FINITE WORDLENGTH EFFECTS IN DIGITAL FILTERS

G. PÉCELI

Department of Measurement and Instrument Engineering,
Technical University, H-1521 Budapest

Summary

The implementation of digital filters involves the use of finite precision arithmetic. This leads to quantization of the filter coefficients and the results of the arithmetic operations. Such quantization operations are nonlinear and cause a filter response substantially different from the response of the underlying infinite-precision model. This paper intends to give an introductory survey of finite wordlength effects and proposals how to reduce them.

Introduction

In recent years many studies investigated the finite wordlength effects in digital filters [1—2]. This paper, intended for the non-specialist, presents through very simple examples the main problems of the digital filter implementation and some of the methods which may improve the overall filter performance.

A recursive digital filter generates an output, y_n, as follows:

$$y_n = f(x_{n-k}, y_{n-j}); \quad k=0, 1, \ldots, M-1; \quad j=1, 2, \ldots, N-1 \qquad (1)$$

where x_n is the present input. Typically, $f(\ )$ approximates but never exactly duplicates a linear function because of limited precision in real implementations. Moreover, recursively applying the nonlinear function, $f(\ )$ can lead to self-sustained oscillations (limit cycles), the parameters of which depend on the type arithmetic, type of quantization, number of quantizers and the filter structure. Due to the relatively small dynamic range even overflow oscillations may occur. These large-scale limit-cycles depend on the overflow arithmetic applied. Generally saturation arithmetic provides better performance than two's complement arithmetic.

In a nonrecursive digital filter, since no feedback is involved, limit cycles do not disturb the overall behaviour of the filter, however, overflow cannot always be avoided even if the input signal level is relatively low.

Quantization of filter coefficients in any kind of digital filters may be considered as a deterministic change of the filter characteristics which can be minimized by a structure and parameter-value dependent quantization scheme.

A third component of the nonideal behaviour derives from the quantization of the internal arithmetic operations. For example, when fixed-point arithmetic is used, the products are usually quantized to the original wordlength. This leads to a fluctuating error, often called quantization noise.

In the following five sections illustrations are given to show some details of these problems. Section 2 demonstrates the zero input limit cycles which even for a first-order recursive filter, may produce a relatively high error signal amplitude. Section 3 presents the widely used quantization noise approach and the correlated noise problem, generally neglected in usual considerations. Section 4 is devoted to the scaling problem which always leads to a compromise between overflow error probability and dynamic range. Section 5 deals with quantization strategies, while section 6 briefly investigates the digital filter structures.

Obviously, in order to achieve an optimum or nearly optimum filter performance, due to nonlinearity, the above-mentioned effects cannot be treated separately, but should be considered in a common framework. Such a framework, however, does not exist and so there are several open questions to be answered in this field.

Zero input limit cycles

In this section, instead of a detailed analysis, only a simple example is given, the generalization is straight-forward. Let us consider the first-order recursive filter with a single quantizer (Fig. 1). In Figure 1. Q denotes a quantizer for which

$$|\alpha y_{n-1} - Q[\alpha y_{n-1}]| \leq \frac{1}{2} 2^{-B} \tag{2}$$

where B is the wordlength. In the particular case when

$$|y_{n-1}| = |Q[\alpha y_{n-1}]| \tag{3}$$

α looks like if it were ± 1. This may happen if

$$|y_{n-1}| \leq \frac{\frac{1}{2} 2^{-B}}{1-|\alpha|} \tag{4}$$

which implies a parasitic (possibly zero frequency) oscillation whenever y_{n-1} differs from zero. The limit given by expression (4) may be considerably high, thus relatively high amplitude oscillations may occur.

As an example, if $B=7$ and $|\alpha|=\frac{127}{128}\approx 0.992$, then $|y_{n-1}|\leq\frac{1}{2}$. A detailed investigation shows that if sampling frequency is too high relative to the filter bandwidth, then $|\alpha|$ will be close tc the unity. This fact, in conjunction with the above example, also emphasizes that sampling frequency relative to filter bandwidth is a very important design parameter.

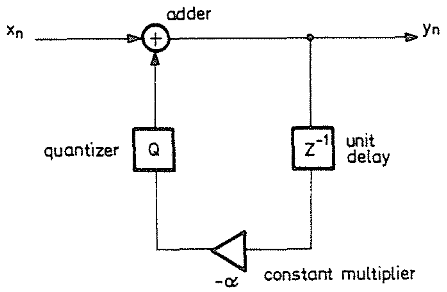

Fig. 1. First-order recursive filter section

The quantization noise model

A great many technical problems can be solved at least to a certain extent if instead of a deterministic approach appropriate noise models are applied. This is the case even considering quantization errors.

