AUTOMATIC SQNR DETERMINATION IN NON-LINEAR AND NON-RECURSIVE FIXED-POINT SYSTEMS

D. Menard, R. Rocher, P. Scalart and O. Sentieys

IRISA/ENSSAT
University of Rennes I
6, rue de kerampont
22300 Lannion, FRANCE
Email:name@enssat.fr

ABSTRACT

Most of the digital signal processing applications are implemented in embedded systems which are based on fixed-point arithmetic. The reduction of the time-to-market requires the automation of the fixed-point specification determination. The accuracy evaluation is one of the most important stage of this process. In this paper, a new methodology for evaluating the quality of non-recursive and non-linear systems is presented. The fixed-point specification accuracy is automatically determined through the computation of the Signal-to-Quantization-Noise-Ratio (SQNR) expression. The theoretical approach used for computing the output noise power is detailed and the methodology developed for automating the accuracy evaluation is presented. Then, the quality of our estimation is evaluated through different experiments.

1. INTRODUCTION

The efficient implementation of digital signal processing (DSP) algorithms in embedded systems requires to limit hardly the cost and the power consumption. Consequently, most of the systems are based on the fixed-point arithmetic to satisfy the embedded system constraints. The manual coding of fixed-point data is an error prone and a time-consuming task as illustrated in [6]. The embedded system time-to-market requires the reduction of the development time with the help of high-level tools allowing the automation of some tasks. Consequently, different methods for determining automatically the fixed-point specification have been proposed [8, 7, 11].

For an hardware or software implementation, the fixed-point conversion process is made-up of three main tasks. The data binary-point position is determined from the dynamic range of each data. Then, the fixed-point data format is optimized under accuracy constraint. For a software implementation, the data word-length and the scaling operation location are optimized to reduce the code execution time [11]. For an hardware implementation, the data word-length is optimized to minimize the chip area. These two optimization processes are achieved under accuracy constraint. Thus, the accuracy evaluation of a fixed-point specification is one of the most crucial task of the fixed-point conversion process. This task must be efficient in term of execution time. Indeed, the accuracy is evaluated many times during the fixed-point format optimization process.

In digital signal processing domain, the most common used criteria for evaluating the fixed-point specification accuracy is the Signal-to-Quantization-Noise-Ratio (SQNR) [8, 7, 9]. Most of the available methodologies for evaluating the SQNR are based on simulation [8, 5, 7]. These approaches lead to long execution times for the fixed-point format optimization process. Indeed, a new simulation is achieved when a fixed-point data format is modified. An alternative to the simulation based method can be an analytical approach which determines the SQNR expression. A method has already been proposed for obtaining automatically, the SQNR expression for linear time-invariant systems [12].

In this paper, a new approach is detailed for computing the SQNR expression in non-linear and non-recursive systems. Some examples of this kind of systems are presented in section 5. This approach allows to obtain an accurate estimation of the output noise power. Compared to previous works, no limiting assumptions are done on the statistical parameters of the system input signals and this method is valid for the different quantization modes. The paper is organized as follows. After an overview of the available methods for evaluating the accuracy, the theoretical concepts of our method are explained in section 3. Then, the techniques used for implementing this method are summarized in section 4. Finally, the quality of the SQNR estimation is analyzed through different experimental results in section 5.

2. RELATED WORKS

Most of the available methodologies for evaluating the fixed-point system accuracy are based on a bit-true simulation of the fixed-point application [8, 5, 4]. Nevertheless, this technique suffers from a major drawback which is the time required for the simulations [5]. The fixed-point mechanisms emulation on a floating-point workstation increases the simulation time compared to a classical floating-point simulation. Moreover, for obtaining an accurate estimation of the noise statistic parameters, a great number of samples must be taken for the simulation. Different techniques [8, 4, 5] have been investigated for reducing this simulation time. This drawback becomes a severe limitation when these methods are used in the process of fixed-point format optimization where multiple simulations are needed for exploring the design-space [8]. For each evaluation of the fixed-point specification accuracy, a new simulation is required.

