# A Noise Model for Evaluating Fixed-point System Performances

D. Menard, R. Rocher, O. Sentieys
IRISA/INRIA, University of Rennes,
6 rue de Kerampont
F-22300 Lannion
daniel.menard@irisa.fr

R. Serizel
K.U.Leuven, ESAT/SISTA
Kasteelpark Arenberg 10
B-3001 Leuven-Heverlee
romain.serizel@esat.kuleuven.be

***Abstract*— In fixed-point conversion process, the implementation cost is optimized under accuracy constraint. The determination of this constraint is one of the main issue of the conversion process. In this paper, a quantization noise model is proposed to evaluate the fixed-point application performances and thus to determine the computation accuracy constraint. The fixed-point system is modelled with a infinite precision version of the system and a unique noise source located at the system output. This noise source model is detailed in this paper. To validate our model, different Digital Signal Processing application benchmarks have been tested and the adequacy between our model and real noises has been measured.**

## I. Introduction

Fixed-point arithmetic is widespread in embedded systems to optimize power consumption and cost. Given that applications are developed with floating-point data types, a fixed-point conversion is required. The finite precision arithmetic leads to a quantization error which modifies the application functionalities and degrades the desired performances.

To maintain application performances, minimal computation accuracy must be guaranteed. In the fixed-point conversion process, the fixed-point specification is optimized such as the implementation cost is minimized as long as application performances are fulfilled. Nevertheless, the performance degradations are not analyzed directly in the conversion process. An intermediate metric is used to measure the computation accuracy. Indeed, the exploration of the fixed-point search space is more complex if the application performances are managed directly.

The global conversion method is decomposed into two main steps. Firstly, a computation accuracy constraint is determined according to the application performances, and secondly the architecture cost is minimized under this accuracy constraint during the fixed-point conversion process. This computation accuracy metric can be the quantization error power or the error bounds. The minimal value determination for the computation accuracy metric is a difficult problem and cannot be defined directly. This accuracy constraint has to be linked to the quality evaluation and performances of the application.

In this paper, a quantization noise model is proposed to evaluate the fixed-point application performances and thus to determine the computation accuracy constraint. In our approach, the metric used to evaluated the computation accuracy is the quantization noise power. The fixed-point system behavior is modelled with an infinite precision version of the system and a unique noise source located at the system output. The accuracy constraint is determined as the maximal value of the noise power which keeps the desired application quality. To our knowledge, no output quantization noise model is available, in the literature, to determine the computation accuracy constraint.

The paper is organized as follows. The fixed-point conversion process and the determination of the accuracy constraint are presented in Section II. The noise model is detailed and justified in Section III. This noise model is validated in Section IV and the adequacy between our noise model and real quantization noises is shown through examples.

## II. Accuracy constraint

### A. Fixed-point conversion process

A fixed-point data is composed of an integer part and a fractional part. The fixed-point conversion aim is to determine the number of bits for each part. Thus, as illustrated in figure 1, this process can be divided in two main modules [1]. The first module corresponds to the determination of the integer part word-length. Thus, firstly the dynamic range is evaluated for each data. Then, these results are used to determine, for each data, the binary-point position which minimizes the integer part word-length and which avoids overflow.

The second module corresponds to the determination of the fractional part word-length. The number of bits for this fractional part defines the computation accuracy. Thus, the data word-lengths are optimized such as the implementation cost is minimized under accuracy constraint. The fractional part word-length determination corresponds to an optimization problem where the implementation cost and the application accuracy must be evaluated. In our approach, the computation accuracy is evaluated through the quantization noise power. The analytical approaches have been favoured to evaluate the computation accuracy. Indeed, compared to the simulation-based techniques, they allow to obtained reasonable optimization time during the fixed-point space exploration. In the case of simulation-based approaches, a new fixed-point simulation is required when a fixed-point data format is modified.

