Statistical Analysis of Round-off Noise in First-Order IIR DSP Systems

Zlatka Valkova-Jarvis[1], Kamelia Nikolova[1] and Venera Dimitrova[1]

Abstract – **The quantisation of multiplication products in digital signal processing (DSP) recursive systems causes parasitic noise, resulting in deterioration of the system characteristics. The low-sensitivity of DSP systems helps to decrease this noise, thereby improving system performance. This work presents the experimental results of an investigation into the quantisation of the multiplication product at all possible placements of poles for some of the most commonly-used low-sensitivity recursive first-order digital systems. The study presented here enables the areas of best performance of each of the examined DSP systems to be determined.**

Keywords – **quantisation effects, round-off errors, first-order LP and HP DSP networks, all-pass first-order digital systems.**

I. INTRODUCTION

Due to recursion, infinite impulse response (IIR) DSP systems have considerable advantages over finite impulse response (FIR) ones. Recursive systems are much preferred in practice owing to their very high efficiency, lower order, faster signal processing, and improved operation at high sampling rates. Despite the above advantages, IIR digital systems have more complex structures, pose greater challenges in design and investigation, may have linear phase responses in only part of the relevant frequency range, and can display instability.

Most DSP applications in telecommunications, especially real-time processing ones, use fixed-point binary arithmetic and sign-magnitude, one's complement or two's complement binary encoding. Despite the narrow dynamic range, fixed point representation has advantages such as lower cost of hardware, higher speed processing and reduced energy consumption [1].

Ideally, all the discrete signals and parameters in both FIR and IIR DSP systems are of infinite length; hence the digital system will be absolutely quiet with no noise or distortion. In practice, delivering a working recursive DSP system based purely on theoretical design requires the use of a limited number of bits to represent the system's coefficients and processed signals. This process is called *quantisation* and transforms the ideal linear DSP system into a real nonlinear DSP system. Hence, finite word length errors will appear, possibly resulting in a catastrophic impact on the performance of IIR digital systems and thus rendering them inoperative. For example, unexpected oscillations may appear in the output of a nonlinear digital system – these are called limit cycles. They are typically a problem for recursive systems with poles located very close to the limit of stability (the unit circle) [1].

Quantisation can be performed either by *rounding* or by *truncation*, both having significant impact on the accuracy of the presentation of the numbers as well as the characteristics of digital circuits.

Quantisation effects are also called *parasitic* and, for fixed-point digital systems, can be summarised as follows: overflow saturation; arithmetic rounding; coefficient quantisation; data scaling; limit cycling. Numerous methods are used for reducing parasitic effects and they are primarily directed towards one particular problem. However, general corrective techniques are also available and take the form of: scaling the input or coefficients, increasing the system's word length, and selecting an alternative digital structure [1], [2].

In DSP the following fixed-point parameters and signals are quantised by rounding or by truncation:

- The input and output signals;
- Signals resulting from intermediate calculations (also called inner products);
- Multiplier coefficients of the digital system.

The DSP system coefficients are quantised only once, remaining unchanged throughout the entire system operation but also causing the digital system's characteristics to diverge from their ideal form. If the system specifications are no longer met, the quantisation design must be optimised by allocating more bits or choosing a less sensitive realisation.

Quantisation of the results of intermediate calculations is a repetitive process that is reiterated on each cycle of the digital system action. As a result, the output of the system accumulates quantisation errors that worsen the DSP system performance. It is of great importance to assess these errors in order to develop methods and approaches to reduce them [2]. One of the most effective approaches to counteract quantisation errors is the use of very low-sensitivity digital structures, such as cascade or parallel combinations of first- and second-order sections.

In addition to being the building blocks of cascade or parallel DSP structures, first-order sections may also be autonomously operating DSP units. Examples of this are, inter alia, digital integrators, differentiators, averagers, oscillators, DC blockers, maximum-flat group delay adjustable and fixed fractional-delay digital filters [1], [3] digital Hilbert transformers; each performing as frequency selective low-/high-pass or all-pass systems [4].