In this section the white noise model of a simple second-order section is presented. Consider the second-order recursive filter with two quantizers (Fig. 2).

The uniform quantizer Q with B quantization steps can be modelled, under certain circumstances, by an additive noise source for which we assume that:

- the error (noise) sequence $\{e_n\}$ is a white noise sequence
- the error sequence has a uniform distribution in each quantization interval
- the error sequence $\{e_n\}$ is uncorrelated with the input $\{x_n\}$

If this model is valid, then

$$-\frac{q}{2} < e_n \leq \frac{q}{2}; \qquad q = 2^{-B}; \qquad \sigma_e^2 = \frac{q^2}{12} \tag{5}$$

where σ_e^2 is the noise variance. The output noise variance σ_y^2 can be calculated using the following formula:

$$\sigma_y^2 = \frac{q^2}{12} 2 \frac{1+r^2}{1-r^2} \frac{1}{1+2r^2 \cos \Theta + r^4} \tag{6}$$

where r is the pole distance from the origin ($r<1$), Θ is the pole angle to the real axis.

As an example, if $\Theta = 0$ and $r = \frac{127}{128}$, then $\sigma_y^2 \approx 64 \sigma_e^2$ which may be unacceptable in many applications.

The white noise model is very often used for its simplicity even if the conditions for its application are not always completely fulfilled.

It is rarely considered that the error sources modelled by white noise sequences may be correlated and that significantly modifies the output error variance. Let us investigate the case (Fig. 3) where a signal is weighted by two different constants α_1 and α_2. Using the white noise model, the weighted samples following the quantization can be described by the quantized value and an additive white noise sequence. The two error sequences, $\{e_1\}$ and $\{e_2\}$ however, will be correlated to an extent dependent on the weighting coefficients. By definition the correlation coefficient

$$r_{12} = \frac{E\{e_1 e_2\}}{E\{e_1^2\} E\{e_2^2\}} \tag{7}$$

The coefficients and their ratio can be expressed in the following form

$$\alpha_1 = Jq; \; \alpha_2 = Kq; \qquad R = \frac{\alpha_2}{\alpha_1} = \frac{K}{J} = \frac{[K]_{pJ}}{[J]_{pK}} \tag{8}$$

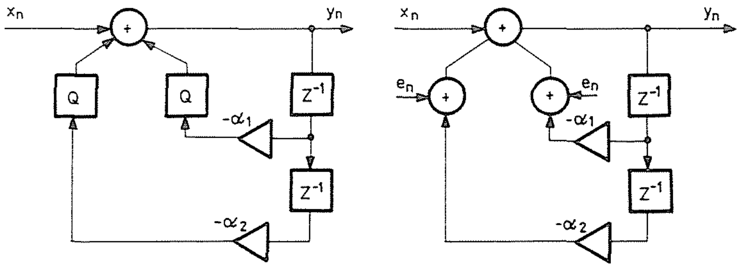

Fig. 2. Block diagram and white noise model of a second-order recursive filter section

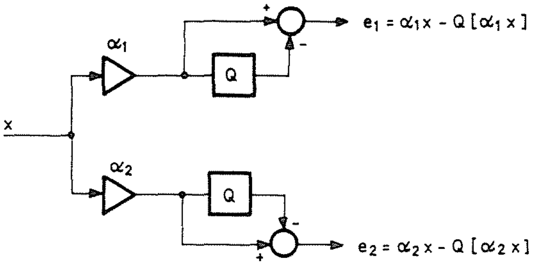

Fig. 3. Correlated quantization noise model

where q is the quantization step, K and J are integers, and $[K]_{pJ}$ denotes the prime factors of K which are not prime factors of J, while $[J]_{pK}$ denotes the prime factors of J which are not prime factors of K.

For the sake of simplicity let us suppose that the input signal was not quantized. In this case

$$r_{12}=\frac{\operatorname{sign}[R]}{[K]_{pJ}[J]_{pK}} \quad \text{if } [K]_{pJ} \text{ and } [J]_{pK} \text{ both are uneven,}$$

$$r_{12}=-\frac{\operatorname{sign}[R]}{2[K]_{pJ}[J]_{pK}} \quad \text{if } [K]_{pJ} \text{ is uneven and } [J]_{pK} \text{ is even}$$

or vice versa. From these expressions the correlation is obviously maximum if $\alpha_1=\alpha_2 \Rightarrow r_{12}=1$, and the correlation is minimum if α_1 and α_2 are relative primes.

As an example, if $\alpha_1=8q$, $\alpha_2=9q$, $r_{12}=-\frac{1}{144}$ and if $\alpha_1=8q$, $\alpha_2=10q$, $r_{12}=-\frac{1}{40}$.