The accuracy constraint corresponds to the maximal value for the noise power which allows to respect the desired application performances or quality. The approach used to obtain this noise power maximal value is presented in the next paragraph.

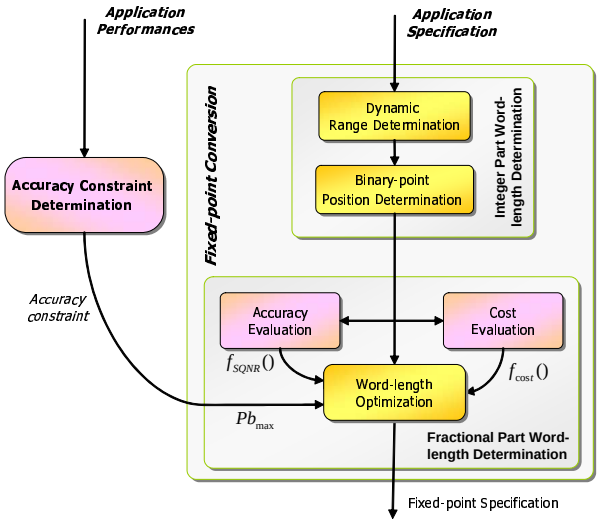


Fig. 1. Fixed-point conversion process

### B. Accuracy constraint determination

The accuracy constraint corresponding to the maximal value of the quantization noise power ($Pb_{max}$) is defined according to the system performance constraints. In our approach, the fixed-point system is modelled by the system infinite precision version and a unique noise source $b_y$ located at the system output. The accuracy constraint is determined from the maximal value of the noise power which allows to keep the desired application performances. The performances are measured by simulation. The system floating-point version is used and the noise $b_y$ is added to the output. The noise model used for $b_y$ is presented in Section III. The power of $b_y$ is increased as long as the measured performances are acceptable. Most of the time, the floating-point simulation has already been developed during the application design step, and the application output samples can be directly used. Therefore, the time required for exploring the noise power values is significantly reduced, and becomes negligible with regards to the global implementation flow.

Figure 2 shows the global process of accuracy constraint determination followed by the fixed-point design process. After the accuracy constraint determination, fixed-point conversion is achieved. The fixed-point specification is determined in order to optimize the implementation and to fulfil the accuracy constraint. This optimized fixed-point specification is simulated to measure the real performance obtained with this specification and to verify if the application performance constraints are still achieved. In the opposite case, the minimal value of the SQNR is adjusted and the fixed-point process is repeated. With this approach, few fixed-point simulations are required.

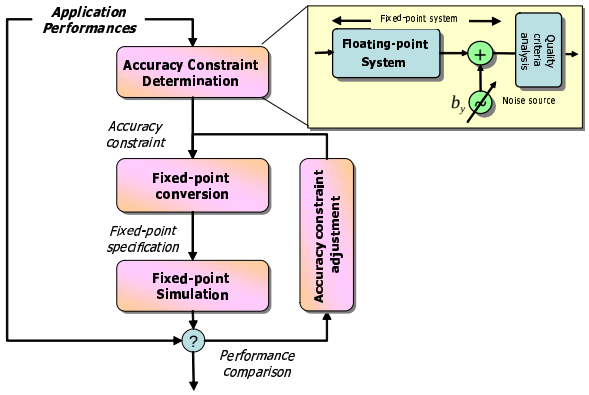


Fig. 2. Accuracy constraint determination and fixed-point design process

## III. Noise Model for Fixed-point System

The use of fixed-point arithmetic introduces an unavoidable quantization error when a signal is quantified. A common used model for signal quantization has been proposed in [2] and in [3]. The quantization of a signal $x$ is modelled by an additive noise $b$. This noise is a uniformly distributed white noise that is uncorrelated with the signal $x$, and independent from the other quantization noises. In this study, the round-off method is used rather than truncation. Quantization by rounding process leads to an error with a zero mean.

