Noise Probability Density Function in Fixed-Point Systems based on smooth operators

Romuald Rocher, Pascal Scalart

Noise Probability Density Function in Fixed-Point Systems based on smooth operators

Romuald Rocher, *Member, IEEE,* and Pascal Scalart, *Member, IEEE*

Abstract*— *To satisfy cost constraints, application implementation in embedded systems requires fixed point arithmetic. Thus, the application defined in floating point arithmetic must be converted into a fixed-point specification. This conversion requires accuracy evaluation to ensure algorithm integrity. Indeed, fixed-point arithmetic generates quantization noises due to the elimination of some bits during a cast operation. These noises propagate through the system and degrade computing accuracy. In this paper, a method based on Generalized Gaussian PDF is presented to model and generate the output noise of the system. The accuracy of the proposed model is evaluated through different experiments.

I. Introduction

Digital signal processing applications are specified in floating point to prevent problems due to computing accuracy. However, to satisfy cost constraints, application implementation in embedded systems requires fixed point arithmetic. Thus, the application defined in floating point arithmetic must be converted into a fixed-point specification. To reduce application time-to-market, tools to automate floating-point to fixed-point conversion are needed. In these tools, an important stage corresponds to accuracy evaluation of fixed-point specification. Indeed, fixed-point arithmetic generates quantization noises due to the elimination of some bits during a cast operation. These noises propagate through the system and modify computing accuracy. Computing accuracy deviations must be limited to ensure algorithm integrity and application performance.

Application accuracy can be evaluated by different manners. On one hand, accuracy evaluation can be obtained with fixed-point simulations [1], [5]. However, these methods require high computing time since a new simulation is required as soon as a format changes in the system. So, these approaches lead to very important optimisation time inside a fixed-point conversion process. On the other hand, a fixed-point specification accuracy can be evaluated with analytical methods [7], [3], [2]. These approaches determine a mathematical expression for the accuracy metric. In this paper, we present a method to study the fixed-point behaviour of a system with a floating-point simulation. Indeed, the fixed-point system can be represented by the floating-point system where a global noise is added at the output. This global noise must represent the quantization noise contributions due to fixed-point arithmetic whose characteristics are analytically determined. So, we need to be able to generate samples following the characteristics (PDF, power) of this global noise. In this way, we consider Generalized Gaussian law which allows us to generate samples following different PDFs (uniform, dirac, gaussian...). The parameters of the Generalized Gaussian are analytically determined from statistics (mean, variance and kurtosis) of the output noise.

This paper is organized as follows. We first introduce quantization noise models. The model of a noise source generation is presented and the propagation of these noises through the system is developed. The considered system is supposed to be composed by smooth operations such as additions, subtractions, multiplications and divisions. A general PDF based on Generalized Gaussian is presented and the approach to compute its parameters is defined. We also propose a method to generate samples following this PDF.

Then, the High Order Moments (HOM) are presented. They are necessary to determine the parameters of the Generalized Gaussian. First, the 3^{rd} (skewness) and 4^{th} (kurtosis) order moments of the quantization noises are computed for the different quantization types (truncation, rounding and convergent rounding). The 1^{st} and 2^{nd} order moments are already known. Then, the computation of the HOM of the output noise are analytically computed from the expression of the output noise.

Finally, the method is applied to different systems such as the FIR and IIR filters and Volterra filters. These results show the accuracy of our model for different noise PDFs.

II. Quantization Noise Models

A. Quantization Noise Sources

The quantization process can be modelled by the sum of the original signal and of a uniformly distributed white noise [11], [8]. This quantization noise is uncorrelated with the signal and the other noise sources. Such a model is valid provided that the signal has a sufficiently large dynamic range compared to the quantization step. According to the quantization mode, the noise Probability Density Function (PDF) will differ. Three quantization modes are usually considered: truncation, conventional rounding, and convergent rounding. In the truncation mode, the Least Significant Bits (LSB) are directly eliminated. The resulting number is always smaller than or equal to the value available before quantization, and, therefore, the quantization noise is always positive. Consequently, the mean of the quantization noise is not equal to zero. To reduce the bias due to truncation, the rounding quantization mode is often used. In conventional rounding, the data are rounded to the nearest value representable in the reduced-accuracy format. For numbers located at the midpoint between two consecutive representable values, the data are rounded-up always to the larger output value. This technique leads to a (small) bias for the quantization noise. To eliminate this quantization noise bias, the convergent rounding can be used as well. In this case,

the numbers located at the midpoint between two consecutive representable values are, with equal probability, rounded to the higher or lower output value.

Let n denotes the number of bits for the fractional part after the quantization process and k the number of bits eliminated during the quantization. The quantization step q after the quantization is equal to $q = 2^{-n}$. The quantization noise mean and variance are given in Table I for the three considered quantization modes [4], [6].