The scaling problem

To keep the probability of internal overflow within acceptable bounds, the dynamic range at certain summing nodes of the filter must be limited. However, the more limited the dynamic range, the greater the effect of roundoff error. The investigations of this interaction between roundoff noise and dynamic range show that the roundoff noise is sensitive to the input signal level, the form of realization and, in case of cascade realizations, sensitive to the way in which terms are grouped in the factorization of the transfer function, as well as sensitive to the sequential ordering of the subfilters.

To find an acceptable input signal level a compromise must be achieved by setting the scale factor of the filter (see Fig. 4), between limited dynamic range and roundoff error. If the form of realization is given, there are three basic suggestions for this scaling problem in the underlying theory. The mathematical background of these methods is the so-called functional analysis which provides effective tools for such and similar problems.

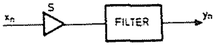

Fig. 4. The scaling problem

l_1 scaling

Let us consider the input/output relationship of the filter using the weighting function method

$$y_n = \sum_{k=0}^{\infty} h_k x_{n-k} \tag{9}$$

where h_k denotes a sample of the discrete weighting function. If y_n should be a limited value, the same is valid for x_n, for every n. If y_n is limited by ± 1 then from the inequality

$$|y_n| \leq \sum_{k=0}^{\infty} |h_k|\,|x_{n-k}| \leq \max|x_n| \sum_{k=0}^{\infty} |h_k| \tag{10}$$

it can be derived that

$$\max|x_n| = \frac{1}{\sum_{k=0}^{\infty} |h_k|} \tag{11}$$

This method always works but it is rather pessimistic and gives unacceptably low x_n levels.

L_∞ scaling

In the frequency domain the following requirement may be fulfilled

$$\max_{-\pi < \omega T < \pi} [SH(z)|_{z=e^{j\omega T}}] < 1 \tag{12}$$

where $H(z)$ denotes the transfer function, S the scale factor, $\omega=2\pi f$ the angular frequency and T the sampling period. This approach is good for sinusoidal input signals, in other cases it is rather optimistic.

L_2 scaling

Based on energy considerations the following method provides a good compromise for the scaling problem. The input/output relationship of the filter in the frequency domain is given by

$$Y(z)=H(z)\,X(z) \tag{13}$$

Using Parseval's formula and Schwartz's inequality

$$\begin{aligned}\sum_{\forall k} y_k^2 &= \int_0^{2\pi} Y(e^{j\omega T})\,Y(e^{-j\omega T})\frac{d\omega T}{2\pi} = \\ &= \int_0^{2\pi} H(e^{j\omega T})\,H(e^{-j\omega T})\,X(e^{j\omega T})\,X(e^{-j\omega T})\frac{d\omega T}{2\pi} \leq \\ &\leq \sum_{\forall k} x_k^2 \int_0^{2\pi} H(e^{j\omega T})\,H(e^{-j\omega T})\,d\omega T\end{aligned} \tag{14}$$

According to this approach the scale factor S should be

$$S=\left[\int_0^{2\pi} H(e^{j\omega T})\,H(e^{-j\omega T})\,d\omega T\right]^{1/2} \tag{15}$$

Quantization strategies

As already stated, the different finite wordlength effects are highly interrelated. The quantization strategy seems to be a key element since an appropriate strategy (together with an appropriate structure) may eliminate the severely nonlinear behaviour of the digital filter.

The first attempt is generally to apply zero memory quantizers because their implementation is much easier than that of any other type. Typical methods are rounding, truncation and "random" rounding where in the usual case a random source selects between rounding and truncation.

A far more effective solution can be obtained if the quantizers have some memory and the so-called controlled rounding can be applied. This approach however, requires considerable additional hardware [4, 5].

But even if we have a highly effective quantization strategy (possibly on chips), we must select an appropriate structure to achieve a satisfactory solution. As an example we mention a very attractive structure, the so-called "coupled loop" one (Fig. 5), which is free of overflow oscillations. In this structure large quantization limit cycles are possible if rounding is applied, while, using truncation, such limit cycles can be avoided [6].

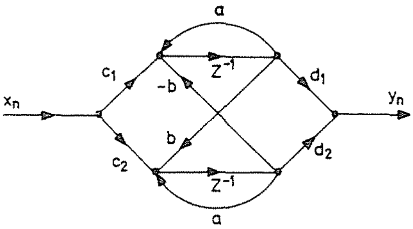

Fig. 5. The coupled loop structure using signal flow graph notation

Filter structures

Any digital filter can be realized in a great number of different structures [1, 8]. These structures, however, differ considerably regarding their accuracy in case of limited wordlength.

