

# Accuracy evaluation of fixed-point based LMS algorithm

Romuald Rocher\*, Daniel Menard, Olivier Sentieys, Pascal Scalart

INRIA/University of Rennes 1, 6 rue de Kerampont, 22300 Lannion, France

## ARTICLE INFO

### Article history:

Available online 24 October 2009

### Keywords:

Adaptive filters

Fixed-point arithmetic

Quantization noise

## ABSTRACT

The implementation of adaptive filters with fixed-point arithmetic requires computation quality evaluation. The accuracy may be determined by computing the global quantization noise power at the system output. In this paper, a new model for evaluating analytically the global noise power in LMS-based algorithms is presented. Thus, the model is developed for LMS and NLMS algorithms. The accuracy of our model is analyzed by simulations.

© 2009 Elsevier Inc. All rights reserved.

## 1. Introduction

Adaptive filters are present in various fields of digital signal processing. They are used in channel or system identification (echo cancellation), in noise reduction, in linear prediction and in equalization to compensate the communication channel distortion. The adaptive filter aim is to estimate a sequence of samples from an observation sequence filtered by a system in which coefficients vary. These coefficients converge toward the optimum coefficients which minimize the mean square error (MSE) between the filtered observation signal and the desired sequence. The different algorithms existing for adaptive filtering are mainly classified in two types corresponding to Least Square Algorithms and Gradient Algorithms. The Least Square Algorithms outperforms the Gradient Algorithms in terms of convergence time and residual mean-square error. Nevertheless, the Gradient Algorithms are the most used in embedded applications because their implementations are more simple than for Least Square Algorithms.

To minimize the cost and the power consumption, digital signal processing applications are implemented in embedded systems with fixed-point arithmetic. However, this simple arithmetic introduces an unavoidable quantization noise when a signal is quantified. These different quantization noise sources are propagated through the system and lead to an output quantization noise. In hardware fixed-point implementation, the goal is to minimize the operator word-length as long as the output quantization noise power is lower than a maximal value. Thus, the analytical expression of the output noise power is the key element in the process of data word-length optimization. In [1], this analytical expression has been integrated in a configurable hardware component generator for LMS algorithm. The operator and memory word-lengths are optimized under an accuracy constraint defined by the user. This analytical approach reduces significantly the noise power evaluation time compared to the techniques based on fixed-point simulation which require very long evaluation time.

Some different models have been proposed to evaluate the output noise power for the LMS algorithm in [2] and [3] and for NLMS algorithm in [4]. These models are presented only for convergent rounding because the quantization noise bias associated with conventional rounding and truncation is not taken into account. The convergent rounding is the more complex quantization mode in terms of hardware implementation and thus it is less often used. The truncation is the most common mode used in embedded systems because its implementation requires no additional hardware. So, the aim of this paper is to propose for the LMS-based algorithms an output noise power expression for all types of quantization laws (truncation, convergent and conventional rounding). The model is presented for LMS, NLMS and Leaky-LMS algorithms.

\* Corresponding author.

E-mail address: rocher@irisa.fr (R. Rocher).

This paper is organized as follows. The basic properties of the LMS algorithm are first recalled in Section 2. Then, the existing models are detailed and their limits discussed. In Section 3, the developed model is explained and the method is clarified. Finally, in Section 4, our model quality is evaluated through different experimentations. This allows to underline its quality and to compare its results with related works.

## 2. Fixed-point gradient-based algorithms

### 2.1. Gradient-based algorithms

The LMS adaptive algorithm addresses the problem of estimating a sequence of scalars  $y_n$  from an  $N$  length vector  $x_n = [x(n), x(n-1), \dots, x(n-N+1)]$  [5]. The linear estimate of  $y_n$  is  $w_n^t x_n$  where  $w_n$  is an  $N$ -length vector which converges to the optimal vector  $w_{opt}$  in the mean-square error (MSE) sense. This optimal vector is equal to  $w_{opt} = R^{-1}p$  where the term  $R$  corresponds to the correlation matrix of the input signal and is equal to  $E[x_n x_n^t]$ . The vector  $p$  represents the intercorrelation between the vector  $x_n$  and the scalar  $y_n$  and is equal to  $E[x_n y_n]$ . The vector  $w_n$  is updated according to the following equation

$$w_{n+1} = w_n + \mu x_n e_n \tag{1}$$

where  $\mu$  is a positive constant representing the adaptation step. The maximum value of  $\mu$  to ensure stability is equal to  $2/\lambda_{max}$  with  $\lambda_{max}$  the maximum eigenvalue of the autocorrelation matrix  $R$  [5]. Another input data representation model has been developed in [6] to study the convergence behavior of the LMS algorithm. According to the value of  $\mu$ , the convergence will be faster or not. In [7], the coefficient probability density function is computed to show the convergence. Moreover, the speed convergence has been studied in [8], for Gaussian input data.

The filter output  $\hat{y}_n$  corresponding to the estimation of  $y_n$  and the error  $e_n$  between the reference signal  $y_n$  and the estimated signal  $\hat{y}_n$  are defined as follows

$$\hat{y}_n = w_n^t x_n \tag{2}$$

$$e_n = y_n - \hat{y}_n \tag{3}$$

To prevent overflow, a new LMS-based algorithm corresponding to the Leaky-LMS has been proposed [9]. In this algorithm, the coefficient update expression is presented in Eq. (4). The leakage factor  $\phi$  is included in the interval  $[0, 1]$ .

$$w_{n+1} = \phi w_n + \mu x_n e_n \tag{4}$$

To assure algorithm convergence, the Normalized LMS (NLMS) has been proposed. In this case, the input signal vector  $x_n$  is normalized by the input data power  $x_n^t x_n$ . This normalization ensures that the adaptation step is included in the interval  $[0, 2]$ . Thus, the coefficient update expression becomes

$$w_{n+1} = w_n + \mu \frac{x_n}{x_n^t x_n} e_n \tag{5}$$

More generally, for these three LMS-based algorithms, the coefficient update expression can be written as follows

$$w_{n+1} = \phi w_n + \mu x_n f(x_n) e_n \tag{6}$$

where the normalization function  $f(x_n)$  is equal to 1 for the LMS and the Leaky-LMS algorithms and to  $\frac{1}{x_n^t x_n}$  for the NLMS algorithm. In the case of the LMS or NLMS algorithm,  $\phi$  is equal to 1.

### 2.2. Fixed-point gradient-based algorithms

The fixed-point gradient-based algorithms are described in this part (Fig. 1). Let  $x'_n$  be the input signal after quantization and  $y'_n$  the quantified desired signal.

$$\begin{aligned} x'_n &= x_n + \alpha_n \\ y'_n &= y_n + \beta_n \end{aligned} \tag{7}$$

The two terms  $\alpha_n$  and  $\beta_n$  are quantization noises. The noise  $\alpha_n$  is an  $N$ -size vector whereas  $\beta_n$  is a scalar. They have means  $m_\alpha$  and  $m_\beta$  and variance  $\sigma_\alpha^2$  and  $\sigma_\beta^2$ .