To reduce dramatically the number of samples used for the accuracy estimation, the stochastic approach proposed initially for the floating point arithmetic has been adapted to the fixed-point arithmetic [2]. For this method, the output error is assumed to be a gaussian noise. The Student's distribution is used to estimate the number of significant bits from a weak number of output samples. This approach suffers from two major drawbacks. Firstly, this method is not valid for all fixed-point specifications because the gaussian noise assumption for the output error is no longer valid if only few error sources predominate. Secondly, this approach has been proposed for the rounding quantization mode but not for the truncation which is the most common mode used in embedded systems.

An alternative to the simulation based method can be an analytical approach. The system output Signal-to-Quantization-Noise-Ratio (SQNR) expression is determined according to the statistical parameters of the different quantization noise sources. In this case, the problem is to determine the output quantization noise power. The major advantage of this kind of technique is the execution time reduction of the fixed-point format optimization process. Indeed, the determination of the SQNR expression is done only once, then, the fixed-point system accuracy is evaluated through the computation of a mathematical expression.

Analytical expressions of the SQNR have been formulated for some DSP applications and more particularly for linear systems as in [10]. In [14], the authors have proposed a SQNR evaluation

methodology based on an analytical approach. For each type of operator the output noise is modelized by the sum of the input noises propagated through the operator and by the noise generated if a cast operation occurs. This model defines the operator output noise variance according to the input noise variance. The variance of the system output noise is obtained by applying this model to each operator during the traversing of the application Signal Flow Graph (SFG). Different restrictive assumptions have been made for defining the expressions of the operator output noise variance. The operator input variables are considered to be independent and centered. This last assumption is valid only for the rounding quantization mode. Given that all the variables inside the system are scaled between -1 and 1, the input signal power is set to its maximal value which is one. Thus, this model is independent of the signal statistical parameters (second order moment and cross-correlation). Nevertheless, these simplifying assumptions lead to a non-realistic estimation of the output noise power.

In this paper a new method for the automatic SQNR evaluation based on an analytical approach is proposed for non-linear and non-recursive systems. In non-recursive systems, the current value $y(n)$ of a variable depends no more on the past samples $y(n-i)$. Thus the application SFG contains no cycle. This new approach uses a realistic noise model which takes into account the different quantization modes (rounding and truncation). Moreover, no assumption on the signal statistical parameters is needed to compute the output noise power. This technique extends our previous works on linear time-invariant systems [12]. For linear systems, the statistical parameters of the noise sources and the transfer functions between the output and the noise sources are automatically computed.

3. THEORETICAL APPROACH

3.1 Noise models

For analyzing the error due to the fixed-point arithmetic two kinds of noise model are needed. The first one determines the quantization noise generated when a signal is quantified and the second one defines the propagation of these noises through the operators.

3.1.1 Quantization noise models

The use of fixed-point arithmetic introduces an unavoidable quantization error when a signal is quantized. A well known model has been proposed by Widrow in [15] for the quantization of a continuous-amplitude signal like in the process of analog-to-digital conversion. The quantization of a signal x is modeled by the sum of this signal and a random variable b_g. This additive noise b_g is a stationary and uniformly distributed white noise that is not correlated with the signal x and the other quantization noises. The validity conditions of this model are based on the signal x characteristic function [13]. Nevertheless, the model is valid as soon as the signal x dynamic range is sufficiently greater than the quantum step size and as soon as the bandwidth of x is enough large [13].

This model has been extended for modeling the computation noise in a system resulting from the elimination of some bits during a cast operation (fixed-point format conversion) if the number of bits eliminated k is sufficiently high [1]. Nevertheless, when k is small, the probability density function (PDF) of the quantization noise can no longer be assumed continuous. In [3], a model based on a discrete PDF is proposed and the first and second-order moments of the quantization noise are given according to the number of bits eliminated.