The output quantization noise is the contribution of the different noise sources. Each noise source is due to the elimination of some bits during a cast operation which follows an arithmetic operation. These different noise sources are independent of each other [3]. These noise sources are propagated through the different operations. A propagation model for each arithmetic operator is proposed in [4]. The operator output noise is a weighted sum of the input noises associated with each operation input. The weights of the sum do not include noise term, because the product between the noise terms can be neglected. Thus, it can be demonstrated that the output quantization noise is a weighted sum of the different noise sources. The contribution of each noise in terms of statistical parameters depends on the fixed-point format after quantization, and the gain between the output and the noise source.

In this context, two extreme cases can be distinguished. In the first case, a quantization noise source predominates in terms of variance compared to the other noise sources. A typical example is an extensive reduction of the number of bits at the system output compared to the other fixed-point format. In this case, the level of this output quantization noise exceeds the other noise source level. Thus, the probability density function of the output quantization noise is very closed to the one of the predominant noise source and can be assimilated to a uniform distribution. In the second case, an important number of independent noise sources have similar statistical

parameters and no noise source predominates. All the noise source are centered, uniformly distributed and independent of each other. By using the central limit theorem, the sum of the different noise sources can be modeled by a centered normally distributed noise.

From these two extreme cases, an intuitive way to modelize the output quantization noises of a complex systems is to use a noise $b$ which is the weighted sum of a gaussian and an uniform noise. Let $f_b$ be the probability density function of the noise $b$. Let $b_n$ be a normally distributed noise with a mean and variance equal respectively to 0 and 1. Let $b_u$ be a uniformly distributed noise in the interval [-1;1]. The noise $b$ is defined with expression 1. The $\beta$ weight is set in the interval $[0;1]$ and allows to represent the different intermediate cases between the two extreme cases presented above. The weight $\upsilon$ fixes the global noise variance.

$$b = \upsilon(\beta \times b_u + (1-\beta) \times b_n) \tag{1}$$

## IV. Validation of the Proposed Model

### A. Validation methodology

The aim of this validation section is to analyze the adequacy between our model and real quantization noises. Our model is valid if a $\beta$ weight can be found to model the noise probability density function with equation 1. The adequacy between the real noise and our model is analyzed with the $\chi^2$ goodness-of-fit test. This test is a statistical tool which can be used to know if an observed noise $b_y$ follows a chosen probability density function $f_b$ [5]. The test is based on the distance between the two probability density functions. If $y_s$ is the observed frequency for bin $s$, $E_s$ is the expected frequency for sample $s$ and $k$ the number of sample $s$, the statistical test is:

$$\chi^2 = \sum_{s=1}^{k} \frac{(b_{y_s} - E_s)^2}{E_s} \tag{2}$$

This statistical test follows a $\chi^2$ distribution with $k-1$ degrees of freedom. Therefore, if the distance is higher than a certain value, then the hypothesis $H_X$ ($b_y$ follows the probability density function $f_b$) is rejected. The significance level of the test is the probability to reject $H_X$ when the hypothesis is true. Choosing a certain value for this level will set the threshold distance for the test. According to [6], the significance level $\alpha$ should be in $[0.001\ 0.05]$.

Concerning the observed noise, there is no *a priori* knowledge of the $\beta$ weight. Thus the $\chi^2$ test has to be used collectively with a searching algorithm. The idea is, for a given observed noise, to find the $\beta$ weight for which the $f_b$ fits the best to the noise. In the interval $[0;1]$, different values of $\beta$ are tested, and, progressively, the search interval is reduced. When the test succeeds, the balance coefficient is found otherwise the noise can not be modelled with our approach. This procedure, called the $\beta$-searching algorithm, allows to check the validity of the model on several examples.