The filter coefficient vector is written as

$$w'_n = w_n + \rho_n \tag{8}$$

where  $\rho_n$  is the error vector of length  $N$  due to the quantization effects. This noise cannot be considered as the noise due to the quantization of a signal. So the statistical characteristics of white noise cannot be applied to  $\rho_n$ . The error in finite precision is given by

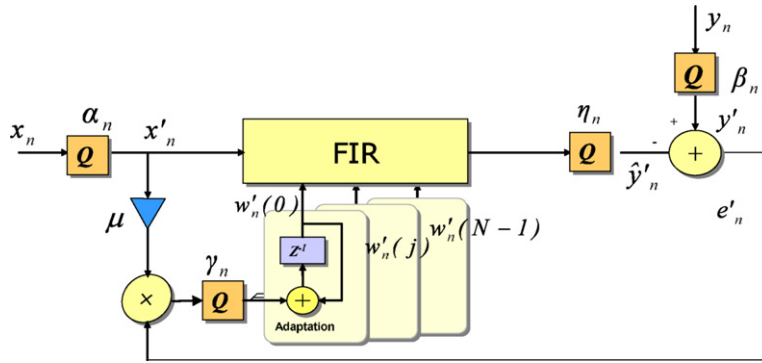


Fig. 1. Gradient-based algorithms.

$$e'_n = y'_n - \hat{y}'_n \quad (9)$$

$$\hat{y}'_n = w_n'^t x'_n + \eta_n \quad (10)$$

with  $\eta_n$  the global noise in the inner product  $w_n'^t x'_n$  and will be developed in the next part.

The updated coefficients expression becomes

$$w'_{n+1} = w'_n + \mu e'_n x'_n + \gamma_n \quad (11)$$

where  $\gamma_n$  is the noise associated with the term  $\mu e'_n x'_n$  and depends on the way the filter is computed. If the Leaky-LMS is under consideration, the coefficient update equation becomes

$$w'_{n+1} = \phi' w'_n + \mu e'_n x'_n + \gamma_n + \psi_n \quad (12)$$

where  $\phi'$  is the fixed-point value of  $\phi$ . If  $\phi$  is a sum of power of two then  $\phi' = \phi$ . More generally, the bias term  $\Delta\phi$  is introduced and is defined as the difference between the fixed-point and real value of the leakage factor  $\phi$ .

$$\Delta\phi = \phi' - \phi \quad (13)$$

Moreover,  $\psi_n$  is the noise generated by the multiplication of  $\phi'$  and  $w'_n$ . In the case of the NLMS algorithm, this equation leads to

$$w'_{n+1} = w'_n + \mu e'_n \frac{x'_n}{x_n'^t x'_n} + \gamma_n \quad (14)$$

To prevent from a division in fixed-point arithmetic, the term  $x_n'^t x'_n$  is approximated by a power of two which greatly simplifies the implementation. Thus, the division is equivalent to a shift of some bits. This method does not introduce a new noise. More generally, the fixed-point coefficient update expression can be written as

$$w'_{n+1} = \phi' w'_n + \mu x'_n f(x'_n) e'_n + \gamma_n + \psi_n \quad (15)$$

The term  $\psi_n$  is equal to zero for algorithms other than Leaky-LMS.

### 2.3. Existing fixed-point LMS models

In this section, the different available models which analyse the fixed-point LMS algorithm are explained. To have a correct behavior, the fixed-point LMS algorithm must assure no overflow as presented in [9]. The LMS operation is made-up of two phases corresponding first to the convergence and then to the steady-state. The modification of the convergence process due to fixed-point arithmetic has been studied in [10] and [11]. The fixed-point arithmetic leads to a slight slowdown. A condition on dynamic to prevent slowdown is proposed in [11]. The LMS output is only exploited at the steady-state when the coefficients have converged towards the optimum coefficients. Thus, the fixed-point effects which decrease the LMS output accuracy are analyzed only at the steady-state after convergence.

Different studies have been done about the fixed-point steady-state analysis of the LMS algorithm. In [3], the expression of the mean square error (MSE) in fixed-point implementation is determined. In that case, the MSE is the second order moment of the difference between the desired signal in infinite precision and the quantified computed output. Thus the MSE is given by the sum of the mean square error in infinite precision and of the noise power which is composed of three terms.

- The error  $\alpha_n$  due to input data quantization filtered by the coefficients  $w_n$  leading to output noise  $\alpha_n^t w_n$ .
- The input sequence  $x_n$  filtered by the deviation  $\rho_n$  of the filter coefficients from their exact values in infinite precision conducting to noise  $\rho_n^t x_n$ .

- The noise  $\eta_n$  inside the filter due to fixed-point arithmetic operations.

The second moment expression of these three terms has been determined. The two first terms are expressed as in the case of linear systems. The last term is more complex. A recurrence is determined on the deviation of the coefficients.

$$\rho_{n+1} = \rho_n + \mu e_n \alpha_n - \mu x_n x_n^t \rho_n - \mu x_n w_n^t \alpha_n + \mu (\beta_n - \eta_n) x_n + \gamma_n \tag{16}$$

Denoting  $\sigma^2$  the quantization noise variance,  $\xi_{\min}$  the minimum MSE, the obtained fixed-point MSE  $\xi'$  expression is given by

$$\begin{aligned} \xi' = \xi_{\min} + \frac{\mu \xi_{\min} \text{Tr}(R)}{2 - \mu \text{Tr}(R)} + \left( \sum_{i=0}^{N-1} w_{\text{opt}_i}^2 + \frac{1}{2} \mu \xi_{\min} N \right) \sigma_{\alpha}^2 \\ + \frac{N \sigma_{\gamma}^2 + \mu^2 (\sigma_{\alpha}^2 \sum_{i=0}^{N-1} w_{\text{opt}_i}^2 + \sigma_{\beta}^2 + \sigma_{\eta}^2) + N \mu^2 \sigma_{\alpha}^2 \xi_{\min}}{2\mu - \mu^2 \text{Tr}(R)} + \sigma_{\eta}^2 \end{aligned} \tag{17}$$

But, few hypotheses are made to reduce the expression complexity. The final result is complex and includes term of second order in  $\mu^2$ .

Moreover, a condition to ensure a correct coefficients convergence is shown in [10]. The study published in [12] uses the same developments but is looking at misadjustment comporment. A study based on [3] has been adapted to the NLMS algorithm in [4]. The fixed-point MSE is determined and a condition to insure convergence is shown.

The model detailed in [2] deals with the MSE like the one before but the method is different. This model also determines the MSE in the case of fixed-point implementation. Only two noises are considered. They correspond to the noise inside the filter due to arithmetic operations and the noise in the multiplication between the input signal and the error  $\mu x_n (y_n - w_n^t x_n)$ . The deviation  $\theta_n$  between the coefficients and their optimum value in the transform domain is introduced  $\theta_n = M[w_{\text{opt}} - w_n]$ , where  $M$  is the matrix defined by  $R = MDM^{-1}$  and  $D$  a diagonal matrix. A recurrence is deduced such as

$$\theta_{n+1} = \theta_n - \mu M x_n e_n' + \gamma_n \tag{18}$$

Then, this recurrence is injected in the equation of the MSE. So, the MSE is determined in the case of finite precision.

$$\xi' = \frac{1}{1 - \frac{\mu}{2} \text{Tr}(R)} \left( \xi_{\min} + \frac{N \sigma_{\gamma}^2}{2\mu} + \sigma_{\eta}^2 \right) \tag{19}$$

This expression leads to the same result as in [3] if the input noise is not considered.

In [13], a model based on quantized energy relation is proposed. With this relation, the fixed-point MSE of different adaptive filters is computed. A result is shown for the LMS algorithm. Nevertheless, this approach is only proposed for white input signal.

