 0 & \text{if } p \neq 1 \end{cases} \tag{11}$$

In this case, the constant $b_i = 0$ and the general Equation 10 reduces to the expression given in Equation 2. This new model extends the previous one as it is more general.

3.5 Noise Propagation Model with Decision Errors

Conditional structures direct data into different paths depending on the conditional value. In the case of fixed-point implementation, quantization noise from various sources transit through different paths and merge at the end of the conditional structure. In addition, decision errors may appear if the condition value suffered from a value degradation due to the quantization noise. An instance in which a real value c is compared with a constant K in a conditional structure is shown in Figure 7.

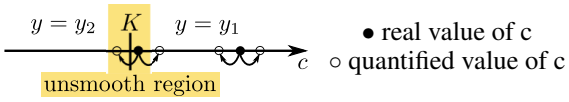


Figure 7: Unsmooth Region Representation

When the condition $c > K$ is true, the data is directed to the *Then* path. This is noted as $y = y_1$. Similarly with y_2, if the condition is false, the data is directed to the *Else* path. If the quantization noise influences c, the value c can traverse the boundary K and generates a decision error. The probability of decision errors increases if c is in an unsmooth region. Inversely, decision errors do not appear if c is far enough from the unsmooth region.

If decision errors are allowed, decision error probabilities are required. These probabilities can be obtained by simulation. However, a new simulation is required for each change in the data format (FWL) influencing the conditional value c. An analytical approach already exists in Shi's work [8].

Decision errors cannot be seen as a noise because they result in large errors at the output of φ node. Thus, the analytical noise model cannot work with decision errors. To solve this problem, the mixed approach proposed in [9] is used to deal with decision errors, using simulation when decision errors occur, and using our analytical method otherwise.

4. EXPERIMENTS

In order to demonstrate the validity of employing this method to study quantization noise, an IMDCT algorithm is chosen. The IMDCT (Inverse Modified Discrete Cosine Transform) is used for decoding mp3 audio streams.

4.1 Benchmark

Our new noise quantization propagation model is compared with results obtained by simulation and considered as a reference. This reference simulation consists in executing the algorithm in Infinite Precision (IP) and in Finite Precision (FP). The noise corresponds to the difference between IP and FP. Matlab is used to execute the IMDCT algorithm in IP and FP. The double precision used in Matlab can be a fair approximation of IP because Matlab uses a 53-bit mantissa. Moreover, Matlab can easily execute the new noise propagation model and extract results. IMDCT is composed of a conditional structure with four paths having a probability α_i. Two experiments are performed. In both experiments, the IMDCT has

18 inputs and 36 outputs. Results of statistical moments are compared to those of the reference simulation using 100,000 iterations of the IMDCT.

In the first experiment, no decision errors appear and two cases for α_i are considered: $\alpha_i = \{0.1, 0.1, 0.7, 0.1\}$ (case 1) and $\alpha_i = \{0.3, 0.3, 0.1, 0.3\}$ (case 2). A 16-bit FWL generates a standard degradation and is chosen for data in paths 1, 2 and 4. The operators in path 3 are assigned with the same number of bits. This assignment (FWL_3) varies from 8 to 24.

The second experiment includes strong decision errors, so the mixed approach is used. In this example, the same IMDCT is used but the value chosen for the paths are influenced by the quantization noise. In this experiment, α_i are ($[0.2, 0.2, 0.4, 0.2]$) and FWL is set to 16 bits. To appreciate the influence of decision errors, the conditional variable deciding of the path is influenced by the noise quantization (Qerr).

4.2 Results

Figure 8 shows an evaluation of the mean of output powers. The noise power is high when FWL_3 is comprised between 8 and 15. When increasing FWL_3, it becomes negligible compared to the quantization noise generated by other paths. This method is more realistic than the worst case and the accuracy of the model compared to the reference simulation is high since the error is less than 9.4%.

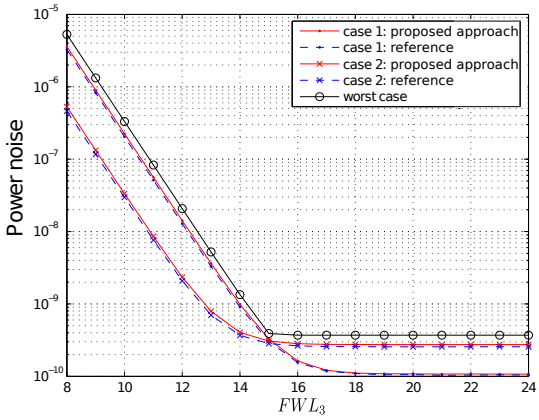


Figure 8: Proposed Approach without Decision Errors

The second experiment shows the effect of decision errors. Figure 9 puts in evidence the big influence of the noise (Qerr) even if the decision error probability P_{err} is small as shown in Table 1.