### B. FIR filter example

To illustrate the model, a 32-tap FIR filter example is under consideration. The signal flow graph of one cell $i$ is presented in Figure 3. The word-length of the input signal ($wl_x$) and the coefficient ($wl_h$) are equal to 16 bits. At the filter output, the data is stored in memory with a word-length $wl_o$ equal to 16 bits. The adder word-length $wl_{add}$ is varying between 16 and 32 bits.

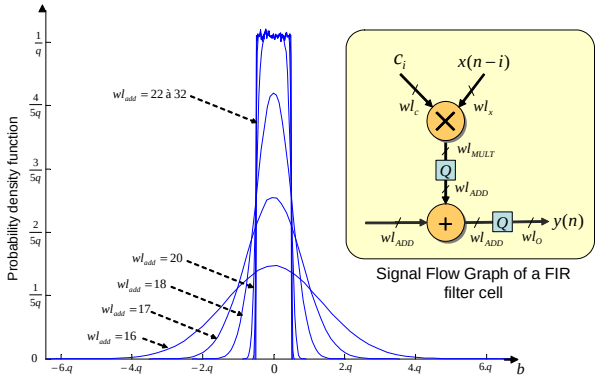


Fig. 3. PDF of the quantization noise $b$

The probability density function of the filter output quantization noise is presented in Figure 3 for different values of $wl_{add}$. The noise is uniform when one source is prevailing (the adder is on 32 bits). As the influence of the sources at the output of the multiplier is increasing (the length is decreasing), the distribution of the output noise tends to become gaussian. Theses simple visual observations can be confirmed using the $\beta$-searching algorithm. Figure 4 depicts the evolution of $\beta$ for different adder word-lengths, varying from 16 to 32 bits. When the output of the multiplier is on 16 or 17 bits, $\beta = 0$, the sources are numerous. Their influence on the system output is a gaussian noise. While the length of the multiplier is increasing, $\beta$ also grows and eventually tends to 1. When $wl_{add}$ is greater than 26 bits, the variance of the noise sources $b_{m_i}$ located at each multiplier output is insignificant compared to the variance of the noise source $b_o$ located at the filter output. Thus, this latter is prevailing. Its influence on the output signal is an uniform white noise.

### C. Benchmarks

To validate our noise model, different DSP application benchmarks have been tested and the adequacy between our model and real noises have been measured. The real noises are obtained through simulations. The output quantization noise is obtained from the difference between the system outputs obtained with a fixed-point and a floating-point simulation. The floating-point simulation which uses double-precision types is considered to be the reference. Indeed, in this case, the error due to the floating-point arithmetic is definitely weaker than the error due to the fixed-point arithmetic. Thus, the floating-point arithmetic errors can be neglected.

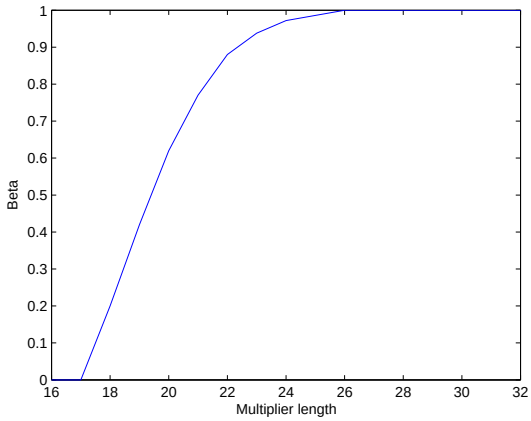


Fig. 4. Weight β found for different adder word-lengths. This weight is obtained with the β-search algorithm presented in section IV-A

| Applications | Test Number $N_t$ | Significance level $\alpha = 0.05$ | $\alpha = 0.001$ |
|---|---|---|---|
| FFT | 16 | 100 % | 100 % |
| IIR 8 Direct form I | 192 | 98 % | 99 % |
| IIR 8 Direct form II | 192 | 100 % | 100 % |
| IIR 8 Transposed form | 192 | 97 % | 99 % |
| Adaptive APA filter | 8 | 87 % | 100 % |
| Volterra filter | 8 | 100 % | 100 % |
| WCDMA receiver | 16 | 100 % | 100 % |
| MP3 | 28800 | 78 % | 87 % |