These different approaches consider that the noise generated during a cast operation<sup>1</sup> are centered. This assumption is only valid for convergent rounding. These models are not accurate for the other quantization modes.

### 3. General noise power expression

The global noise power expression is detailed in this part. The model is proposed for all quantization laws. Thus, quantization noise statistics are first introduced for the different quantization laws. Then, the output noise expression is presented and its power expression is computed and is applicable to LMS-based algorithms. To determine output noise power expression, different hypotheses will be made.

#### Hypotheses.

- Quantization noises are independent from signal terms and other noises. This comes from quantization noise model.
- Coefficient deviation  $\rho_n$  is supposed uncorrelated with input data  $x_n$ . This hypothesis has been verified by experimentations and is deduced from recursive relation about  $\rho_n$  presented in this part.
- Error  $e_n$  is supposed to have zero-mean and to be uncorrelated with other signal terms. This last hypothesis is generally applied in adaptive filtering domain and has been verified by experimentations.

<sup>1</sup> Reduction of the number of bits.

**Table 1**  
Quantization noise first and second order moment for the three quantization modes.

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

### 3.1. Quantization noise model

A data quantization can be modeled by the sum of the data and a uniformly distributed white noise [14,15]. This white noise (or quantization noise) is uncorrelated with the signal and other noise sources. According to the type of quantization, the noise distribution will differ. Three quantization modes can be considered. It corresponds to truncation, conventional rounding and convergent rounding. In truncation mode the different bits are directly eliminated. The resulting number is always smaller or equal to the number before quantization and thus the quantization noise is always positive. Consequently, the quantization noise mean is not equal to zero. To reduce the bias due to truncation, the rounding quantization mode can be used. In conventional rounding, the data are rounded to the nearest value representable in the reduced-precision format. For numbers located at the midpoint between two consecutive representable values, the data are rounded-up always to the higher output value. This technique leads to a bias for the quantization noise. To eliminate the quantization noise bias, the convergent rounding can be used. In this case, the numbers located at the midpoint between two consecutive representable values are equiprobably rounded to the higher or lower output value.

Let  $n$  be the number of bits for the fractional part after the quantization process and  $k$  the number of bit eliminated during the quantization. The quantization step  $q$  after the quantization is equal  $q = 2^{-n}$ . By following the technique presented in [16] for truncation, the quantization noise mean and variance are presented in Table 1 for the three quantization modes.

### 3.2. Output noise expression

The global noise  $b_y$  represents the noise at the filter output and corresponds to the difference between the fixed-point estimate filter output  $\hat{y}'_n$  and the real estimate filter output  $\hat{y}_n$ .

$$b_y = \hat{y}'_n - \hat{y}_n \quad (20)$$

Introducing Eqs. (10) and (2) in Eq. (20), the global noise  $b_{y_n}$  expression becomes

$$b_y = w_n^t x_n' + \eta_n - w_n^t x_n \quad (21)$$

Using Eqs. (7) and (8) a new expression is obtained where the cross term including the product of noise terms is neglected. Indeed, the expression (22) is obtained

$$b_y = w_n^t \alpha_n + \rho_n^t x_n + \eta_n \quad (22)$$

The interest is in the noise power. The global noise power is given by the two order moment of expression (22)

$$E[b_y^2] = E[(\alpha_n^t w_n)^2] + E[(\rho_n^t x_n)^2] + E[\eta_n^2] + 2E[\alpha_n^t w_n \rho_n^t x_n] + 2E[\alpha_n^t w_n \eta_n] + 2E[\rho_n^t x_n \eta_n] \quad (23)$$

The cross terms can be neglected with respects to other terms, leading to the next expression

$$E[b_y^2] = E[(\alpha_n^t w_n)^2] + E[(\rho_n^t x_n)^2] + E[\eta_n^2] \quad (24)$$

The global noise power is divided into 3 terms which will be developed in the next sections.

#### 3.2.1. Expression of the term $E[(\alpha_n^t w_n)^2]$

The term  $E[(\alpha_n^t w_n)^2]$  is associated to the noise due to the propagation of the input data quantization noise  $\alpha_n$ . This quantization noise vector is not correlated with the coefficient vector  $w_n$ . Thus, the term  $E[(\alpha_n^t w_n)^2]$  is equal to

$$E[(\alpha_n^t w_n)^2] = \text{Tr}(E[\alpha_n \alpha_n^t] E[w_n w_n^t]) \quad (25)$$

Given that  $\alpha_n$  is a white-noise vector, the term  $E[\alpha_n \alpha_n^t]$  can be written as follows

$$E[\alpha_n \alpha_n^t] = \sigma_\alpha^2 I_N + m_\alpha^2 \mathbf{1}_N \quad (26)$$

where  $I_N$  is the  $N$ -size identity matrix and  $\mathbf{1}_N$  the  $N$ -size matrix filled with 1. So, Eq. (27) can be simplified by

$$E[(\alpha_n^t w_n)^2] = \sigma_\alpha^2 \text{Tr}(E[w_n w_n^t]) + m_\alpha^2 \text{Tr}(E[w_n w_n^t] \mathbf{1}_N) \quad (27)$$

At the steady-state,  $w_n$  can be approximated by the optimal coefficients  $w_{\text{opt}}$ . In this case, the term  $E[(\alpha_n^t w_n)^2]$  can be simplified as follows

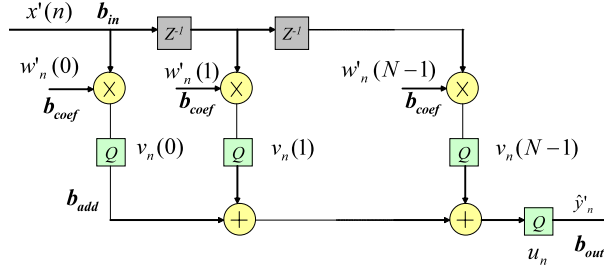


Fig. 2. Filter implementation.

$$\text{Tr}(E[w_n w_n^t]) = \sum_{i=0}^{N-1} w_{\text{opt}_i}^2 \tag{28}$$

$$\text{Tr}(E[w_n w_n^t] 1_N) = \left( \sum_{i=0}^{N-1} w_{\text{opt}_i} \right)^2 \tag{29}$$

The term  $E[(\alpha_n^t w_n)^2]$  can be deduced such as

$$E[(\alpha_n^t w_n)^2] = \sigma_\alpha^2 \sum_{i=0}^{N-1} w_{\text{opt}_i}^2 + m_\alpha^2 \left( \sum_{i=0}^{N-1} w_{\text{opt}_i} \right)^2 \tag{30}$$

The expression (30) shows that this term corresponds to the noise  $\alpha_n$  filtering with a Linear Time Invariant (LTI) FIR where coefficients are equal to  $w_{\text{opt}}$ .

### 3.2.2. Expression of the term $E[\eta_n^2]$

The second term  $E[\eta_n^2]$  depends on the specific implementation chosen for the computation of the filter output (Fig. 2). As all terms  $u_n$  and  $v_n$  are independent, the power of  $\eta_n$  is given by the next expression

$$E[\eta_n^2] = \sigma_u^2 + \sum_{i=0}^{N-1} \sigma_{v(i)}^2 + \left( m_u + \sum_{i=0}^{N-1} m_{v(i)} \right)^2 \tag{31}$$

The noise  $u_n$  are defined by the difference between output bits number  $b_{\text{out}}$  and adders bits number  $b_{\text{add}}$  whereas noises  $v_n$  depends on the difference between adders bits number  $b_{\text{add}}$  and the sum of input data bits number  $b_{\text{in}}$  and coefficients bits number  $b_{\text{coef}}$ .