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 I

FIRST- AND SECOND-ORDER MOMENTS FOR THE THREE CONSIDERED QUANTIZATION MODES.

B. Output Noise Model

The system output noise $b_y(n)$ is the sum of all contributions as presented in Figure 1 and expressed by [7]

$$b_y(n) = \sum_{i=1}^{N_e} \sum_{k=0}^{n} h_i(k) b_i(n-k) \tag{1}$$

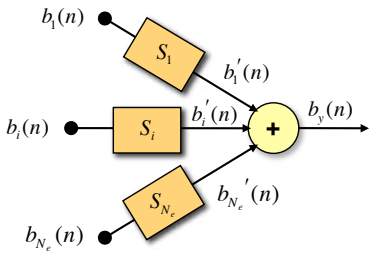

Fig. 1. System noise model for N_e input noises $b_i(n)$ and output noise $b_y(n)$

where $h_i(k)$ is the time varying impulse response of system S_i between the noise source $b_i(n)$ and the output noise $b_y(n)$. So, as shown in Figure 2, a fixed point system can be represented by a floating-point system whose output is perturbed (in an additive way) by a noise term by(n) modelling all quantization noise contributions.

Since fixed-point simulations are very time consuming, it is really interesting to be able to model the fixed-point behaviour using a global noise that is added to the output of a floating-point simulation.

C. Generalized Gaussian PDF

This noise must have the same characteristics as the global noise due to all quantization sources (mean, power, PDF). So,

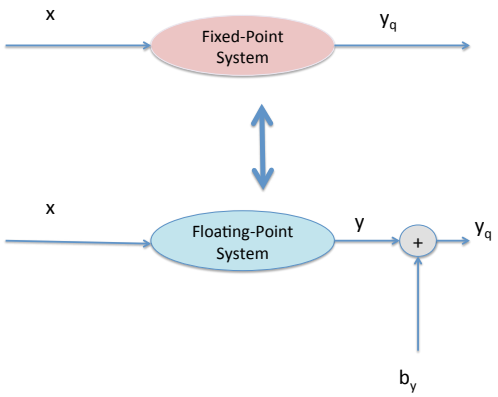

Fig. 2. Equivalence between a fixed-point simulation and a floating-point simulation with an additive noise

to generate this noise, we use the Generalized Gaussian PDF [9] defined by

$$f_G(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-|\frac{x-\mu}{\alpha}|^\beta} \tag{2}$$

where β and α are parameters defining the shape and the scale respectively, and where μ is the mean value of the random variable x.

The term Γ is the Gamma function defined by

$$\Gamma(k) = \int_0^\infty e^{-x} x^{k-1} dx \tag{3}$$

The variance of the Generalized Gaussian random variable can be expressed as a function of parameters α and β, or equivalently parameter α is defined by:

$$\alpha = \sqrt{\frac{\Gamma(3/\beta)}{\sigma^2\Gamma(1/\beta)}} \tag{4}$$

with σ^2 the variance of the variable x. Its 4^{th} order moment (normalized kurtosis $\frac{\kappa}{\sigma^4}$) can be computed [9] and is equal to

$$\frac{\kappa}{\sigma^4} = \frac{E(x-\mu)^4}{\sigma^4} = \frac{\Gamma(5/\beta)\Gamma(1/\beta)}{\Gamma(3/\beta)^2} \tag{5}$$

From [10] , β can be approximated as

$$\beta = \sqrt{\frac{5}{\frac{\kappa}{\sigma^4} - 1.865}} - 0.12 \tag{6}$$

for $1.865 < \frac{\kappa}{\sigma^4} < 30$. For values less than 1.865, the signal x can be modelled with a uniform PDF, whereas for values higher than 30, it can be represented as a dirac at zero.

So, with the determination of the mean, the variance and the kurtosis of the output noise, all parameters defining the Generalized Gaussian PDF can be computed. First, β is computed from the knowledge of the kurtosis value and (6). Then, α can be determined from the values of β and σ, and thus, the PDF is completely expressed. So, in the section

III-B, we will present a method to compute analytically these parameters.

As we are interested in the generation of samples $x(n)$ following a Generalized Gaussian PDF, we demonstrate in the Appendix that this operation can easily realized by the multiplication of two random variables $u(n)$ and $y(n)$ as :

$$x = u.y^{1/\beta} \tag{7}$$

where $u(n)$ are uniform data over $[-\alpha, \alpha]$ and $y(n)$ are samples following the PDF $\Gamma(1/\beta + 1, 1)$ at the power $1/\beta$.