3.1.2 Propagation noise models

In this section, the propagation noise models are defined for the elementary arithmetic operations. These models define the operator output noise as a function of the operator inputs. An operator with two inputs X and Y and one output Z is under consideration. The input X and Y and the output Z are made up respectively of a signal x, y and z and a quantization noise b_x, b_y and b_z. Thus, for an adder, the expressions of the output signal z and the output quantization noise b_z are

$$Z = X \pm Y \quad \Rightarrow \quad \begin{cases} z = x \pm y \\ b_z = b_x \pm b_y \end{cases} \tag{1}$$

For the multiplication the expressions of z and b_z are

$$Z = X \times Y \quad \Rightarrow \quad \begin{cases} z = xy \\ b_z = b_x y + b_y x + b_x b_y \end{cases} \tag{2}$$

The term $b_x b_y$ represents the product of two quantization noises which is much smaller than the two other terms, then it is neglected in the following.

For the division, the output noise b_z expression is obtained by achieving a first-order decomposition in Taylor/Mac-Laurin series

$$Z = \frac{X}{Y} \quad \Rightarrow \quad \begin{cases} z = \frac{x}{y} \\ b_z = b_x . \frac{1}{y} - b_y . \frac{x}{y^2} \end{cases} \tag{3}$$

For the four elementary operators presented above, the operator output noise b_z is the weighted sum of the input noises b_x and b_y associated respectively with the first and second input of the operation. Thus, the function f_γ expressing the output noise b_z from the input noises is defined as follows for each kind of operation γ ($\gamma \in \{+, -, \times, \div\}$)

$$b_z = f_\gamma(b_x, b_y) = \alpha^{(1)} . b_x + \alpha^{(2)} . b_y \tag{4}$$

The terms $\alpha^{(1)}$ and $\alpha^{(2)}$ are associated with the noise located respectively on the first and second input of the operation. They are obtained only from the signal x and y and include no noise term.

3.2 Expression of the system output noise power

3.2.1 Expression of the system output noise b_y

Let consider, a non-recursive system made up of N_e inputs x_j and one output y. For multiple-output system our method is applied for each output. Let $\widehat{y}$ be the fixed-point version of the system output. The use of fixed-point arithmetic gives rise to an output computation error b_y which is defined as the difference between $\widehat{y}$ and y. This error is due to two types of noise sources. An input quantization noise is associated with each input $\widehat{x_j}$. When a cast operation occurs, some bits are eliminated and a quantization noise is generated. The same model, presented in section 3.1.1, is used for these two types of noise sources. Each noise source is a stationary and uniformly distributed white noise that is uncorrelated with the signals and the other noise sources. Thus, no distinction between these two types of noise sources is done in the rest of the study. A single type of quantization noise source b_{q_i} is considered.

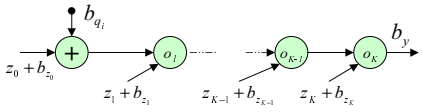

Figure 1: Computation graph example

To study the contribution to the global output noise b_y of the different quantization noise sources let consider the example presented in figure 1. A quantization noise source b_{q_i} is present and is propagated through the K operations o_k. The different noises b_{z_k} represent the propagation of the other quantization noise sources. The output noise b_y expression is obtained by replacing each operation o_k by its noise model $f_{\gamma_k}(b_x, b_y)$ defined in section 3.1.2. Thus, the output noise b_y can be expressed as follows

$$b_y = f_{\gamma_K}\left(f_{\gamma_{K-1}}\left(\ldots f_{\gamma_1}(b_{q_i} + b_{z_0}, b_{z_1}), b_{z_{K-1}}\right), b_{z_K}\right) \tag{5}$$

By introducing the expression of each operator output noise, given in equation 4, the system output noise is equal to

$$b_y = (b_{z_0} + b_{q_i}) \left(\prod_{k=1}^{K} \alpha_k^{(1)} \right) + b_{z_1} \alpha_1^{(2)} \left(\prod_{k=2}^{K} \alpha_k^{(1)} \right) ... + b_{z_K} \alpha_K^{(2)} \tag{6}$$