### 3.2.3. Expression of the term $E[(\rho_n^t x_n)^2]$

The last term  $E[(\rho_n^t x_n)^2]$  is more complex since  $\rho_n$  is not a quantization noise. Subtracting Eqs. (15) and (6), it leads to

$$w'_{n+1} - w_{n+1} = \phi(w'_n - w_n) + \mu e'_n x'_n f(x'_n) - e_n x_n f(x_n) + \Delta\phi w_n + \gamma_n + \psi_n \tag{32}$$

For recall, the finite-precision error is given in Eq. (9). With Eqs. (7) and (10), it leads to

$$\begin{aligned} e'_n &= y'_n - \hat{y}'_n = y_n + \beta_n - w_n^t x'_n - \eta_n \\ &= \underbrace{y_n - w_n^t x_n - \rho_n^t x_n - w_n^t \alpha_n + \beta_n - \eta_n}_{e_n} \end{aligned} \tag{33}$$

With Eqs. (8), (7) and (33), and neglecting the second order noise terms, Eq. (32) becomes

$$\begin{aligned} \rho_{n+1} &= \phi \rho_n + \mu e_n f(x_n) \alpha_n - \mu x_n x'_n f(x_n) \rho_n - \mu x_n f(x_n) w_n^t \alpha_n \\ &\quad + \mu(\beta_n - \eta_n) x_n f(x_n) + \mu e_n x_n f(\alpha_n) + \Delta\phi w_n + \gamma_n + \psi_n \end{aligned} \tag{34}$$

or else

$$\rho_{n+1} = F_n \rho_n + b_n \tag{35}$$

where

$$F_n = \phi I_N - \mu x_n x_n^t f(x_n) \tag{36}$$

and

$$b_n = \mu e_n f(x_n) \alpha_n - \mu x_n f(x_n) w_n^t \alpha_n + \mu (\beta_n - \eta_n) x_n f(x_n) + \mu e_n x_n f(\alpha_n) + \Delta \phi w_n + \gamma_n + \psi_n \quad (37)$$

The term  $f(\alpha_n)$  is equal to zero in the case of the LMS algorithm. If the NLMS algorithm is under consideration, this term represents the input noise  $\alpha_n$  propagation through the normalization term. The noise propagation through the division operator is given in [17] and leads to

$$f(\alpha_n) = -\frac{x_n^t \alpha_n + \alpha_n^t x_n}{(x_n^t x_n)^2} \quad (38)$$

A recursion with the term  $\rho_n$  has been found. This recursion is the same as in [3] and [12]. Introducing the matrix  $P_n = E[\rho_n \rho_n^t]$ , Eq. (39) can be obtained

$$P_{n+1} = E[b_n b_n^t] + E[F_n \rho_n b_n^t] + E[b_n \rho_n^t F_n] + E[F_n \rho_n \rho_n^t F_n] \quad (39)$$

This expression is composed by four terms which are developed in the next paragraphs.

*Computation of the term  $E[b_n b_n^t]$  of  $E[(\rho_n^t x_n)^2]$ .* The term  $E[b_n b_n^t]$  is developed in this section. It represents the autocorrelation matrix of  $b_n$  and can be computed using Eq. (37) and neglecting the cross terms because of their weak power. Moreover, quantization noise model lets us deduce the non-correlation between the noise terms and the other terms. The error  $e_n$  is supposed uncorrelated with other signal terms. This hypothesis has been verified by experimentations. So, the term  $E[b_n b_n^t]$  can be developed as follows

$$\begin{aligned} E[b_n b_n^t] &= \mu^2 E[e_n^2 f(x_n)^2 \alpha_n \alpha_n^t] + \mu^2 E[x_n f(x_n)^2 w_n^t \alpha_n \alpha_n^t w_n x_n^t] + \Delta \phi^2 E[w_n w_n^t] \\ &\quad + E[\gamma_n \gamma_n^t] + E[\psi_n \psi_n^t] + \mu^2 E[x_n^2 f(x_n)^2 (\beta_n^2 + \eta_n^2) x_n^t] + \mu^2 E[e_n^2 x_n f(\alpha_n)^2 x_n^t] \\ &= \mu^2 \xi E[f(x_n)^2] (\sigma_\alpha^2 I_N + m_\alpha^2 1_N) + \mu^2 E[x_n f(x_n)^2 w_n^t (\sigma_\alpha^2 I_N + m_\alpha^2 1_N) w_n x_n^t] \\ &\quad + \Delta \phi^2 E[w_n w_n^t] + (\sigma_\gamma^2 + \sigma_\psi^2) I_N + (m_\gamma^2 + m_\psi^2) 1_N \\ &\quad + \mu^2 E[x_n^2 f(x_n)^2 x_n^t] (\sigma_\eta^2 + \sigma_\beta^2 + m_\eta^2 + m_\beta^2) + \mu^2 \xi E[x_n f(\alpha_n)^2 x_n^t] \end{aligned} \quad (40)$$

where  $\xi = E[e_n^2]$  is the MSE.

*Computation of the terms  $E[F_n \rho_n b_n^t]$  and  $E[b_n \rho_n^t F_n]$  of  $E[(\rho_n^t x_n)^2]$ .* As  $b_n$  is the sum of noise terms, it is independent from the other terms. Moreover,  $F_n$  and  $\rho_n$  are non-correlated. Indeed, from Eq. (35), it is clear that  $\rho_n$  depends on  $F_{n-1}$  but does not depend on  $F_n$ . This hypothesis has been verified by simulations. Thus, with these two hypotheses, the first term  $E[F_n \rho_n b_n^t]$  can be simplified into

$$E[F_n \rho_n b_n^t] = E[F_n] E[\rho_n] E[b_n^t] \quad (41)$$

The computation of Eq. (41) requires the knowledge of  $E[\rho_n]$ . Thus, with average of Eq. (35) used at the steady-state, the term  $E[\rho_n]$  is equal to

$$E[\rho_n] = (I_N - E[F_n])^{-1} E[b_n] \quad (42)$$

The term  $E[b_n]$  is deduced from Eq. (37) in which  $e_n$  has zero-mean in general and the coefficients mean is equal to optimum coefficients

$$E[b_n] = -\mu E[x_n f(x_n) w_n^t] m_\alpha + \mu (m_\beta - m_\eta) E[x_n f(x_n)] + m_\gamma + m_\psi + \Delta \phi w_{\text{opt}} \quad (43)$$

Finally, introducing Eq. (42) in Eq. (41), the next expression is obtained

$$E[F_n \rho_n b_n^t] = E[F_n] (I_N - E[F_n])^{-1} E[b_n] E[b_n^t] \quad (44)$$

Using that  $F_n = \phi I_N - \mu x_n x_n^t f(x_n)$ , it leads to

$$E[F_n \rho_n b_n^t] = (\phi I_N - \mu E[x_n x_n^t f(x_n)]) ((1 - \phi) I_N + \mu E[x_n x_n^t f(x_n)])^{-1} E[b_n] E[b_n^t] \quad (45)$$