In this work, classic frequency selective low-pass-cum-high-pass and all-pass first-order low-sensitivity digital systems are studied. The comparative analysis conducted here helps to define the best performance areas of the investigated bilinear DSP systems with respect to the errors due to quantisation of multiplication products. Our analysis allows recommendations to be made for the optimum use of the first-order digital sections which were investigated, these being currently regarded as the best in terms of sensitivity.

[1]The authors are with the Faculty of Telecommunications at Technical University of Sofia, 8 Kl. Ohridski Blvd, Sofia 1000, Bulgaria; e-mails: zvv@tu-sofia.bg, ksi@tu-sofia.bg, vdimitrova@tu-sofia.bg.

II. MULTIPLICATION ROUND-OFF NOISE ANALYSIS – A THEORETICAL BACKGROUND

The nonlinear nature of the process of quantisation makes it difficult to assess the effects of the finite word length on the performance of DSP systems. Hence, the analysis of parasitic effects can only be estimated by making a number of assumptions regarding the quantisation being linear in nature. Statistical assessment of the quantisation errors is a method that, following these assumptions, allows quantisation of the values represented by fixed point to be regarded as a linear process with a given accuracy.

However in such an approximate approach the results should also be treated as approximate.

The statistical assessments of quantisation errors that are most often used are:

- Probability Density Function $P(e)$ (PDF);
- Variance σ_e^2 or its square root standard deviation σ_e of the quantisation error;
- Signal-to-Noise Ratio (SNR).

In order to regard the quantisation of a signal $m(n)$, resulting in the occurrence of the quantisation error $e(n)$, as a linear process, the following assumptions are necessary:

1. The error noise signal $e(n)$ is uncorrelated with the quantised signal $m(n)$.

2. The random discrete sequence of the quantisation error $e(n)$ is an independent process and its samples are uncorrelated with one another.

3. The Probability Density Function $P(e)$ of the quantisation error of each sample of the discrete error sequence $e(n)$ has a normal distribution and constant value in a range having a width equal to the quantisation step size δ .

These stipulations make practical sense if the values of the discrete quantised signal $m(n)$ vary considerably from sample to sample, which can occur when the quantisation step is small.

The inner DSP products result from two arithmetic operations - summation and multiplication – both produce fixed-point results with longer word length; these then have to be quantised, especially in the case of real-time signal processing.

When summing fixed-point binary numbers, any increase will occur at the most significant bit thus possibly resulting in an integer part; the quantisation of this integer will lead to overflow – catastrophic in its effect on recursive digital systems.

In this work, we will investigate the quantisation of an inner signal resulting from intermediate multiplication. The word length of the multiplication product is the sum of those of the multiplication factors. In line with the stipulations listed above, a linear statistical noise model of multiplication product quantisation (Fig. 1) can be developed. Assuming that quantisation by rounding is used, which is most often the case in practice, the model is called the *round-off noise model*.

The signal $m(n)$ and the constant a are quantised to B bits before being multiplied. The discrete sequence of the multiplication $u(n)=am(n)$ is quantised by rounding in order to reduce its word length to B bits. The quantisation device is replaced by the simple linear model shown in Fig. 1, under which the quantised multiplication $\hat{u}(n)$ will be the result of the sum of non-quantised multiplication $u(n)$ and the quantisation error $e(n)$ (more commonly called *quantisation noise*, because it really does have the characteristics of noise). The generated parasitic noise signal $e(n)$ is injected at a node after each multiplication quantisation in the digital structure and is referred to as *round-off noise* [2].