Each quantization noise source b_{q_i} leads to a noise, called b'_{q_i}, located at the system output. This noise is the product of the input quantization noise source b_{q_i} and the different signals α_k associated with each operation involved in the propagation of the noise source b_{q_i}. If the noise b_{q_i} is propagated through K operations o_k, the expression of b'_{q_i} is as follows

$$b'_{q_i} = b_{q_i} \prod_{k=1}^{K} \alpha_k^{(\iota_k)} = b_{q_i} \eta_i \qquad \text{with} \quad \eta_i = \prod_{k=1}^{K} \alpha_k^{(\iota_k)} \tag{7}$$

The term (ι_k) ($\iota_k = 1$ or 2) defines the input of the operation o_k which propagate the considered noise. The term η_i represents a signal obtained from the different signals $\alpha_k^{(\iota_k)}$ located at the operation o_k inputs. If there is more than one path between the system output and the noise source b_{q_i}, the paths having a different number of delay operations are processed separately. In this case, the noise source b_{q_i} is duplicated for each group of paths having the same number of delay operations. Given that these new sources are white noise, they are not correlated each other.

Equation 7 is applied to each noise source. From the generalization of equation 6, the output noise b_y can be expressed as the sum of the different noise source contributions b'_{q_i}. For a non-recursive system made-up of N_s quantization noise sources, the output noise b_y can be expressed as follows

$$b_y = \sum_{i=0}^{N_s-1} b'_{q_i} = \sum_{i=0}^{N_s-1} b_{q_i} \eta_i \tag{8}$$

3.2.2 Output noise power

The output noise power P_{b_y}, corresponding to the second order moment of b_y is given from equation 8 by

$$P_{b_y} = E(b_y^2) = \sum_{i=0}^{N_s} E(b'^2_{q_i}) + 2 \sum_{i=0}^{N_s} \sum_{\substack{j=0 \\ j>i}}^{N_s} E(b'_{q_i} b'_{q_j}) \tag{9}$$

Each term b'_{q_i} is replaced by its expression given in equation 7. From the quantization noise properties, each noise source b_{q_i} is not correlated with any signal η_i and with the other noise sources b_{q_j}, thus equation 9 can be written as follows

$$P_{b_y} = \sum_{i=0}^{N_s} E(b_{q_i}^2) E(\eta_i^2) + 2 \sum_{i=0}^{N_s} \sum_{\substack{j=0 \\ j>i}}^{N_s} E(b_{q_i}) E(b_{q_j}) E(\eta_i \eta_j) \tag{10}$$

For linear non-recursive systems, this approach gives the same results as those obtained in [12]. The computation of the noise power expression presented in equation 10 requires the knowledge of the statistical parameters associated with the noise sources b_{q_i} and the signal η_i. The second order moment $E(\eta_i^2)$ and the cross-correlation $E(\eta_i \eta_j)$ between the signals η_i and η_j are estimated statistically from the samples obtained with a floating-point simulation. These statistical parameters are independent of the fixed-point specification. Thus, the application can be simulated only once. The methodology developed for obtaining automatically these different parameters is summarized in the next section.

3.3 Example: vector normalization

Our approach is illustrated with the vector normalization example. Let X be a length N column vector given by $[x(0)\ x(1)\ \ldots\ x(N-1)]^t$. Let Y be the vector X normalized by its power defined by X^tX

$$Y = \frac{X}{X^tX} \tag{11}$$

Let $\widehat{X}$, be the quantized terms of X and B_q its quantization noise vector given by $B_q = [b_{q_0} b_{q_1} ... b_{q_{N-1}}]^t$. Thus, $\widehat{X}$ is the sum of X and B_q. For simplifying the presentation of this example, it is considered that no noise is generated inside the computation. Let $\widehat{Y}$ be the output fixed-point vector. Thus, by replacing each arithmetic operation by its noise model presented in 3.1.2, the global noise vector B_y associated to Y is given by

$$B_y = \widehat{Y} - Y = \frac{B_q}{X^tX} - \frac{2B_q^t XX}{X^tX^2} \tag{12}$$

The expression of the i^{th} element of the vector B_y is as follows

$$\begin{aligned} B_y(i) &= \frac{b_{q_i}}{X^tX} - \frac{2 \sum_{j=0}^{N-1} b_{q_j} x(j) x(i)}{X^tX^2} \\ &= b_{q_i} \left(\frac{1}{X^tX} - \frac{x(i)^2}{X^tX^2} \right) - 2 \sum_{\substack{j=0 \\ j \neq i}}^{N-1} b_{q_j} \frac{x(j)x(i)}{X^tX^2} \end{aligned} \tag{13}$$

