TRADEOFF BETWEEN ROUNDOFF AND OVERFLOW ERRORS IN DIGITAL FILTER REALIZATIONS

José L. Sanz-González

Universidad Politécnica de Madrid (Dpto. SSR)
ETSI de Telecomunicación-UPM, Ciudad Universitaria, 28040 Madrid, Spain.
e-mail: jlsanz@gc.ssr.upm.es

ABSTRACT

This paper is concerned with a linked analysis of overflow and roundoff errors in fixed-point digital filter realizations. Upper bounds for the overflow error power are obtained, having considered saturation quantizer characteristics. Also, upper bounds for the overflow probability are given in order to overflow power be lower than roundoff noise power. Finally, computer simulation results support the theoretical ones, and some of these results are presented in curves for the optimal state-space digital realizations of Butterworth, Chebyshev and elliptic filters.

1. INTRODUCTION

Any digital filter structure can be represented by a computable signal flow graph [1,2], where branch transmittances are constants or unit delays and there are no delay-free loops; on the other hand, nodes represent sequences or their z-transforms. In the signal flow graph, two impulse responses can be associated with the *ith*-node: the sequence $f_i(k)$ is the unit-impulse response from the filter input to the *ith*-node and the sequence $g_i(k)$ is the unit-impulse response from the *ith*-node to the filter output.

The quantization of the operation results in digital filters conveys a nonlinear process. Consequently, a digital filter is a nonlinear system; but, under some mild conditions, it can be modeled by a linear system with error sources in each filter node. The error source $e_i(k)$ at the *ith*-node can be characterized statistically in different ways, depending on the arithmetic to be used.

The error sequence, $\Delta y(k)$, at the filter output can be obtained by

$$\Delta y(k) = \sum_{i=1}^{n} g_i(k) * e_i(k) \qquad (1)$$

where * means convolution operation and n is the number of nodes. The error source $e_i(k)$ has to include the effects of roundoff and overflow. As is well known, if the input sequence is white the roundoff errors are well characterized statistically; on the other hand, overflow errors are not so easy to characterize in the statistical sense.

The output error power is the variance of $\Delta y(k)$ and from (1), assuming that roundoff errors are uncorrelated with overflow errors, we have

$$\sigma_{\Delta y}^2 = \sigma_r^2 + \sigma_{of}^2 \qquad (2)$$

where σ_r^2 is the roundoff error power and σ_{of}^2 is the overflow error power, and both depend on the arithmetic and the quantizer type. Fixed-point arithmetic and quantizers of saturation characteristic will be considered in this paper. This type of quantizer is also used in A/D converters for signal processing [3].

2. THEORETICAL FORMULAS

The *ith*-node error source $e_i(k)$ for saturation quantizers is given by

$$e_i = e_{ri} \cdot u_s(T - |x_i|) + (x_i - T \cdot sgn(x_i)) \cdot u_s(|x_i| - T) \qquad (3)$$

where the first term of (3) is the roundoff error and the second is the overflow error, x_i is the *ith*-node variable and it is really $x_i(k)$, $u_s(\cdot)$ is the unit-step function, $sgn(\cdot)$ is the sign function, i.e. $u_s(x)=(sgn(x)+1)/2$, and $2T$ is the quantizer range.

If $2T$ is the quantizer dynamic range, we define the "overflow probability" P_i of the *ith*-node as follows

$$P_i = Pr\{|x_i| > T\}, \quad i=1,2,...,n \tag{4}$$

From (3) and (2), supposing that $e_{ri}(k)$, $i=1,2,...,n$, are white and uncorrelated sequences, the roundoff noise power σ_r^2 can be expressed as follows

$$\sigma_r^2 = \sum_{i=1}^{n}(1-P_i)\sigma_{ri}^2 \cdot \sum_{k=0}^{\infty}(g_i(k))^2 = \sum_{i=1}^{n}(1-P_i)\sigma_{ri}^2 \cdot \|g_i\|_2^2 \tag{5}$$

where P_i was defined in (4), σ_{ri}^2 is the variance of e_{ri} and $\|g_i\|_2$ is the l_2-norm of $g_i(k)$, i.e. $\|g_i\|_2^2=\sum_k|g_i(k)|^2$

On the other hand, supposing an input Gaussian sequence, it is proved in the Appendix that the overflow error power σ_{of}^2 can be bounded by

$$\sigma_{of}^2 \le 2T^2 \cdot \left(\sum_{i=1}^{n}\|g_i\|_1\right)^2 \cdot \sup_i\left(\frac{P_i}{(T/\sigma_i)^4}\right) \tag{6a}$$

where σ_i^2 is the variance of the *ith*-node variable x_i, and $\|g_i\|_1$ is the l_1-norm of $g_i(k)$, i.e. $\|g_i\|_1=\sum_k|g_i(k)|$.