To determine  $E[x_n x_n^t f(x_n)]$ , the average principle described in [4] leads to

$$E[x_n x_n^t f(x_n)] = R E[f(x_n)] \quad (46)$$

With the same method, the term  $E[b_n \rho_n^t F_n]$  can be computed with the following expression

$$E[b_n \rho_n^t F_n] = E[b_n] E[b_n^t] (\phi I_N - \mu R E[f(x_n)]) ((1 - \phi) I_N + \mu R E[f(x_n)])^{-1} \quad (47)$$

Computation of the fourth term  $E[F_n \rho_n \rho_n^t F_n]$  of  $E[(\rho_n^t x_n)^2]$ . As  $F_n = \phi I_N - \mu x_n x_n^t f(x_n)$ ,  $E[F_n \rho_n \rho_n^t F_n]$  can be expressed as follows

$$E[F_n \rho_n \rho_n^t F_n] = E[(\phi I_N - \mu RE[f(x_n)]) \rho_n \rho_n^t (\phi I_N - \mu RE[f(x_n)])] \quad (48)$$

Nevertheless,  $\rho_n$  and  $x_n$  are supposed uncorrelated since, from Eq. (35),  $\rho_n$  depends on  $x_{n-1}$  but not on  $x_n$ . So, developing Eq. (48), and using [18] and [19], it leads to

$$E[F_n \rho_n \rho_n^t F_n] = \phi^2 P_n - 2\mu \phi RE[f(x_n)] P_n + \mu^2 RE[f(x_n)^2] \text{tr}(R P_n) \quad (49)$$

The term in  $\mu^2$  is smaller than the two other terms so  $\mu^2 RE[f(x_n)^2] \text{tr}(R P_n)$  can be neglected. Thus,  $E[F_n \rho_n \rho_n^t F_n]$  can be written as

$$E[F_n \rho_n \rho_n^t F_n] = \phi^2 P_n - 2\mu \phi RE[f(x_n)] P_n \quad (50)$$

Computation of  $E[(\rho_n^t x_n)^2]$ . Grouping Eqs. (50), (47), (45) and (40) in Eq. (40), the next expression is obtained

$$P_{n+1} = \phi^2 P_n - 2\mu \phi RE[f(x_n)] P_n + E[b_n b_n^t] + E[b_n] E[b_n^t] (\phi I_N - \mu RE[f(x_n)]) ((1 - \phi) I_N + \mu RE[f(x_n)])^{-1} \\ + (\phi I_N - \mu RE[f(x_n)]) ((1 - \phi) I_N + \mu RE[f(x_n)])^{-1} E[b_n] E[b_n^t] \quad (51)$$

At the steady-state,  $P_{n+1} = P_n$ . So Eq. (51) becomes

$$P_n ((1 - \phi^2) I_N + 2\mu \phi RE[f(x_n)]) = E[b_n b_n^t] + E[b_n] E[b_n^t] (\phi I_N - \mu RE[f(x_n)]) ((1 - \phi) I_N + \mu RE[f(x_n)])^{-1} \\ + (\phi I_N - \mu RE[f(x_n)]) ((1 - \phi) I_N + \mu RE[f(x_n)])^{-1} E[b_n] E[b_n^t] \quad (52)$$

Multiplying by the inverse of  $((1 - \phi^2) I_N + 2\mu \phi RE[f(x_n)])$  and by  $R$ , it leads to

$$R P_n = R ((1 - \phi^2) I_N + 2\mu \phi RE[f(x_n)])^{-1} (E[b_n b_n^t] + E[b_n] E[b_n^t] (\phi I_N - \mu RE[f(x_n)])) \\ \times ((1 - \phi) I_N + \mu RE[f(x_n)])^{-1} (\phi I_N - \mu RE[f(x_n)]) ((1 - \phi) I_N + \mu RE[f(x_n)])^{-1} E[b_n] E[b_n^t] \quad (53)$$

Using the trace operator and its commutativity property, it leads to

$$\text{Tr}(R P_n) = \text{Tr}(R ((1 - \phi^2) I_N + 2\mu \phi RE[f(x_n)])^{-1} (E[b_n b_n^t] + E[b_n] E[b_n^t] (\phi I_N - \mu RE[f(x_n)])) \\ \times ((1 - \phi) I_N + \mu RE[f(x_n)])^{-1} (\phi I_N - \mu RE[f(x_n)]) ((1 - \phi) I_N + \mu RE[f(x_n)])^{-1} E[b_n] E[b_n^t]) \quad (54)$$

Moreover, using non-correlation between  $\rho_n$  and  $x_n$ , the next equation can be written

$$\text{Tr}(R P_n) = E[(\rho_n^t x_n)^2] \quad (55)$$

Using the trace operator properties  $\text{Tr}(AB) = \text{Tr}(BA)$  for matrix  $A$  and  $B$ , the term  $E[(\rho_n^t x_n)^2]$  can be obtained from Eq. (54)

$$E[(\rho_n^t x_n)^2] = \text{Tr}(R ((1 - \phi^2) I_N + 2\mu \phi RE[f(x_n)])^{-1} (E[b_n b_n^t] + 2E[b_n] E[b_n^t] (\phi I_N - \mu RE[f(x_n)])) \\ \times ((1 - \phi) I_N + \mu RE[f(x_n)])^{-1}) \quad (56)$$

This term corresponds to the input signal filtered by the deviation on the coefficients.

### 3.2.4. Global noise power

According to the previous analysis, the global noise power  $P_b$  can be written as

$$P_b = \sigma_\alpha^2 \sum_{i=0}^{N-1} w_{\text{opt}i}^2 + m_\alpha^2 \left( \sum_{i=0}^{N-1} w_{\text{opt}i} \right)^2 + m_\eta^2 + \sigma_\eta^2 + \text{Tr}(R ((1 - \phi^2) I_N + 2\mu \phi RE[f(x_n)]))^{-1} \\ \times (E[b_n b_n^t] + 2E[b_n] E[b_n^t] (\phi I_N - \mu RE[f(x_n)])) ((1 - \phi) I_N + \mu RE[f(x_n)])^{-1}) \quad (57)$$

## 4. Simplified model for LMS, NLMS and Leaky-LMS

### 4.1. Case of the LMS algorithm

In the case of the LMS algorithm, some simplifications can be made:  $\phi = 1$ ,  $f(x_n) = 1$  and  $f(\alpha_n) = 0$ .



Moreover,  $b_n$  can be approximated by  $\gamma_n$  ( $\gamma_n$  is the noise associated with the term  $\mu e_n^t x_n^t$ ). Indeed,  $b_n$  is composed of several terms in which  $\gamma_n$  is the most important since the other terms are products of weak power terms. So, the term  $E[b_n b_n^t]$  is approximated by

$$E[b_n b_n^t] = \sigma_\gamma^2 I_N + m_\gamma^2 1_N \tag{58}$$

With these hypotheses, the global noise power becomes

$$Pb = \sigma_\alpha^2 \sum_{i=0}^{N-1} w_{opt_i}^2 + m_\alpha^2 \left( \sum_{i=0}^{N-1} w_{opt_i} \right)^2 + (m_\eta^2 + \sigma_\eta^2) + m_\gamma^2 \frac{\sum_{i=1}^N \sum_{j=1}^N R_{ij}^{-1}}{\mu^2} + \frac{N(\sigma_\gamma^2 - m_\gamma^2)}{2\mu} \tag{59}$$

This model is presented for quantization by truncation and rounding. In the case of rounding, the means of  $\eta_n$  and  $\gamma_n$  are not equal to zero since they represent the quantization of a discrete signal. From Eq. (57),  $m_\alpha$  is the only term to be equal to zero in rounding quantization.