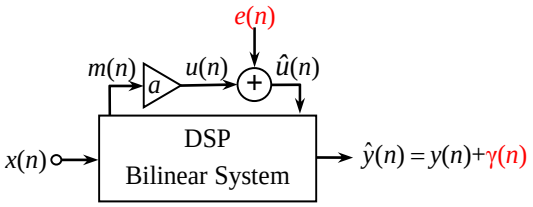


Fig.1: Linear statistical noise model of multiplication product quantisation by rounding

The statistical quantisation noise model is regarded as a linear process when the samples of the signal $m(n)$ have fast changing values. In this case, the quantisation noise signal samples are uncorrelated with each other, i.e. $e_i(n)$ and $e_i(k)$ are statistically independent for $n \neq k$. Hence, the round-off noise $e(n)$ can be modelled as stationary white noise, uniformly distributed in the relevant interval. Since one multiplication is performed in the first-order DSP systems considered here, only one round-off noise source will appear. This error source develops a white noise $\gamma(n)$ at the output of the bilinear digital system, which is added to the output signal $y(n)$, resulting in a output signal mixed with noise $\hat{y}(n)$ (Fig.1).

When a fixed-point multiplication product is encoded by any of the three binary codes (sign-magnitude, one's complement or two's complement), round-off error will be normally distributed in the interval $[-\delta/2 \div \delta/2]$. When the quantisation error is normalised to the quantisation step size: $e_n=e/\delta$, the limits of the intervals of uniform distribution change. For the normalised round-off error the interval will be $[-1/2 \div 1/2]$.

The method that is used most frequently for a statistical representation of the quantisation error is determining variance σ_e^2 or its square root standard deviation σ_e. In the case of fixed-point representation, the uniformly distributed round-off noise $e(n)$ has zero mean value and the variance σ_e^2 [1]:

$$\sigma_e^2 = \frac{\delta^2}{12} = \frac{2^{-2B}}{12}. \qquad (1)$$

B is the word length in bits (without the sign bit). The quantisation step size δ (step of sampling) is the maximum value of the modulus of the error for fixed-point fractional numbers.

The variance of the noise that reaches the bilinear digital system output as a result of the multiplication quantisation has a steady-state (nominal) value σ_γ^2 given by:

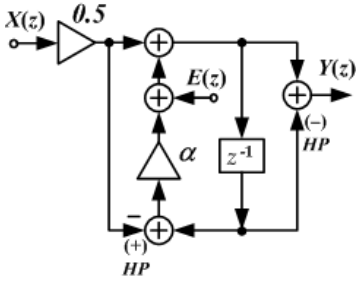
$$\sigma_\gamma^2 = \sigma_e^2 \frac{1}{2\pi j} \oint G(z) G\left(z^{-1}\right) z^{-1} dz = \sigma_e^2 \, \sigma_{\gamma,n}^2 , \qquad (2)$$

where $G(z)$ is defined as noise transfer function (NTF), i.e. the transfer function from the round-off noise source $E(z)$ to the digital system output $Y(z)$. Clearly, the NTF depends on the structure of the digital filter. $\sigma^2_{\gamma,n}$ denotes the noise gain and is also called normalised output noise variance [2], [5].

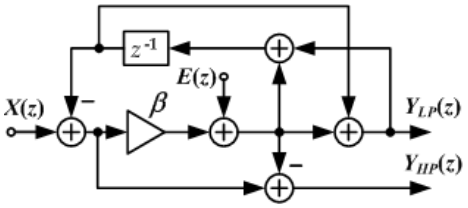
III. NOISE MODELS OF FIRST-ORDER LOW-SENSITIVITY DSP SYSTEMS

As discussed, the use of very low-sensitivity digital structures is one of the most effective approaches to ensure successful DSP system performance in real-world conditions and to counteract the effects of finite word length. Different digital structures do not show the same sensitivity to coefficient accuracy and thus quantisation of coefficients results in unacceptable distortions in frequency response. In this work we investigate low-sensitivity bilinear digital structures with respect to the multiplication round-off noise for all possible pole locations.

In Fig. 2, graphs of signal flow due to multiplication product quantisation in universal LP/HP first-order MHNS (Fig. 2a) and LS1b (Fig. 2b) digital systems are shown [3].