Var(Qerr)	7.9×10^{-8}	1.3×10^{-6}	2×10^{-5}	3.3×10^{-4}
P_{err}	3×10^{-4}	7×10^{-4}	3.3×10^{-3}	1.2×10^{-2}

Table 1: Decision errors probability

The mixed approach gives a good result in a realistic time since the biggest part of execution time is spent in simulation to compute decision errors.

5. CONCLUSION

Automatic tools for fixed-point conversion require an analytic approach to study accuracy degradation. All control

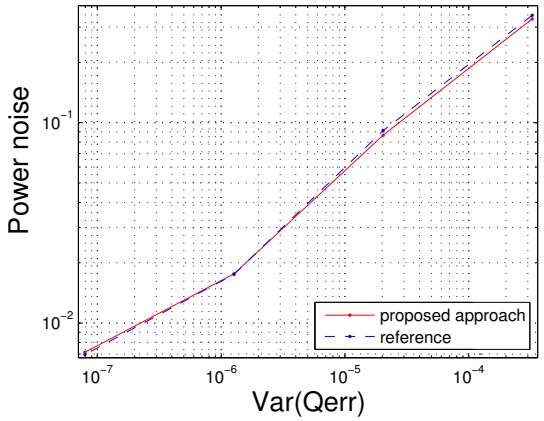


Figure 9: Proposed Approach with Decision Errors

structures, such as conditional structures, are not well supported by existing approaches. Conditional structures are generated by *if-then-else* or *switch* statement in programming languages.

The proposed analytical method supports conditional structures and provides the expression of the noise power when noise source are propagated through different conditional structure alternatives. Our approach is coupled with a mixed approach to handle decision errors. It combines both results of analytic and simulation-based methods. Two experiments with and without decision errors illustrate the quality of the proposed method. Estimation modeling error is less than 9.4%. Finally, a new method implementable in our fixed-point conversion tool is available. It extends existing approaches to evaluate analytically the quantization noise power.

6. ACKNOWLEDGMENTS

This work was supported by the Nano 2012 R&D research program in collaboration with STmicroelectronics.

REFERENCES

[1] B. Widrow, I. Kollar, and M.-C. Liu, "Statistical theory of quantization," *IEEE Transactions on Instrumentation and Measurement*, vol. 45, no. 2, pp. 353 –361, Apr. 1996.

[2] P. Belanovic and M. Rupp, "Automated floating-point to fixed-point conversion with the fixify environment," *IEEE International Workshop on Rapid System Prototyping*, pp. 172–178, 2005.

[3] H. Keding, M. Willems, M. Coors, and H. Meyr, "Fridge: a fixed-point design and simulation environment," *Design, Automation and Test in Europe, Proceedings, DATE*, pp. 429 –435, feb. 1998.

[4] G. Caffarena, J. A. López, A. Fernandez, and C. Carreras, "Sqnr estimation of fixed-point dsp algorithms," *Journal on Advance Signal Processing, Special issue on Design Methods for DSP Systems, EURASIP*, vol. 2010, 2010.

[5] R. Rocher, D. Menard, P. Scalart, and O. Sentieys, "Analytical accuracy evaluation of Fixed-Point Systems," *EUSIPCO*, September 2007.

[6] D. Menard, R. Rocher, and O. Sentieys, "Analytical fixed-point accuracy evaluation in linear time-invariant systems," *IEEE Transactions on Circuits and Systems I: Regular Papers, TCSI*, vol. 55, no. 10, pp. 3197 –3208, nov. 2008.

[7] C. Shi and R. Brodersen, "A perturbation theory on statistical quantization effects in fixed-point dsp with non-stationary inputs," *Proceedings of the International Symposium on Circuits and Systems. ISCAS.*, vol. 3, pp. 373–378, may. 2004.

[8] ——, "Floating-point to fixed-point conversion with decision errors due to quantization," *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP.*, vol. 5, pp. 41–44, may 2004.

[9] K. Parashar, R. Rocher, D. Menard, and O. Sentieys, "A hierarchical methodology for word-length optimization of signal processing systems," *International Conference on VLSI Design*, pp. 318–323, 2010.

[10] N. Doi, T. Horiyama, M. Nakanishi, and S. Kimura, "Minimization of fractional wordlength on fixed-point conversion for high-level synthesis," *ASP-DAC*, pp. 80–85, 2004.