Also, if the overflow error sequences $e'_i(k)$ $(i=1, 2,...,n)$ are white and uncorrelated with $e'_j(k)$ for $i\neq j$, we have from (A1) and (A4) of the Appendix

$$\sigma_{of}^2 \le 2T^2 \cdot \left(\sum_{i=1}^{n}\|g_i\|_2^2\right) \cdot \sup_i\left(\frac{P_i}{(T/\sigma_i)^4}\right) \tag{6b}$$

The bound (6b) is verified in some structures for low overflow probabilities P_i (say $P_i<10^{-3}$).

For a given filter structure, it is reasonable to adjust the quantizer dynamic range in order that $\sigma_{of}^2 \le \sigma_r^2$; otherwise, overflow errors mask roundoff errors. Then, considering (5) and (6a) or (6b) for $\sigma_{of}^2 < \sigma_r^2$, the following inequalities hold

$$\sup_i\left(\frac{P_i}{(T/\sigma_i)^4}\right) \le \frac{\sum_{i=1}^{n}(1-P_i)\sigma_{ri}^2 \cdot \|g_i\|_2^2}{2T^2 \cdot \left(\sum_{i=1}^{n}\|g_i\|_1\right)^2} \le \sup_i \frac{\sigma_{ri}^2}{2T^2} \tag{7}$$

The second inequality of (7) is a necessary condition for $\sigma_{of}^2 < \sigma_r^2$, and the first inequality is sufficient.

If the sequence samples are coded by b-bits (sign included) in fixed-point representation, assuming double precision accumulators (rounding after summations), we have

$$\sigma_{ri}^2 = \frac{1}{3}\cdot(T\cdot 2^{-b})^2 \tag{8}$$

where $2T\cdot 2^{-b}$ is the quantization step size.

Finally, from (8) and (7), after some manipulations, we can obtain an upper bound for the overflow probability

$$P_i \le \frac{2^{-2(b-1)}}{3}\cdot(b-4)^2 \qquad i=1, 2, ..., n \tag{9}$$

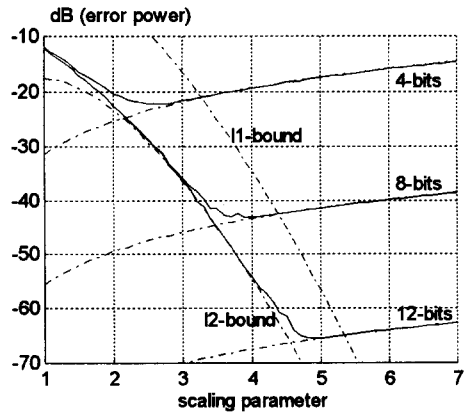
whenever $b>6$.

3. COMPUTER RESULTS

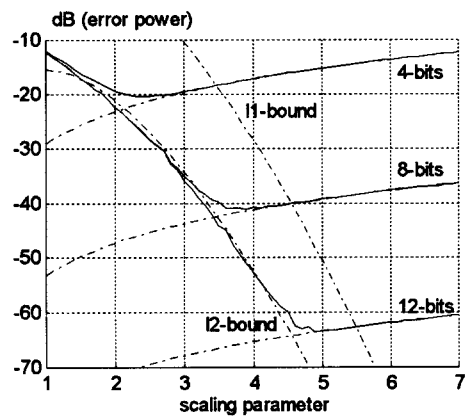
In order to test formulas (6) and (9), some computer simulations were carried out, and some results will be presented in the following.

First, the overflow error power can be estimated from a computer simulation of the filter, whenever the roundoff noise power be vanished. So, in fixed-point arithmetic, using a large number of bits (say 32 bits), the overflow error sequence is obtained by the difference of two output sequences. One sequence corresponding to the filter output with high precision and high dynamic range realization (double precision floating-point), the other output sequence corresponds to high precision fixed-point arithmetic (32-bit fixed-point). The overflow error power is estimated from this error sequence with the appropriate length, in order to obtain a good power estimation. For example, if the scaling parameter δ ($\delta=T/\sigma_i$) is high (low overflow probability), then a large sequence length is required (e.g. if $\delta=5$, we need more than 10^7 samples). Finally, we can plot the overflow error power (σ_{of}^2) versus the scaling parameter (δ) and, of course, a monotonically decreasing curve has to be obtained.