(a)



(b)

Fig.2: Statistical round-off multiplication noise model of bilinear LP/HP DSP systems (a) MHNS and (b) LS1b

Since round-off noise models contain one noise source $E(z)$, the noise transfer functions of bilinear LP/HP MHNS and LS1b systems are as follows:

$$G_{MHNS}^{LP}(z)=\frac{Y(z)}{E(z)}=\frac{1+z^{-1}}{1-\alpha z^{-1}}; \quad (3)$$

$$G_{LS1b}^{LP}(z)=\frac{Y_{LP}(z)}{E(z)}=\frac{1+z^{-1}}{1-(1-2\beta)z^{-1}}; \quad (4)$$

$$G_{MHNS}^{HP}(z)=\frac{Y(z)}{E(z)}=\frac{1-z^{-1}}{1-\alpha z^{-1}}; \quad (5)$$

$$G_{LS1b}^{HP}(z)=\frac{Y_{HP}(z)}{E(z)}=\frac{1-z^{-1}}{1-(1-2\beta)z^{-1}}. \quad (6)$$

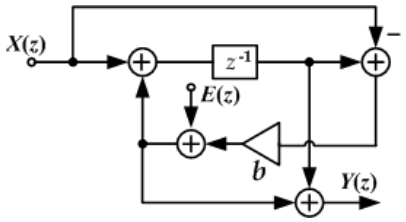
Statistical round-off noise models for the three most low-sensitivity bilinear digital systems are depicted in Fig. 3 [4].

The all-pass bilinear sections (Fig. 2) noise transfer functions are as follows:

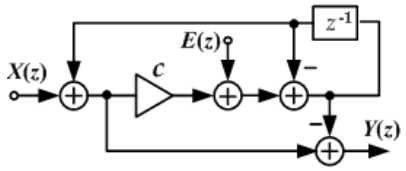
$$G_{MH1}(z)=\frac{Y(z)}{E(z)}=\frac{1+z^{-1}}{1-bz^{-1}}; \quad (7)$$

$$G_{SV}(z)=\frac{Y(z)}{E(z)}=\frac{-1+z^{-1}}{1+(1-c)z^{-1}}; \quad (8)$$

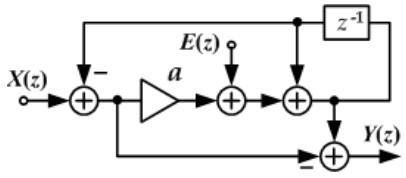
$$G_{ST1}(z)=\frac{Y(z)}{E(z)}=\frac{1+z^{-1}}{1-(1-a)z^{-1}}. \quad (9)$$



(a)



(b)



(c)

Fig.3: Statistical round-off multiplication noise model of bilinear All-pass DSP systems (a) MH1, (b) SV and (c) ST1

IV. EXPERIMENTAL RESULTS

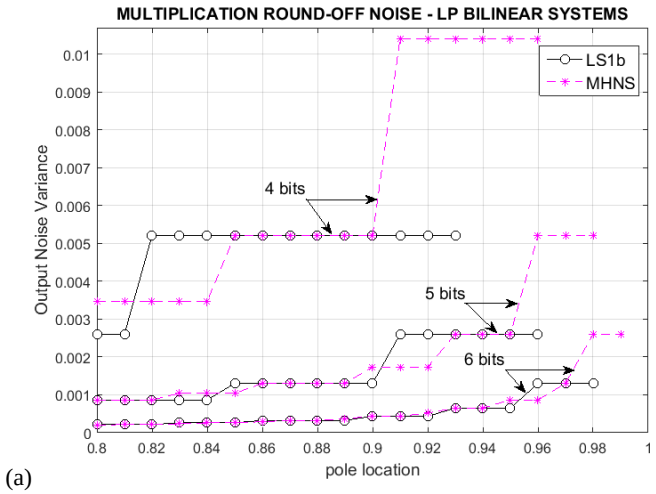
A. Low-pass / High-pass bilinear sections