III. High Order Moments

In this section, we present a model to compute the HOM of the output noise. Indeed, to determine the Generalized Gaussian corresponding to the output noise $b_y(n)$ the moments $1-4$ must be evaluated. This evaluation must be analytical to prevent a fixed-point simulation. First, HOM of quantization noises are computed. Then, we can use these values to compute the HOM for the output noise.

A. Input noise High Order Moments

In this subsection the HOM of the quantization are determined for the 3 types. We consider a quantization of k eliminated bits and we note $\Delta = 2^{-n-k}$ the quantization step before the quantization and $q = 2^{-n}$ is the quantization step after the quantization. The term n is the bit number of the fractional part after quantization.

1) Quantization by truncation: The noise PDF is represented Figure 3 and is equal to

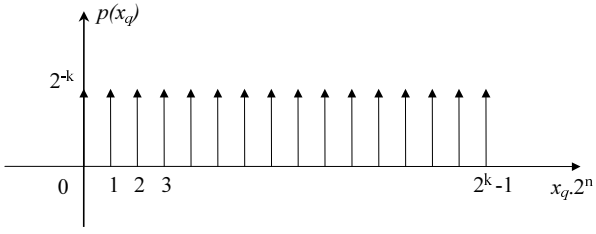

Fig. 3. PDF of a noise generated by quantization by truncation

$$p_b(x) = \frac{1}{2^k} \sum_{j=0}^{2^k-1} \delta(x - j.\Delta) \tag{8}$$

So its 3^{rd} order moment $E(x^3)$ is equal to

$$\begin{aligned} E(x^3) &= \int_{\mathbb{R}} p_b(x) x^3 dx \\ &= \frac{\Delta^3}{2^k} \sum_{j=0}^{2^k-1} j^3 \\ &= \frac{\Delta^3}{2^k} \frac{2^{2k}(2^k-1)^2}{4} \\ &= \frac{\Delta^3}{4}(2^{3k} - 2^{(2k+1)} + 2^k) \\ &= \frac{q^3}{4}(1 - 2.2^{-k} + 2^{-2k}) \end{aligned} \tag{9}$$

The skewness γ is defined by

$$\begin{aligned} \gamma &= E(x-\mu)^3 \\ &= E(x^3) - 3\mu\sigma^2 - \mu^3 \end{aligned} \tag{10}$$

Applying values in Table I and (9), the skewness is equal to zero.

The 4^{th} order moment $E(x^4)$ is equal to

$$\begin{aligned} E(x^4) &= \int_{\mathbb{R}} p_b(x) x^4 dx \\ &= \frac{\Delta^4}{2^k} \sum_{j=0}^{2^k-1} j^4 \\ &= \frac{\Delta^4}{2^k}(\sum_{j=0}^{2^k} j^4 - 2^{4k}) \\ &= \frac{\Delta^4}{2^k}\Big(\frac{2^k(2^k+1)(6.2^{3k} + 9.2^{2k} + 2^k - 1)}{30} - 2^{4k}\Big) \\ &= \Delta^4\Big(\frac{(2^k+1)(6.2^{3k} + 9.2^{2k} + 2^k - 1)}{30} - 2^{3k}\Big) \\ &= \Delta^4\Big(\frac{6.2^{4k} - 15.2^{3k} + 10.2^{2k} - 1}{30}\Big) \\ &= \frac{q^4}{5} - \frac{q^4.2^{-k}}{2} + \frac{q^4.2^{-2k}}{3} - \frac{q^4.2^{-4k}}{30} \\ &= \frac{q^4}{5}(1 - \frac{5}{2}2^{-k} + \frac{5}{3}2^{-2k} - \frac{1}{6}2^{-4k}) \end{aligned} \tag{11}$$

The kurtosis κ is defined by

$$\begin{aligned} \kappa &= E(x-\mu)^4 \\ &= E(x^4) - 4\mu\gamma - 6\mu^2\sigma^2 - \mu^4 \end{aligned} \tag{12}$$

Applying values in Table I and (11), the kurtosis is equal to

$$\kappa = \frac{q^4}{80}(1 - \frac{10}{3}2^{-2k} + \frac{7}{3}2^{-4k}) \tag{13}$$

More generally, the kurtosis is normalized by σ^4 to be able to compare the PDF of different signal leading to

$$\frac{\kappa}{\sigma^4} = 1.8\frac{1 - \frac{10}{3}2^{-2k} + \frac{7}{3}2^{-4k}}{1 - 2.2^{-2k} + 2^{-4k}} \tag{14}$$

2) Quantization by rounding: In this subsection, we compute the HOM of noises generated by quantization by rounding. Its PDF is presented in Figure (4) and is equal to

$$p_b(x) = \frac{1}{2^k} \sum_{j=-2^{k-1}}^{2^{k-1}-1} \delta(x - j.\Delta) \tag{15}$$