However, if the implementation is made in convergent rounding quantization, the means of  $\eta_n$  and  $\gamma_n$  are equal to zero leading to

$$Pb = \sigma_\alpha^2 \sum_{i=1}^N w_{opt_i}^2 + \sigma_\eta^2 + \frac{N\sigma_\gamma^2}{2\mu} \tag{60}$$

In that case, the expression is quite similar to the model in [3] and [2] but more tractable (no second order term in  $\mu^2$ ).

#### 4.2. The NLMS algorithm

For the NLMS algorithm, the simplifivative hypotheses can be made:

$$\phi = 1, \quad f(x_n) = \frac{1}{x_n^t x_n} \quad \text{and} \quad f(\alpha_n) = -\frac{x_n^t \alpha_n + \alpha_n^t x_n}{(x_n^t x_n)^2}$$

The term  $E[f(x_n)]$  can be determined using [6]. Let  $\tau_i$  be the correlation coefficient between  $x^2(n)$  and  $x^2(n-i)$  for  $i \in [1 : N-1]$  defined as

$$\tau_i = \frac{E[x^2(n-i)x^2(n)] - E[x^2]^2}{E[x^4] - E^2[x^2]} \tag{61}$$

Using the term  $\tau_i$  and denoting  $r_n = x_n^t x_n$ ,  $E[\frac{1}{r_n}]$  is equal to

$$\begin{aligned} E\left[\frac{1}{r_n}\right] &= \frac{1}{E[r_n]} E\left(\frac{1}{1 - (1 - \frac{r_n}{E[r_n]})}\right) \\ &= \frac{1}{\text{Tr}(R)} \sum_{i=0}^{\infty} E\left(1 - \frac{r_n}{E[r_n]}\right)^i \\ &\approx \frac{\chi}{\text{Tr}(R)} \end{aligned} \tag{62}$$

with

$$\chi = \left( \frac{\psi(N + 2 \sum_{i=1}^{N-1} (N-i)\tau_i) + 2 \sum_{i=1}^{N-1} (N-i)(1-\tau_i)}{N^2} \right)$$

where  $\psi = \frac{E[x^4]}{E[x^2]^2}$  represents the kurtosis of the input signal.

Consequently, the global noise power can be written as

$$\begin{aligned} Pb &= \sigma_\alpha^2 \sum_{i=0}^{N-1} w_{opt_i}^2 + m_\alpha^2 \left( \sum_{i=0}^{N-1} w_{opt_i} \right)^2 + (m_\eta^2 + \sigma_\eta^2) \\ &\quad + \frac{\text{Tr}(R)^2 \text{Tr}(E[b_n]E[b_n^t]R^{-1})}{\chi^2 \mu^2} + \frac{\text{Tr}(R)(\text{Tr}(E[b_n b_n^t]) - E[b_n]E[b_n^t])}{2\mu \chi} \end{aligned} \tag{63}$$

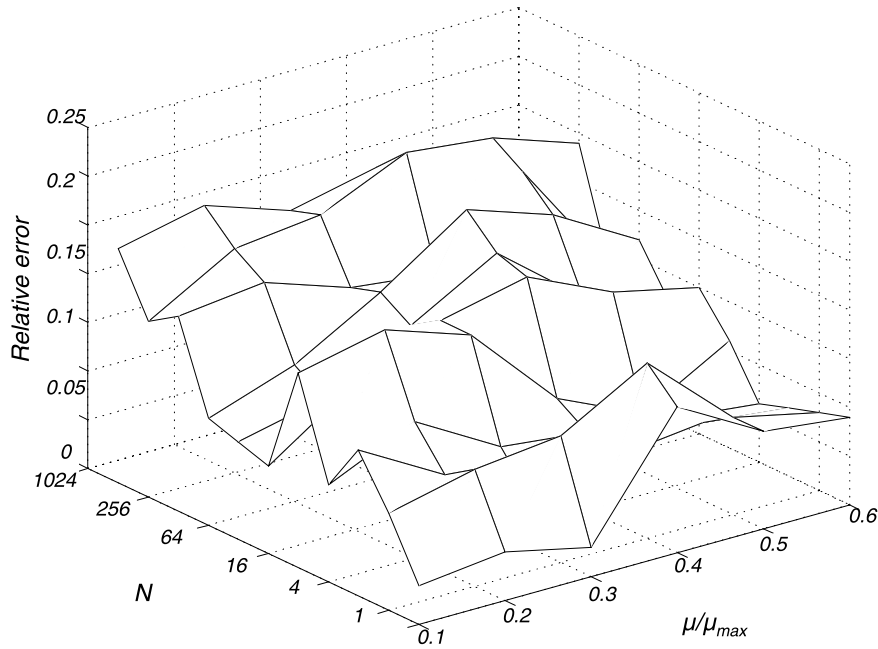


Fig. 3. Relative error for rounding quantization.

## 5. Accuracy estimation quality

In this section, experiments have been conducted to analyse the accuracy of our model for evaluating the fixed-point noise power in LMS algorithms. The input signal chosen is an AR(1) process defined by

$$x(n+1) = \beta x(n) + u(n) \quad (64)$$

where  $u(n)$  is a white noise with zero mean, with variance  $\sigma_u^2$  and  $\beta \in [0, 1[$ . So, the input signal can be very correlated ( $\beta \rightarrow 1$ ) or not ( $\beta \rightarrow 0$ ). For these simulations, tests are made for quantization by truncation and rounding. The relative error between the noise power obtained with simulations and the estimated noise power with our model is computed. For these simulations,  $\mu$  can vary from 0 to  $0.6\mu_{\max}$ . Indeed, the filter coefficients convergence is ensured if  $\mu < \mu_{\max}$ . But, in reality, to be sure that the coefficients do not diverge, a limit of  $0.6\mu_{\max}$  is chosen. However, as  $\mu_{\max}$  depends on the length filter,  $\mu$  is represented by  $\frac{\mu}{\mu_{\max}}$  to be normalized. Moreover, the filter length  $N$  varies from 1 to 1024. The input signal is fairly correlated ( $\beta = 0.5$ ) for the two simulations.

### 5.1. Evaluation of the model accuracy

Fig. 3 shows the relative error between the real and the estimated noise power in rounding quantization. This relative error is smaller than 25% which is a good result since it represents a difference of 1 dB between the output quantization noise power estimated by simulation and the power given by our model. So, this new developed model leads to satisfying results for the case of quantization by rounding.

Fig. 4 represents the relative error in the case of quantization by truncation. As in the rounding quantization case, our model leads to an accurate estimation of the noise power. The relative error is smaller than 20%.

Fig. 5 shows relative error for an LMS algorithm for  $\mu = \frac{\mu_{\max}}{2}$ . Results are presented for different input data correlations and according to LMS size  $N$ . Input data can be uncorrelated ( $\beta = 0.05$ ), fairly correlated ( $\beta = 0.5$ ) or very correlated ( $\beta = 0.95$ ). In all cases, relative error is less than 30% which represents a difference less than 1 significant bit between real noise power and the one estimated by our method.

Fig. 6 illustrates the relative error for the Leaky-LMS algorithm in the case of quantization by truncation. The filter length varies from 1 to 64. The results are very good since the maximum relative error is about 10%.