The universal LP/HP MHNS and LS1b are limit-cycle-free DSP systems, stable when the coefficients are $\alpha\in(-1\div1)$ and $\beta\in(1\div0)$. As per Eq. (2), the nominal round-off noise variances for LP and HP MHNS and LS1b systems (Eqs. (3), (4), (5) and (6)) are respectively as follows:

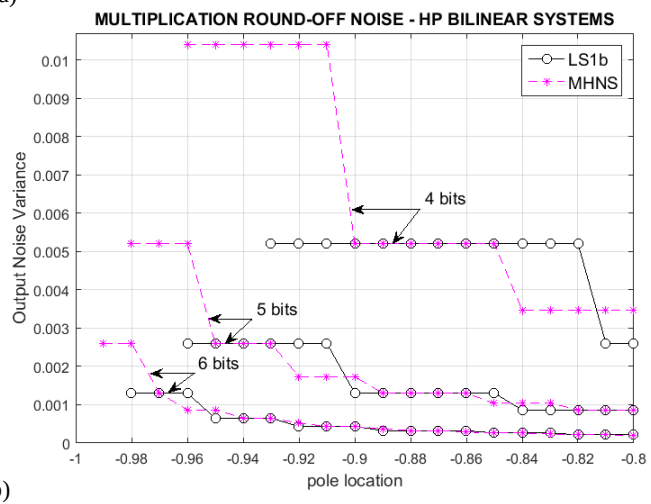
$$\sigma^2_{LP\ MHNS}=\sigma_e^2\frac{2}{1-\alpha}; \quad \sigma^2_{LP\ LS1b}=\sigma_e^2\frac{1}{\beta}; \quad (10)$$

$$\sigma^2_{HP\ MHNS}=\sigma_e^2\frac{2}{1+\alpha}; \quad \sigma^2_{HP\ LS1b}=\sigma_e^2\frac{1}{1-\beta}. \quad (11)$$

They are experimentally explored when the bilinear system fixed-point coefficient and the multiplication inner product are quantised by rounding to very short word length - 4, 5 and 6 bits, including the sign bit. Experimental results for round-off variance versus pole location are shown in Fig. 4a - for LP outputs (Eq. (10)), and in Fig. 4b - for HP outputs (Eq. (11)). The experiments are conducted on the pass-band frequency range, which is the essential one. One expected result is that the shorter the word length, the higher the value of the round-off variance. Due to its lower pass-band sensitivity [3], the LS1b bilinear system demonstrates lower multiplication round-off variance in comparison to the MHNS system for both LP and HP outputs. The shorter the word length, the bigger is the difference between these two sections. When the round-off quantisation removes more bits, the level of the round-off variance remains unchanged for larger pole location intervals. The variance levels change less frequently in the case of the low-sensitivity LS1b system.



(a)



(b)

Fig. 4: Output round-off noise variance for bilinear (a) LP and (b) HP MHNS and LS1b DSP systems for different word lengths.

B. All-pass bilinear sections

Analytical expressions of output round-off noise variance for the bilinear all-pass systems in Fig. 3 (Eqs. (7), (8) and (9)) are likewise calculated:

$$\sigma^2_{MH1} = \sigma^2_e \frac{2}{1-b}; \quad \sigma^2_{SV} = \sigma^2_e \frac{2}{c}; \quad \sigma^2_{ST1} = \sigma^2_e \frac{2}{a}. \qquad (12)$$

Despite the word length, each all-pass system shows the best result for the pole location interval where it is most used in practice (Fig. 5). The MH1 bilinear system has low round-off variance for the pole located in (-0.5÷0.5). Actually, in the last third of this interval, the SV system achieves a lower variance value, but this bilinear section is primarily useful when the pole is in (0.5÷1) where the SV section's variance is anyway the lowest. The ST1 all-pass system performs best when the pole is in the interval (-1÷-0.5), which corresponds to its existing implementation.