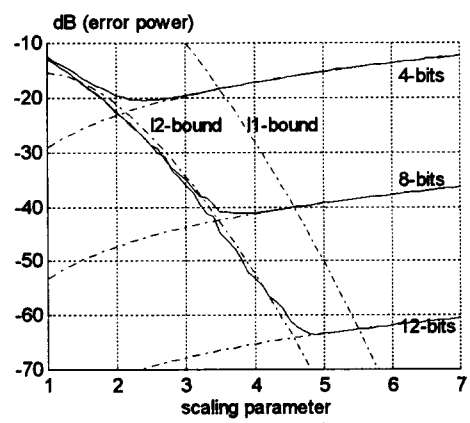
For our computer simulation, we have considered Butterworth, Chebyshev and elliptic filters realized by optimal [1,2,4] state-space structures (minimum roundoff noise structures) with fixed-point arithmetic and saturation-characteristic quantizer. The input sequence is a realization of a white Gaussian process with unit power. The results are presented in curves of σ^2 vs δ, where σ^2 is the error power and $\delta=T/\sigma_i$, and some of them are shown in Figures 1 and 2,



(a) Butterworth filter, order 4

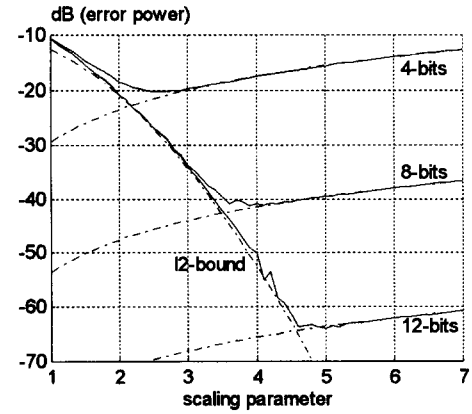


(b) Chebyshev filter, order 4

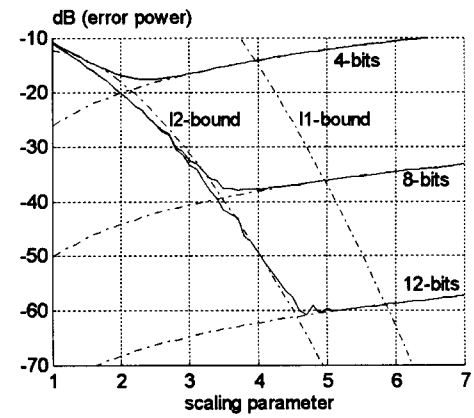


(c) Elliptic filter, order 4

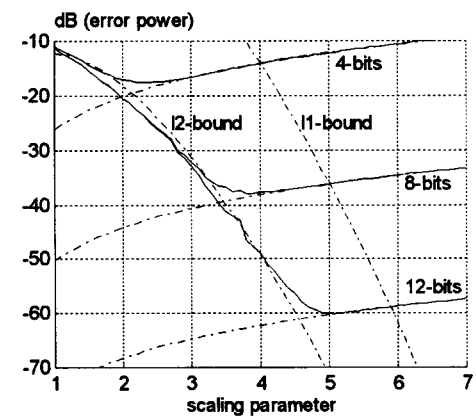
Figure 1. Error Power (σ^2) vs. scaling parameter (δ). Simulation results (continuous lines) and computation results of formulas (discontinuous lines), 4*th*-order lowpass filters.



(a) Butterworth filter, order 8



(b) Chebyshev filter, order 8



(c) Elliptic filter, order 8

Figure 2. Error Power (σ^2) vs. scaling parameter (δ). Simulation results (continuous lines) and computation results of formulas (discontinuous lines), 8*th*-order lowpass filters.

where lowpass filters have been considered with optimal state-space realizations and double precision accumulators (rounding after summations). All filters have the passband cutoff frequencies $\omega_c=0.2\pi$ and the stopband cutoff frequencies $\omega_s=0.3\pi$, and passband ripple of 0.5dB.