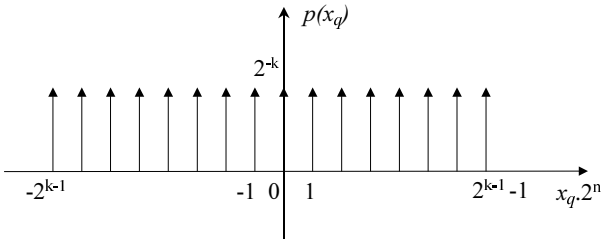

Fig. 4. PDF of a noise generated by quantization by rounding

Its 3^{rd} order moment is computed as follows

$$\begin{aligned} E(x^3) &= \int_{\mathbb{R}} p_b(x)x^3dx \\ &= \frac{\Delta^3}{2^k} \sum_{j=-2^{k-1}}^{2^{k-1}-1} j^3 \\ &= \frac{\Delta^3}{2^k} 2^{(3k-3)} \\ &= \frac{\Delta^3}{8} 2^{2k} \\ &= \frac{q^3}{8} 2^{-k} \end{aligned} \tag{16}$$

(17)

It skewness is equal to

$$\gamma = 0 \tag{18}$$

The 4^{th} order moment is equal to

$$\begin{aligned} E(x^4) &= \int_{\mathbb{R}} p_b(x)x^4dx \\ &= \frac{\Delta^4}{2^k} \sum_{j=-2^{k-1}}^{2^{k-1}-1} j^4 \\ &= \frac{\Delta^4}{2^k} \Big(2 \sum_{j=0}^{2^{k-1}} j^4 - 2^{4(k-1)} \Big) \end{aligned} \tag{19}$$

Using the fact that $q = \Delta.2^k$, this expression is developed and leads to

$$E(x^4) = \frac{q^4}{80}(1 + \frac{20}{3}2^{-2k} - \frac{8}{3}2^{-4k}) \tag{20}$$

So, its kurtosis is computed

$$\kappa = \frac{q^4}{80}(1 - \frac{10}{3}2^{-2k} + \frac{7}{3}2^{-4k}) \tag{21}$$

It is equal to the kurtosis of a noise generated by truncation.

3) Quantization by convergent rounding: For quantization by convergent rounding, the PDF is given in Figure (5) and is equal to

$$p_b(x) = \frac{1}{2^K} \Big(\sum_{j=-2^{k-1}+1}^{2^{k-1}-1} \delta(x - j.\Delta) + \frac{\delta(x - 2^{k-1}.\Delta)}{2} + \frac{\delta(x + 2^{k-1}.\Delta)}{2} \Big) \tag{22}$$

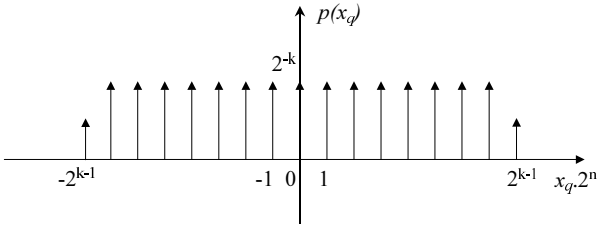

Fig. 5. PDF of a noise generated by quantization by convergent rounding

Its 3^{rd} order moment is given by the following relation

$$\begin{aligned} E(x^3) &= \int_{\mathbb{R}} p_b(x)x^3dx \\ &= \frac{\Delta^3}{2^k} \Big(\sum_{j=-2^{k-1}-1}^{2^{k-1}-1} j^3 - \frac{2^{k-1}}{2} + \frac{2^{k-1}}{2} \Big) \\ &= 0 \end{aligned} \tag{23}$$

It skewness is equal to

$$\gamma = 0 \tag{24}$$

The term $E(x^4)$ is also developed

$$\begin{aligned} E(x^4) &= \int_{\mathbb{R}} p_b(x)x^4dx \\ &= \frac{\Delta^4}{2^k} \Big(\sum_{j=-2^{k-1}-1}^{2^{k-1}-1} j^4 + 2\frac{2^{4(k-1)}}{2} \Big) \\ &= \frac{\Delta^4}{2^k} \Big(2 \sum_{j=0}^{2^{k-1}} j^4 - 2^{4(k-1)} \Big) \end{aligned} \tag{25}$$

Using the same appraoch as before, it leads to

$$E(x^4) = \frac{q^4}{80}(1 + \frac{20}{3}2^{-2k} - \frac{8}{3}2^{-4k}) \tag{26}$$

Its kurtosis κ is equal to

$$\kappa = \frac{q^4}{80}(1 + \frac{20}{3}2^{-2k} - \frac{8}{3}2^{-4k}) \tag{27}$$

The normalized kurtosis is given by

$$\frac{\kappa}{\sigma^4} = 1.8 \frac{1 + \frac{20}{3}2^{-2k} - \frac{8}{3}2^{-4k}}{1 + 4.2^{-2k} + 4.2^{-4k}} \tag{28}$$

All previous computation of HOM of quantization noises are summarized in the Table II.