TABLE I

ADEQUACY BETWEEN OUR MODEL AND REAL NOISES

For each application, different output noises have been obtained by evaluating several fixed-point specifications and different application parameters. The number of output noises analyzed for one application is defined through the term $N_t$. For these different applications based on arithmetic operations, the input and the output word-length are fixed to 16 bits. The different fixed-point specifications are obtained by modifying the adder input and output word-length.

The results are presented in Table I for two significance level α corresponding to the boundary values (0.05 and 0.001). For each application, the number $N_s$ of real noise which can be modelled with our noise model are measured. The adequacy between our model and real noises is measured with the metric Γ defined as the ratio between $N_s$ and $N_t$. This metric corresponds to the ratio of output noises for which a β weight can be found to model the noise probability density function with equation 1.

The results show that our noise model can be applied to most of the real noises obtained for different applications. For some applications, like FFT, FIR, WCDMA receiver and Volterra filter, a β weight can be found. These four applications are non-recursive and the three first applications are linear time-invariant systems.

For the eight-order infinite impulse filter, almost all the noises (97%-100%) can be modelled with our approach. For these filters, 90% of the output quantization noise are modelled with a β weight equal to 0. Thus, the output noise is a purely normally distributed noise. In linear time-invariant systems, the output noise $b'_g$ due to the noise $b_g$ corresponds to the convolution of the noise $b'_g$ with $h_g$. This term $h_g$ is the impulse response of the transfer function between the noise source and the output. Thus, the output noise is the weighted sum of the delayed version of the noise $b_g$. The noise $b_g$ is a uniformly distributed white noise, thus the delayed versions of the noise $b_g$ are uncorrelated (the samples are not independent and thus the central limit theorem can not be applied directly). Even if only one noise source is located in the filter, the output noise is a sum of non-correlated noises and this output noise tends to have a gaussian distribution.

For the MP3 coder, when the level is 0.001 the test is successful about 87% of the time (78% when α is 0.05). The fairly high percentages tend to show that it is relevant to use this model. The second observation is that none of the found β values is higher than 0.5. The noises are mostly gaussian, and no source is prevailing.

## V. CONCLUSION

The accuracy constraint determination is one of the main issue in the floating-point to fixed-point conversion process. In this paper, a noise model has been proposed to model complex application fixed-point behavior. A system output noise is modelled from a gaussian noise and a uniform noise. The different experiments show that this model is adequate in most of the case for different DSP applications. Now, different experiments will be conducted to show the efficiency of our approach to determine the computation accuracy constraint and the adequacy between the desired performances and the real performances measured after the fixed-point conversion. Moreover, the case of quantization by truncation must be studied to find a noise model valid for this quantization law.

## REFERENCES

[1] D. Menard, D. Chillet, and O. Sentieys, "Floating-to-fixed-point Conversion for Digital Signal Processors," *EURASIP Journal on Applied Signal Processing, Special issue on Design Methods for DSP Systems*, vol. 2006, pp. 1–19, january 2006.

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

[3] 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*, vol. 25, no. 5, pp. 442–448, Oct. 1977.

[4] D. Menard, P. Quemerais, and O. Sentieys, "Influence of fixed-point DSP architecture on computation accuracy," in *11th European Signal Processing Conference (EUSIPCO 02)*, Toulouse, France, September 2002, pp. 587–590.

[5] D. E. Knuth, *The Art of Computer Programming, 2nd Ed. (Addison-Wesley Series in Computer Science and Information)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1978.

[6] A. J. Menezes, S. A. Vanstone, and P. C. V. Oorschot, *Handbook of Applied Cryptography*. Boca Raton, FL, USA: CRC Press, Inc., 1996.