### 5.2. Comparisons with the other models

Our model has been compared with the two others models presented before [3,2]. For this simulation,  $N$  varies from 1 to 128 and  $\mu$  is fixed at  $\frac{\mu_{\max}}{2}$ . The results are presented in Fig. 7. This test is made in the case of quantization by rounding for which the two other proposed models are presented.

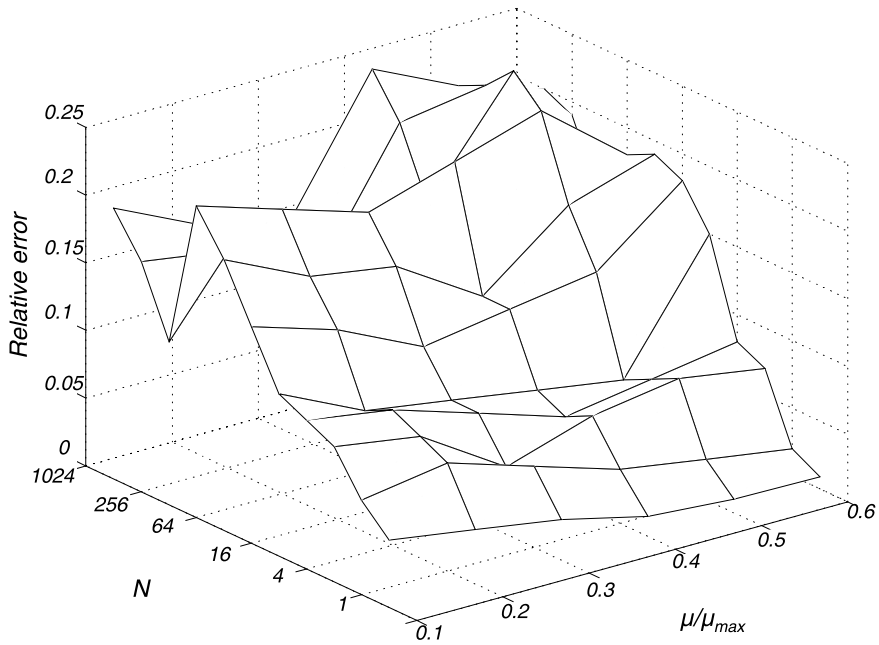


Fig. 4. Relative error for truncation quantization.

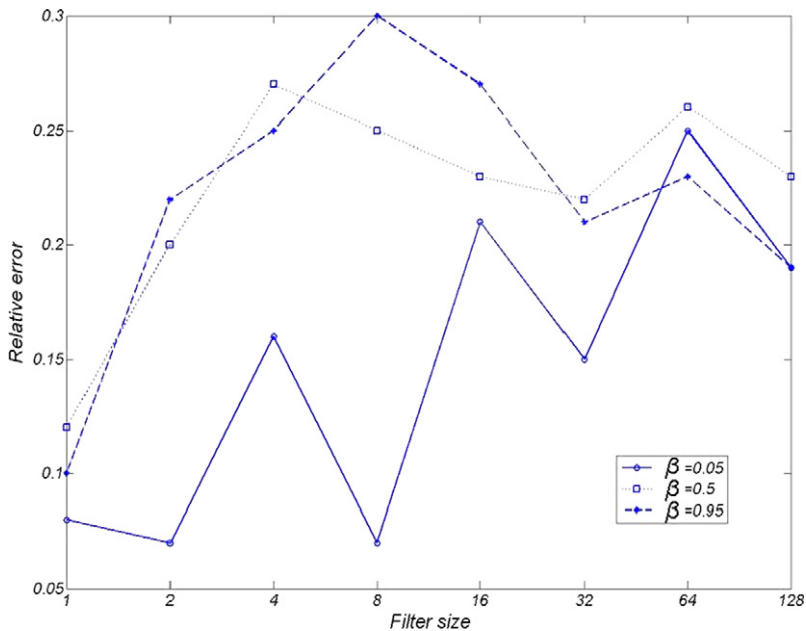


Fig. 5. Relative error for truncation quantization for LMS algorithm for  $\mu = \frac{\mu_{max}}{2}$ .

The model in [2] is the less accurate because it does not integrate the input noise. Our model has better results than the model in [3]. Our simplifications are not prejudicial for the estimation quality. Moreover, the terms we have added in our model let us have a better result. In some cases, the models [2] and [3] are not accurate because they do not integrate the terms  $m_{\eta}^2$ ,  $m_{\gamma}^2 \frac{\sum_{i=1}^N \sum_{k=1}^N (R_{ki}^{-1})}{\mu^2}$  and  $-\frac{Nm_{\gamma}^2}{2\mu}$  for a rounding quantization. In the case of quantization by truncation, these two models lead to a relative error equal to 100%. As a quantization by truncation is the most used law in embedded systems, this result shows the interest of our model.

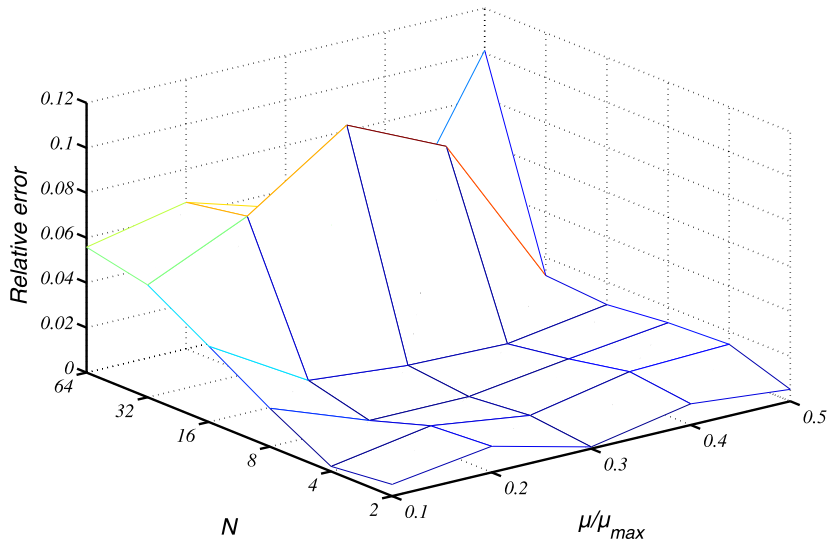


Fig. 6. Relative error for truncation quantization for Leaky-LMS algorithm.

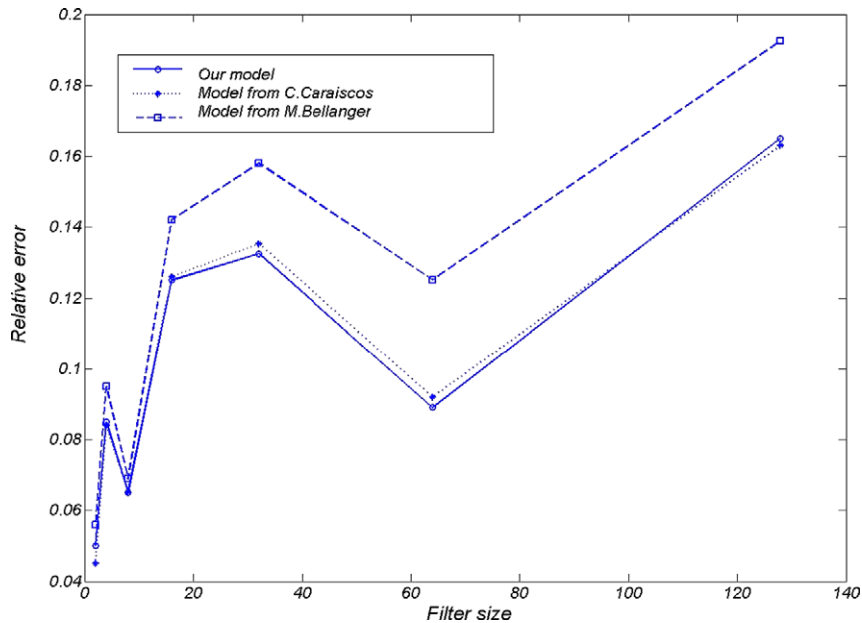


Fig. 7. Comparison between the three models.