Quantization mode	Truncation and Conventional rounding	Convergent rounding
Skewness γ	0	0
Normalized Kurtosis $\frac{\kappa}{\sigma^4}$	$1.8\frac{1-\frac{10}{3}2^{-2k}+\frac{7}{3}2^{-4k}}{1-2.2^{-2k}+2^{-4k}}$	$1.8\frac{1+\frac{20}{3}2^{-2k}-\frac{8}{3}2^{-4k}}{1+4.2^{-2k}+4.2^{-4k}}$

TABLE II
3^{rd} AND 4^{th}-ORDER MOMENTS FOR THE THREE CONSIDERED QUANTIZATION MODES.

It can be noted that, for a only one eliminated bit ($k = 1$) the normalized kurtosis is equal to 1 for quantization by truncation and rounding and equal to 2 for convergent rounding. For a high number of eliminated bits $k \to \infty$, the normalized kurtosis converges to 1.8, which is the classical value for uniform signals with continuous amplitude.

B. Output noise High Order Moments

In this subsection, the computation of the skewness and kurtosis of the output noise $b_y(n)$ is presented. The expression of $b_y(n)$ (1) is repeated here

$$b_y(n) = \sum_{i=1}^{N_e} \sum_{k=0}^{n} h_i(k) b_i(n-k) \tag{29}$$

[7] gives mean μ_y and variance σ_y^2 of output noise $b_y(n)$ equal to

$$\begin{aligned}
\mu_y &= \sum_{i=1}^{N} \mu_i \sum_{k=0}^{n} E[h_i(k)] \\
\sigma_y^2 &= \sum_{i,j=1}^{N^2} \mu_i \mu_j \sum_{k,m=0}^{n^2} E[h_i(k) h_j(m)] \\
&- \sum_{i,j=1}^{N^2} \mu_i \mu_j \sum_{k,m=0}^{n^2} E[h_i(k)] E[h_j(m)] \\
&+ \sum_{i=1}^{N} \sigma_i^2 \sum_{k=0}^{n} E[h_i^2(k)]
\end{aligned} \tag{30}$$

Using statistical characteristics of quantization noises, the 3^{rd} order moment $E[b_y^3]$ of the output noise is equal to

$$\begin{aligned}
E[b_y^3] &= 3 \sum_{i,j}^{N^2} \mu_i \sigma_j^2 \sum_{k,m=0}^{n^2} E[h_i(k) h_j^2(m)] \\
&+ \sum_{i,j,p} \mu_i \mu_j \mu_p \sum_{k,m,l=0}^{n^3} E[h_i(k) h_j(m) h_p(l)]
\end{aligned} \tag{31}$$

Its skewness value γ_y is obtained applying (10) to (31) and (30) leading to (32) page 6

It can be noted that for a LTI system, γ_y is equal to zero. So, in that case, the output noise PDF is always symmetric.

With the same approach, we can compute the 4^{th} order moment $E[b_y^4]$ of the output noise given by (33) page 6. Given that the skewness of the quantization noises $b_i(n)$ is equal to zero, it does not appear in the expression. Its kurtosis value κ_y is obtained applying (12) to (33) and (30) leading to (34) page 6

The HOM of the output noise $b_y(n)$ are determined analytically. These values allow us to determine completely Generalized Gaussian corresponding to the output noise. In the next section we present some experiments of our model.

IV. RESULTS

In this section, some experiments are carried out to validate our model on FIR and IIR filters and on a 2^{nd} order Volterra filter.

For the FIR filter, we first experiment a filter with 2 coefficients 1.5 and 0.5 where the inputs are quantized on 8 bits and the output on 8 bits with quantization by truncation. Figure 6 shows that our model is accurate. In that case, the PDF of the output noise is not uniform and not gaussian. So our model can estimate this case. On Figure 7, we apply a FIR filter with 16 coefficients where inputs are quantized on 16 bits and the output on 8 bits. Here, the output noise predominates leading to a uniform PDF as we can see in the figure.