Thus, the output noise can be expressed according to the noise sources b_{q_j} as follows

$$B_y(i) = \sum_{j=0}^{N-1} b_{q_j} \eta_{i,j} \quad \text{with} \quad \eta_{i,j} = \begin{cases} \frac{1}{X^tX} - \frac{x(i)^2}{X^tX^2} & \text{si } i = j \\ -2 \sum_{\substack{j=0 \\ j \neq i}}^{N-1} \frac{x(j)x(i)}{X^tX^2} & \text{si } i \neq j \end{cases} \tag{14}$$

The output noise power expression can be computed from equation 14 with the approach presented in section 3.2.2.

4. AUTOMATIC COMPUTATION OF THE SQNR EXPRESSION

The goal of this methodology is to compute the output SQNR expression of an application by using an analytical method based on the theoretical approach presented in the previous section. The different stages of this method are presented in figure 2. The front-end transforms the original application representation into a unique intermediate representation, called G_s, corresponding to the application Signal Flow Graph (SFG). It specifies the behaviour of the algorithm at the fixed-point level. The back-end determines the SQNR expression according to the fixed-point data format which are considered as variables.

For the back-end, the goal of the first stage is to represent the application at the quantization noise level through the graph called G_{sn}. The different noise sources are included in the graph G_s. The first and second order moments of each quantization noise source are determined from the model presented in section 3.1.1 according to the number of bit eliminated and the quantization mode used (truncation or rounding). Then, each operator is replaced by its noise propagation model.

The aim of the second stage is to determine the different parameters required for computing the expression of the output noise power. The graph G_{sn} is traversed to define the expression of the signal η_i associated with each noise source b_{q_i} (eq. 7). A floating-point simulation is achieved to collect the different signals inside

the system. From these results, the different samples of the signal η_i are determined from its expression obtained in the previous step. Then, the statistical parameters $E(\eta_i^2)$ and $E(\eta_i.\eta_j)$ are determined statistically from the simulation results.

In the third stage, the SQNR expression is computed from the output noise power expression as defined in equation 10.

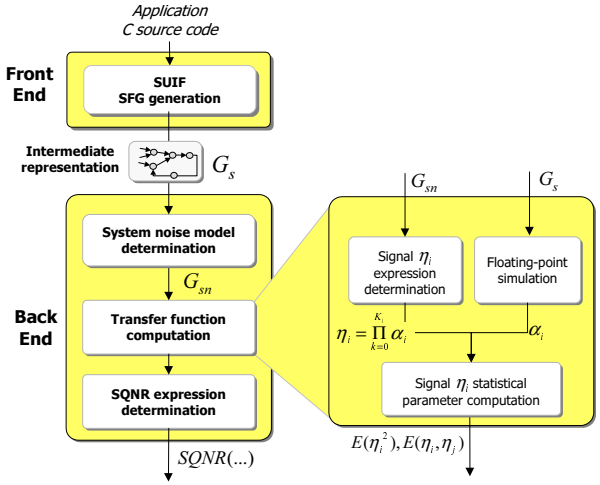

Figure 2: Methodology description

5. EVALUATION OF THE ESTIMATION QUALITY

The quality (accuracy) of the quantization noise power estimation has been evaluated through the measurement of the relative error between our estimation based on an analytical approach and the estimation based on simulation. The results obtained for several non-recursive and non-linear applications are presented in figure 3. For each application, different values of the relative errors are given. They correspond to the test of different fixed-point specifications. The applications correspond to the power computation of a real or a complex signal, the auto-correlation on N_s samples and a vector normalization which is used in the Normalized LMS filter. The Voltera filter tested is a second-order non-linear filter. The output $y(n)$ is computed from the following temporal equation

$$\begin{aligned} y(n) = & a_2 x(n-2) + a_1 x(n-1) + a_{22} x^2(n-2) \\ & + a_{11} x^2(n-1) + a_{21} x(n-1) x(n-2) \end{aligned} \quad (15)$$