## 6. Conclusion

In this paper, a new model for evaluating the noise power in a fixed-point implementation of the LMS-based algorithms is presented. This approach has for main advantage to be more tractable than the models [3] and [2] and to be proposed for all types of quantization. Moreover, this model is more general and can be applied to algorithms such as NLMS or Leaky-LMS. Obtained relative error are less than 25% showing the accuracy of our model. This model lets us define an IPs generator [1], in which a virtual component is optimized in an arithmetic point of view under accuracy constraint.

## References

- [1] R. Rocher, D. Menard, N. Herve, O. Sentieys, Fixed-point configurable hardware components, Article ID 23197, 2006, 13 pp.
- [2] M. Bellanger, Analyse des signaux et filtrage numérique adaptatif, Collection technique et scientifique des télécommunications, vol. 1, 1st edition, Masson, 1989.
- [3] C. Caraiscos, B. Liu, A roundoff error analysis of the LMS adaptive algorithm, IEEE Transactions on Acoustics, Speech and Signal Processing 32 (1) (February 1984) 34–41.

- [4] P.S. Chang, A.N. Willson, A roundoff error analysis of the normalized LMS algorithm, *IEEE Signal, Systems and Computers* 2 (1996) 1337–1341.
- [5] S. Haykin, *Adaptive Filter Theory*, 2nd edition, Prentice Hall Inc., 1991.
- [6] D.T.M. Slock, On the convergence behaviour of the LMS and the normalized LMS algorithms, *IEEE Transactions on Signal Processing* 41 (1993) 2811–2825.
- [7] N.J. Bershad, On the probability density function of the LMS adaptive filter weights, *IEEE Transactions on Acoustics, Speech and Signal Processing* 37 (1) (1989) 43–55.
- [8] A. Feuer, E. Weinstein, Convergence analysis of LMS filters with uncorrelated Gaussian data, *IEEE Transactions on Acoustics, Speech and Signal Processing* 33 (1) (May 1985) 222–230.
- [9] J.M. Cioffi, Limited-precision effects in adaptive filtering, *IEEE Transactions on Circuits Systems* 7 (1987) 821–833.
- [10] J.E. Mazo, R.D. Gitlin, M.G. Taylor, On the design of gradient algorithm for digitally implemented adaptive filters, *IEEE Transactions on Circuit Theory* 2 (1973) 125–136.
- [11] R. Gupta, A.O. Hero, Transient behaviour of fixed-point LMS adaptation, in: *IEEE Conference on Acoustics, Speech and Signal Processing*, 2000, pp. 376–379.
- [12] S.T. Alexander, Transient weight misadjustment properties for the finite precision LMS algorithm, *IEEE Transactions on Acoustics, Speech and Signal Processing* 35 (1987) 1250–1258.
- [13] N.R. Yousef, A.H. Sayed, Fixed-point steady-state analysis of adaptive filters, *International Journal of Adaptive Control and Signal Processing* 17 (2003) 237–258.
- [14] B. Widrow, Statistical analysis of amplitude quantized sampled-data systems, *Trans. AIEE, Part II: Applications and Industry* 79 (1960) 555–568.
- [15] A. Sripad, 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) (Oct. 1977) 442–448.
- [16] G. Constantinides, P. Cheung, W. Luk, Truncation noise in fixed-point SFGs, *IEEE Electronics Letters* 35 (23) (November 1999) 2012–2014.
- [17] D. Menard, R. Rocher, P. Scalart, O. Sentieys, SQNR determination in non-linear and non-recursive fixed-point systems, in: *XII European Signal Processing Conference (EUSIPCO'04)*, Vienna, Austria, September 2004, pp. 1349–1352.
- [18] J.E. Mazo, On the independance theory of equalizer convergence, *Bell Syst. Tech. J.* 58 (1979) 963–993.
- [19] L.L. Horowitz, K.D. Senne, Performance advantage of complex LMS for controlling narrow-band adaptive arrays, *IEEE Transactions on Acoustics, Speech and Signal Processing* 3 (1981) 722–736.

**Romuald Rocher** received the Engineering degree and M.Sc. degree in Electronics and Signal Processing Engineering from ENSSAT, University of Rennes in 2003 and the Ph.D. degree in Signal Processing and Telecommunications from the University of Rennes, in 2006. Since 2008, he is an Assistant Professor at the University of Rennes (IUT Lannion) and a member of the CAIRN (Computing Architectures embedding Reconfigurable resources for eEnergy-efficient systems-on-chip) research team at the IRISA/INRIA Laboratory. His research interests include floating-to-fixed-point conversion and adaptive filters.

**Daniel Menard** received the Engineering degree and M.Sc. degree in Electronics and Signal Processing Engineering from the University of Nantes Polytechnic School in 1996, the Ph.D. degree in Signal Processing and Telecommunications from the University of Rennes, in 2002. From 1996 to 2000, he was a Research Engineer at the University of Rennes. He is currently an Associate Professor of Electrical Engineering at the University of Rennes (ENSSAT) and a member of the CAIRN (Computing Architectures embedding Reconfigurable resources for eEnergy-efficient systems-on-chip) research team at the IRISA/INRIA Laboratory. His research interests include implementation of signal processing and mobile communication applications in embedded systems, floating-to-fixed-point conversion, low power architectures and arithmetic operator design.

**Olivier Sentieys** received the M.Sc. and Ph.D. degrees in Electrical Engineering from the University of Rennes, in 1990 and 1993 respectively. After completing his Habilitation thesis in 1999, he joined the University of Rennes (ENSSAT) and IRISA Laboratory, France, as a full Professor of Electronics Engineering, in 2002. He is leading the CAIRN Research Team at INRIA Institute (national institute for research in computer science and control) and is a Cofounder of Aphycare Technologies, a company developing smart sensors for biomedical applications. His research interests include design of mobile communication systems, finite arithmetic effects, low-power and reconfigurable architectures, and cooperation in mobile systems. He is the author or coauthor of more than 80 journal publications or published conference papers and holds 4 patents.

**Pascal Scalart** received the M.S. degree in signal processing engineering from the University of Rennes, in 1989, and the Ph.D. degree in signal processing and telecommunications from the University of Rennes in 1992. In 1993, he held a Postdoctoral position at Laval University, Québec, Canada, engaging in research on digital signal processing for communications. From February 1994 until March 2003, he was with the Research and Development Center of France Télécom, Lannion, France, where he has been involved in research on speech signal processing for multimedia applications in the field of speech enhancement and adaptive filtering techniques (echo cancellation, noise reduction, beamforming, etc.). In 2001, he joined the Department of Electrical Engineering of the University of Rennes, ENSSAT, France, where he became a professor in 2003. He is a member of the CAIRN research team at the IRISA laboratory. His current research interests are in the area of signal processing for speech processing and for digital communications.