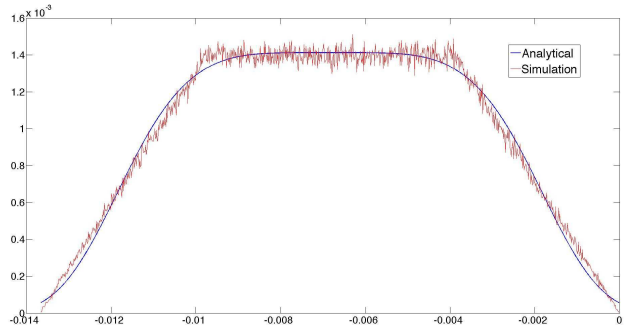

Fig. 6. Output noise of a FIR2 with input on 8 bits

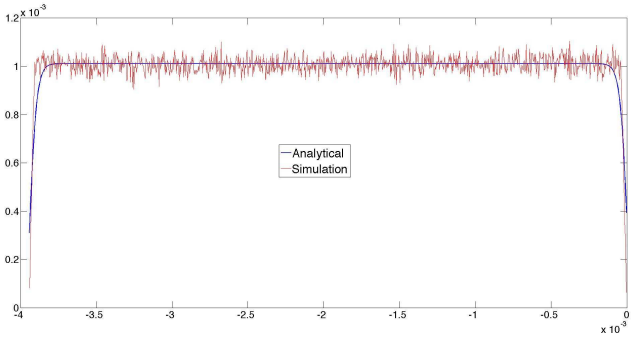

Fig. 7. Output noise of a FIR16 with input on 16 bits and output on 8 bits

Then, we have tested our model on an IIR filter with 2 coefficients where input are quantized on 8 bits and output on 8

$$\begin{aligned}\gamma_y &= 3\sum_{i,j}^{N^2}\mu_i\sigma_j^2\sum_{k,m=0}^{n^2}\Big(E[h_i(k)h_j^2(m)]-E[h_i(k)]E[h_j^2(m)]\Big)\\ &+ \sum_{i,j,p}\mu_i\mu_j\mu_p\sum_{k,m,l=0}^{n^3}\Big(E[h_i(k)h_j(m)h_p(l)]-E[h_i(k)]E[h_j(m)]E[h_p(l)]\Big)\\ &- 3\sum_{i,j,p}\mu_i\mu_j\mu_p\sum_{k,l,m=0}^{n^3}\Big(E[h_i(k)h_j(m)]E[h_p(l)]-E[h_i(k)]E[h_j(m)]E[h_p(l)]\Big) \quad (32)\end{aligned}$$

$$\begin{aligned}E[b_y^4] &= \sum_i\kappa_i\sum_{k=0}^{n}E[h_i^4(k)]+3\sum_{i,j}^{N^2}\sigma_i^2\sigma_j^2\sum_{k,m=0}^{n^2}E[h_i^2(k)h_j^2(m)]\\ &+ 3\sum_i^N\sigma_i^4\sum_{k\neq m=0}^{n^2}E[h_i^2(k)h_j^2(m)]+6\sum_{i,j,p}^{N^3}\mu_i\mu_p\sigma_j^2\sum_{k,m,l=0}^{n^3}E[h_i(k)h_j^2(m)h_p(l)]\\ &+ \sum_{i,j,p,q}\mu_i\mu_j\mu_p\mu_q\sum_{k,m,l,r=0}^{n^4}\sum_{l=0}^{n}E[h_i(k)h_j(m)h_p(l)h_q(r)] \quad (33)\end{aligned}$$

$$\begin{aligned}\kappa_y &= \sum_i\kappa_i\sum_{k=0}^{n}E[h_i^4(k)]+3\sum_{i,j}^{N^2}\sigma_i^2\sigma_j^2\sum_{k,m=0}^{n^2}E[h_i^2(k)h_j^2(m)]\\ &+ 3\sum_i^N\sigma_i^4\sum_{k\neq m=0}^{n^2}E[h_i^2(k)h_j^2(m)]+6\sum_{i,j,p}^{N^2}\mu_i\mu_p\sigma_j^2\sum_{k,m,l=0}^{n^3}\Big(E[h_i(k)h_j^2(m)h_p(l)]-E[h_i(k)]E[h_j^2(m)]E[h_p(l)]\Big)\\ &- 6\sum_{i,j,p,q}\mu_i\mu_j\mu_p\mu_q\sum_{k,m,l,r=0}^{n^4}\Big(E[h_i(k)]E[h_j(m)]E[h_p(l)h_q(r)]-E[h_i(k)]E[h_j(m)]E[h_p(l)]E[h_q(r)]\Big)\\ &- 4\sum_{i,j,p,q}\mu_i\mu_j\mu_p\mu_q\sum_{k,m,l,r=0}^{n^4}\Big(E[h_i(k)]E[h_j(m)h_p(l)h_q(r)]-E[h_i(k)]E[h_j(m)]E[h_p(l)]E[h_q(r)]\Big)\\ &+ \sum_{i,j,p,q}\mu_i\mu_j\mu_p\mu_q\sum_{k,m,l,r=0}^{n^4}\Big(E[h_i(k)h_j(m)h_p(l)h_q(r)]-E[h_i(k)]E[h_j(m)]E[h_p(l)]E[h_q(r)]\Big) \quad (34)\end{aligned}$$

bits with quantization by truncation. The recursion leads to an accumulation of quantization noise which tends to a gaussian as it is shown on Figure 8. Our model is also accurate in that case.