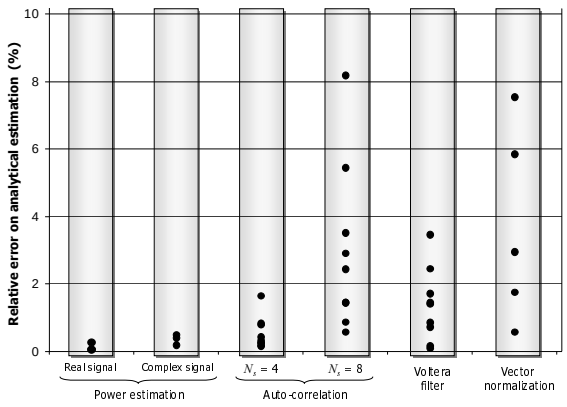

Figure 3: Relative error on P_{b_y} estimation

These results underline the quality of the noise power estimation. The maximal value of the relative error is weaker than 8%. For most of the experiments, the relative error is smaller than 4%. The estimation accuracy is definitively sufficient for our application corresponding to the design of fixed-point systems.

6. CONCLUSION

A new methodology for determining the SQNR expression in a non-linear and non-recursive fixed-point systems based on an analytical approach has been presented. This method extends the one proposed in [12] for linear time-invariant systems. The quality of our noise power estimation has been evaluated through different experiments. The results underline the accuracy of the estimation. This approach provides a significant improvement compared to the simulation based methods. With our method the time required for optimizing a fixed-point specification is definitively lower. It allows a complete design space exploration and the determination of an optimized solution.

REFERENCES

[1] C. Barnes, B. N. Tran, and S. Leung. On the Statistics of Fixed-Point Roundoff Error. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(3):595–606, 1985.

[2] J.-M. Chesneaux, L.-S. Didier, and F. Rico. Fixed CADNA library. In *RNC'5*, September 2003.

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

[4] M. Coors, H. Keding, O. Luthje, and H. Meyr. Fast Bit-True Simulation. In *DAC 01*, Las Vegas, US, June 2001.

[5] L. De Coster, M. Ade, R. Lauwereins, and J.A. Peperstraete. Code Generation for Compiled Bit-True Simulation of DSP Applications. In *ISSS 98*, Taiwan, December 1998.

[6] T. Grötker, E. Multhaup, and O.Mauss. Evaluation of HW/SW Tradeoffs Using Behavioral Synthesis. In *ICSPAT 96*, Boston, October 1996.

[7] H. Keding, M. Willems, M. Coors, and H. Meyr. FRIDGE: A Fixed-Point Design And Simulation Environment. In *DATE 1998*, 1998.

[8] 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.

[9] K. Kum and W. Sung. Word-Length Optimization For High Level Synthesis of Digital Signal Processing Systems. In *SiPS'98*, pages 142–151, Boston, October 1998.

[10] B. Liu. Effect of Finite Word Length on the Accuracy of Digital Filters - A Review. *IEEE Transaction on Circuit Theory*, 18(6):670–677, November 1971.

[11] D. Menard, D. Chillet, F.Charot, and O. Sentieys. Automatic Floating-point to Fixed-point Conversion for DSP Code Generation. In *CASES 2002*, Grenoble, October 2002.

[12] D. Menard and O. Sentieys. A methodology for evaluating the precision of fixed-point systems. In *ICASSP 2002*, Orlando, May 2002.

[13] A. Sripad and D. L. Snyder. A Necessary and Sufficient Condition for Quantization Error to be Uniform and White. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(5):442–448, Oct. 1977.

[14] J. Toureilles, C. Nouet, and E. Martin. A Study on Discrete wavelet transform implementation for a high level synthesis tool. In *EUSIPCO'98*, Rhodes, Greece, Septembre 1998.

[15] B. Widrow. Statistical Analysis of Amplitude Quantized Sampled-Data Systems. *Trans. AIEE, Part. II:Applications and Industry*, 79:555–568, 1960.