V. Conclusion

Regardless of their undoubted advantages, IIR digital systems suffer from parasitic effects due to quantisation of multiplier coefficients, signals and inner products of digital processing. Since the low-sensitivity of DSP systems provides better resistance to parasitic effects, this work investigates the round-off noise of the lowest sensitivity bilinear digital LP, HP and all-pass systems. Relevant assumptions are made about the source of quantisation noise that make the described linear noise model possible. The experiments performed confirm the lower the sensitivity the better the behaviour of the system in the frequency range which is important for the system.

The experimental work conducted here can be effectively employed for DSP systems of other types and orders, which will be helpful when choosing the best system for a particular application.

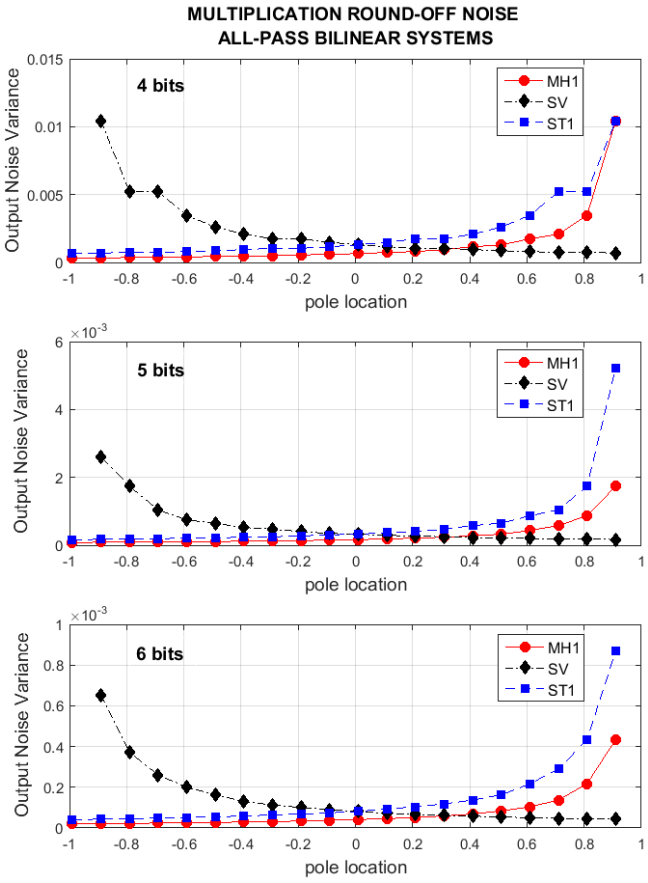


Fig. 5: Output round-off noise variance for bilinear all-pass MH1, SV and ST1 DSP systems for different word length

References

[1] V. Udayashankara, *Modern Digital Signal Processing*, PHI Learning Pvt. Ltd., 2012.

[2] S. K. Mitra, K. Hirano and H. Sakaguchi, "A simple method of computing the input quantization and multiplication roundoff errors in a digital filter**",** IEEE Trans. on Acoustics, Speech and Signal Processing, vol. ASSP-22, No. 5, pp. 326-329, Oct. 1974.

[3] G. Stoyanov, M. Kawamata Z. Valkova, "New first- and second-order very low-sensitivity bandpass/bandstop complex digital filter sections" Proc. IEEE Region 10th Annual Conf. "TENCON'97", Brisbane, Australia, vol. 1, pp. 61-64, 1997.

[4] G. Stoyanov, and K. Nikolova, „Improved accuracy and low-sensitivity design of digital allpass based Hilbert Transformers," Proc. Int. Conf. TELSIKS'2013, Nish, Serbia, 2013, pp. 51-60.

[5] Nikolova Z., D. Romanska, "Multiplication Products Quantization Noise Analysis for Orthogonal Complex IIR Digital Filters", ICEST'08, Nish, Serbia, pp. 540 – 543, 25-27 June 2008.