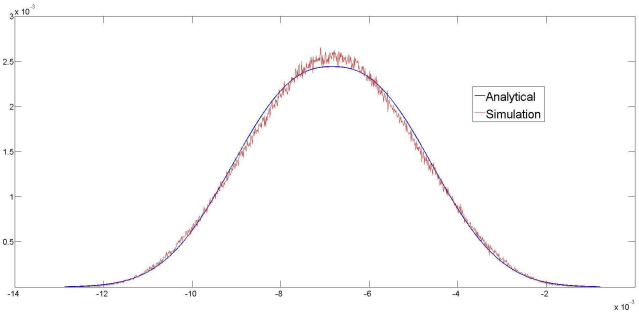

Fig. 8. Output noise of a IIR2 with input on 8 bits and output on 8 bits

Finally, we have applied our model on a 2^{nd} order Volterra filter whose equation is given by

$$\begin{aligned}y(n) &= a_{11}x^2(n)+a_{22}x^2(n-1)\\ &+ a_1x(n)+a_2x(n-1)+a_{21}x(n)x(n-1) \quad (35)\end{aligned}$$

For this experiment, inputs are quantized on 16 bits and the output on 16 bits with quantization by rounding. Figure 9 illustrates our results. The PDF is not really a gaussian (due to non-linearities in the filter) but our model is still accurate. These experiments demonstrates the validity of our model for different output noise PDF and for different quantization types.

V. Conclusion

In this paper, a model to generate the noise on the output of a system due to fixed-point arithmetic is proposed. This model is based on Generalized Gaussian PDF whose parameters are

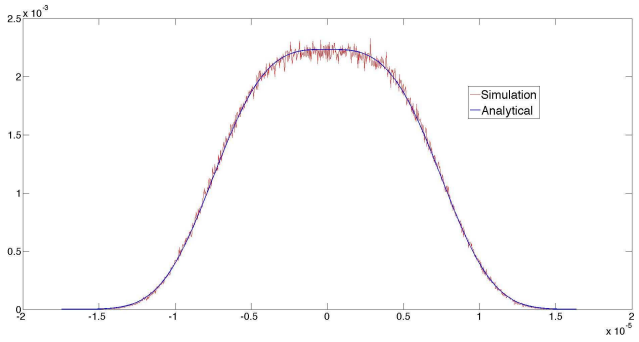

Fig. 9. Output noise of a volterra filter with inputs on 16 bits and output on 16 bits

determined by HOM of the output noise. These HOM are analytically determined which allows to evaluate the fixed-point behaviour of a system without a fixed-point simulation that are really time consuming. Our model has been tested on various applications (FIR filter, IIR filter, Volterra filter) to demonstrate its accuracy. The model will be tried to be extended to asymmetric PDFs.

APPENDIX

In this appendix, the demonstration of the Generalized Gaussian generator is presented. Let X be a random variable with Generalized Gaussian PDF defined as:

$$f_G(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-|\frac{x-\mu}{\alpha}|^\beta} \tag{36}$$

Theorem 1: The generation of samples $x(n)$ following this PDF can be obtained using the following equality

$$x = \mu + u.y^{(1/\beta)} \tag{37}$$

where $u(n)$ are uniform distributed samples on $[-\alpha, \alpha]$, and $y(n)$ are samples following a gamma PDF $\Gamma(1 + 1/\beta, 1)$

Proof:

To demonstrate the theorem, we have to use the three following properties

- The PDF $f_Z(z)$ of $Z = X + c$ where c is a constant and X a random variable is equal to

$$f_Z(z) = f_X(x - c) \tag{38}$$

- The PDF of a random variable $Z = XY$, where X and Y are independent is defined by the following relation.

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|x|} f_X(x) f_Y(z/x) dx \tag{39}$$

- The PDF of $Z = h(X)$ where h is an invertible function is equal to

$$f_Z(z) = \left| \frac{1}{h' o h^{-1}(z)} \right| f_X(h^{-1}(z)) \tag{40}$$

So if $Z = X^a$ where a is a constant

$$f_Z(z) = \left| \frac{1}{a z^{1-1/a}} \right| f_X(z^{1/a}) \tag{41}$$

Samples Y follow a Gamma PDF with coefficients $\Gamma(1 + 1/\beta, 1)$. Their PDF f_Y is equal to

$$f_Y(y) = \frac{1^{(1+1/\beta)}}{\Gamma(1/\beta + 1)} y^{(1/\beta)} e^{-y} = \frac{\beta}{\Gamma(1/\beta)} y^{(1/\beta)} e^{-y} \tag{42}$$