Since in the literature there are a lot of publications which investigate structural problems, here only a simple example is given. In Fig. 6 there is a simple second-order section with a transfer function:

$$H(z)=\frac{z^{-2}}{1-m_1 z^{-1}-m_2 z^{-2}}=\frac{z^{-2}}{1-2r\cos\Theta\, z^{-1}+r^2 z^{-2}} \tag{16}$$

where m_1, m_2 are real parameters, r, Θ are the pole location parameters. The parameter sensitivities of the pole location parameters can be expressed as follows

$$S^r_{m_1}=\frac{\partial r}{\partial m_1}=0;\; S^r_{m_2}=\frac{\partial r}{\partial m_2}=-\frac{1}{2r}$$

$$S^{\Theta}_{m_1}=\frac{\partial\Theta}{\partial m_1}=-\frac{1}{2r\sin\Theta};\; S^{\Theta}_{m_2}=\frac{\partial\Theta}{\partial m_2}=-\frac{1}{2r^2\,\mathrm{tg}\,\Theta} \tag{17}$$

These sensitivity values will be considerably high if the pole pair is near the unit circle and the real axis. This means that any quantization error in the parameters will cause a large deviation in the pole location and thus destroy overall filter performance.

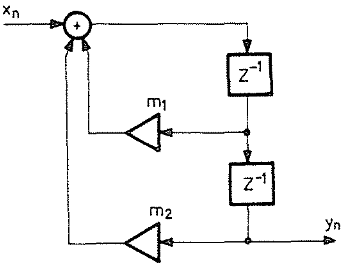

Fig. 6. A simple second-order section

If we introduce free parameters this situation can be changed. Let us introduce a new complex variable, *w* and two real parameters, *c* and *d* and combine the new variable and the parameters in the following manner [7]

$$z=\frac{w}{c}+d \tag{18}$$

If we substitute this variable into the transfer function (16), and fix the introduced parameters in the following way

$$c=\frac{1}{r\sin\Theta}; \qquad d=r\cos\Theta \tag{19}$$

then the pole location sensitivities can be expressed as

$$\begin{aligned} S^{r}_{d}&=\cos\Theta & S^{r}_{c^{-1}}&=\sin\Theta \\ S^{\Theta}_{d}&=-\frac{\sin\Theta}{r} & S^{\Theta}_{c^{-1}}&=\frac{\cos\Theta}{r} \end{aligned} \tag{20}$$

This transformation solves the above sensitivity problem and gives a new structure which is equivalent with the coupled loop structure of Figure 5. In the literature several similar approaches are given [2, 3, 9].

Conclusion

This paper intended to survey briefly the finite wordlength effects in digital filters which, due to the basically nonlinear behaviour, are strongly interrelated and therefore difficult to handle in a common framework.

A great many investigations, provide, however, useful proposals for reducing errors caused by these effects. The importance and effectiveness of these methods and proposals was illustrated by very simple examples.

References

1. RABINER, L. B.—GOLD, B.: Theory and Application of Digital Signal Processing. Prentice Hall, Inc. Englewood Cliffs, N. J. 1975.
2. BARNES, C. W.: Roundoff Noise and Overflow in Normal Digital Filters. IEEE Trans. on Circuits and Systems, Vol. CAS—26, No. 3. March 1979. pp. 154—159.
3. BARNES, C. W.—MIYAWAKI, T.: Roundoff Noise Invariants in Normal Digital Filters. IEEE Trans. on Circuits and Systems, Vol. CAS—29, No. 4, April 1982. pp. 251—256.
4. LAWRENCE, V. B.—LEE, E. A.: Quantization Schemes for Recursive Digital Filters. Proc. of ISCAS 82 (Rome), pp. 690—694.
5. ABU-EL-HAIJA, A. I.—PETERSON, A. M.: An Approach to Eliminate Roundoff Errors in Digital Filters. 1978 IEEE Int. Conf. on Acoustics, Speech and Signal Processing. 78CH1285—6 ASSP pp. 75—78.
6. JACKSON, L. B.: Limit Cycles in State-Space Structures for Digital Filters. IEEE Trans. on Circuits and Systems, Vol. CAS—26, No. 1. January 1979. pp. 67—68.
7. SZCZUPAK, J.—MITRA, S. K.: On Digital Filter Structures with Low Coefficient Sensitivities. Proc. of the IEEE, Vol. 66, No. 9. Sept. 1978. pp. 1082—1083.
8. PÉCELI, G.: Optimum Digital Filter Structures for Measurement Data Processing. ACTA IMEKO 1979. pp. 289—300.
9. MULLIS, C. T.—ROBERTS, R. A.: Synthesis of Minimum Roundoff Noise Fixed-Point Digital Filters. IEEE Trans. on Circuits and Systems, Vol. CAS—23. No. 9, Sept. 1976. pp. 551—562.

Dr. Gábor PÉCELI H-1521 Budapest