In the Figures we plot theoretical curves of σ_r^2 for 4, 8 and 12 bits (discontinuous lines), given by formula (5). Also, theoretical curves of σ_{of}^2 bounds are shown (discontinuous lines), obtained from (6a) for l_1-bound and from (6b) for l_2-bound, where the former is much more conservative than the latter. Finally, experimental curves (continuous lines) of the total error $\sigma_{\Delta y}^2=\sigma_{of}^2+\sigma_r^2$ are given and, of course, for low δ-values $\sigma_{\Delta y}^2\approx\sigma_{of}^2$ and for high δ-values $\sigma_{\Delta y}^2\approx\sigma_r^2$.

As it can be seen from each Figure, a single decreasing curve corresponding to overflow error power σ_{of}^2 was obtained (independent of the number of bits) and it is very tied at the l_2-bound. Also, these experimental results confirm our theory about the dependence of the number of bits and the scaling parameter δ (or overflow probability P_i) required for diminishing the total error power (e.g. 8 bits require $\delta\approx3.5$ and 12 bits require $\delta\approx4.5$), and can be contrasted with expression (9) taking into account that [5]

$$P_i=\sqrt{2/\pi}\cdot\exp(-\delta^2/2)\cdot\left(\frac{1}{\delta}-\frac{1}{\delta^3}+\frac{3}{\delta^5}-\ldots\right)$$

$i=1, 2, \ldots, n$; *for* $\delta>2$

Moreover, the δ-value for minimum total error power $\sigma_{\Delta y}^2$ is approximately independent of filter type and filter order, and only depends on the number of bits.

Finally, it is preferable to overestimate the scaling parameter in order to prevent overflows, although the roundoff noise power be slightly greater; for example, if theoretically it corresponds $\delta=5$, then take $\delta=6$.

APPENDIX

In order to obtain an upper bound for overflow error power, we consider expressions (3) and (1), and from (2), supposing stationarity, we have

$$\sigma_{of}^2=\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}g_i(r)\cdot g_j(s)\cdot E\{e_i'(0)\cdot e_j'(r-s)\} \quad \text{(A1)}$$

where $E\{\cdot\}$ means expectation, $e_i'(k)$ is the overflow error sequence in the *ith*-node and given by the second term of (3), i.e. $e_i'=(x_i-T\cdot sgn(x_i))\cdot u_s(|x_i|-T)$.

By the Schwartz's inequality, we have

$$|E\{e_i'(0)\cdot e_j'(r-s)\}|\le \sup_i E\{(e_i'(0))^2\} \quad \text{(A2)}$$

Consequently, from (A1) and (A2), we have

$$\sigma_{of}^2\le\left(\sum_{i=1}^{n}\sum_{k=0}^{\infty}|g_i(k)|\right)^2\cdot\sup_i E\{(e_i')^2\} \quad \text{(A3)}$$

On the other hand, supposing that the node variable x_i is Gaussian,

$$E\{(e_i')^2\}=\sigma_i^2\cdot\sqrt{2/\pi}\cdot\int_{T/\sigma_i}^{\infty}(x-T/\sigma_i)^2\cdot\exp(-x^2/2)\,dx$$

and integrating by parts, we have for $T/\sigma_i>2$

$$E\{(e_i')^2\}\le 2T^2\cdot\frac{P_i}{(T/\sigma_i)^4} \quad \text{(A4)}$$

Finally, from (A3) and (A4) we obtain (6a).

ACKNOWLEDGMENT.

This work was supported in part by the "Comisión Interministerial de Ciencia y Tecnología" (CICYT) under grant TIC93-0052 of the "Plan Nacional de I+D".

REFERENCES

[1] R. A. Roberts and C.T. Mullis, "Digital Signal Processing." Reading, MA: Addison-Wesley, 1987.

[2] S.K. Mitra and J. F. Kaiser, "Handbook for digital Signal Processing." John Wiley & Sons, NY. 1993.

[3] J.L. Sanz-González, "Nonparametric rank detector on quantized radar video signals". IEEE Trans. Aerosp. Electron. Syst. Vol. AES-26, pp. 969-975, Nov. 1990.

[4] J.L. Sanz-González, F. López-Ferreras and D. Andina, "Roundoff noise results on optimal and block-optimal digital filter structures". Proc. IEEE 1993 Int. Symp. on Circuits and Systems (ISCAS'93), Vol. 1, pp. 611-614, Chicago, Illinois, May 1993.

[5] A.M. Cohen, "Numerical Analysis". McGraw-Hill Book Company (UK) Limited, Maidenhead, Berkshire, England, 1977.