Then applying (41), the PDF $f_V(v)$ of samples $V = Y^{1/\beta}$ is equal to

$$f_V(v) = \frac{\beta^2}{\Gamma(1/\beta)|v^{1-\beta}|} v e^{-v^\beta} \tag{43}$$

Using (39), the PDF of a random variable $Z = U.V = U.Y^{(1/\beta)}$, where U is uniformly distributed on $[-\alpha, \alpha]$ is equal to

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|x|} \frac{\beta^2}{\Gamma(1/\beta)|x^{1-\beta}|} x e^{-x^\beta} \frac{1}{2\alpha} \mathbf{1}_{]|z/\alpha|,\infty]} dx \tag{44}$$

A Gamma PDF is defined on $\mathbb{R}+$. We consider the case where $z > 0$ so

$$\begin{aligned} f_Z(z) &= \int_0^{\infty} \frac{1}{x} \frac{\beta^2}{\Gamma(1/\beta) x^{1-\beta}} x e^{-x^\beta} \frac{1}{2\alpha} \mathbf{1}_{[z/\alpha,\infty]} dx \\ &= \int_{z/\alpha}^{\infty} \frac{\beta^2}{\Gamma(1/\beta)} x^{\beta-1} e^{-x^\beta} \frac{1}{2\alpha} dx \\ &= \frac{\beta^2}{2\alpha\Gamma(1/\beta)} \int_{z/\alpha}^{\infty} x^{\beta-1} e^{-x^\beta} dx \\ &= \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(z/\alpha)^\beta} \end{aligned} \tag{45}$$

Now, we consider the case where $z < 0$. The previous expression is developed as follows

$$\begin{aligned} f_Z(z) &= \int_0^{\infty} \frac{1}{x} \frac{\beta^2}{\Gamma(1/\beta) x^{1-\beta}} x e^{-x^\beta} \frac{1}{2\alpha} \mathbf{1}_{[-z/\alpha,\infty]} dx \\ &= \int_{-z/\alpha}^{\infty} \frac{\beta^2}{\Gamma(1/\beta)} x^{\beta-1} e^{-x^\beta} \frac{1}{2\alpha} dx \\ &= \frac{\beta^2}{2\alpha\Gamma(1/\beta)} \int_{-z/\alpha}^{\infty} x^{\beta-1} e^{-x^\beta} dx \\ &= \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(-z/\alpha)^\beta} \end{aligned} \tag{46}$$

So, in the general case, adding a mean value μ and using (38), it can be written as

$$f_Z(z) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|z-\mu|/\alpha)^\beta} dx \tag{47}$$

■

REFERENCES

[1] P. Belanovic and M. Rupp. Automated Floating-point to Fixed-point Conversion with the fixify Environment. In *IEEE Rapid System Prototyping (RSP'05)*, pages 172–178, Montreal, Canada, 2005.

[2] G. Caffarena, J.A. Lopez C. Carreras, and A. Fernandez. SQNR Estimation of Fixed-Point DSP Algorithms. *EURASIP Journal on Advances in Signal Processing*, 2010.

[3] J.M. Chesneaux, L.S. Didier, and F. Rico. The fixed cadna library. *Real Number and Computers (RNC'03)*, pages 215–229, September 2003.

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

[5] S. Kim, K. Kum, and W. Sung. Fixed-Point Optimization Utility for C and C++ Based Digital Signal Processing Programs. In *Workshop on VLSI and Signal Processing '95*, Osaka, November 1995.

[6] Daniel Menard, David Novo, Romuald Rocher, Francky Catthoor, and Olivier Sentieys. Quantization Mode Opportunities in Fixed-Point System Design. In *18th European Signal Processing Conference (EUSIPCO-2010) (2010)*, pages 542–546, Aalborg Danemark, 08 2010. EURASIP.

[7] R. Rocher, D. Menard, O. Sentieys, and P. Scalart. Analytical approach for numerical accuracy estimation of fixed-point systems based on smooth operations. *Transactions on Circuits and Systems I*, 2012.

[8] 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, October 1977.

[9] A. Tesei and C. S. Regazzoni. The asymmetric generalized gaussian function: A new hos-based model for generic noise pdfs. In *Proceedings of the 8th IEEE Signal Processing Workshop on Statistical Signal and Array Processing (SSAP '96)*, SSAP '96, pages 210–, Washington, DC, USA, 1996. IEEE Computer Society.

[10] A. Tesei and C. S. Regazzoni. Use of fourth-order statistics for non-gaussian noise modelling: The generalized gaussian pdf in terms of kurtosis. 1996.

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