ELIMINATION OF LIMIT CYCLES DUE TO TWO'S COMPLEMENT QUANTIZATION IN NORMAL FORM DIGITAL FILTERS

Guo Fang Xu[1] *Tamal Bose*[1] *Jim Schroeder*[2] *

[1]Department of Electrical Engineering
University of Colorado
Denver, Colorado 80217-3364, U.S.A.
[2]Cooperative Research Center for Sensor Signal & Information Processing
University of South Australia
Technology Park Adelaide, The Levels 5095 SA, AUSTRALIA

ABSTRACT

Normal form digital filters are attractive due to their desirable properties when implemented in finite wordlength arithmetic. These filters are free from all overflow limit cycles and quantization limit cycles when magnitude truncation is used. However, when two's complement truncation (TCT) quantization is used, limit cycles can still exist. In this paper, it is shown that when block structures are used, normal form digital filters can be made free of limit cycles due to TCT quantization. It is shown that this can be done with a small block size. An algorithm is also presented to find the minimum block size required for a given filter. Some examples are given to illustrate the results.

1. INTRODUCTION

When digital filters are realized in processors, finite wordlength effects are inevitable. These effects are nonlinear in nature and make the filters susceptible to limit cycles. Limit cycles can arise in digital filters due to the effects of overflow or quantization. Both of these effects have received widespread attention in the literature. There are different kinds of overflow and quantization. In this paper, we focus on two's complement truncation quantization in normal form digital filters.

The effects of magnitude truncation (MT) and roundoff (R) type of quantizers have been extensively studied. A relatively very small number of papers have appeared on the case of two's complement truncation (TCT) quantization. As pointed out in [1], the shortage of contributions may perhaps be attributed to the peculiarities of the related nonlinear characteristic $Q\{x\}$, which is not suited to the direct application of such inequalities as $Q\{x\} \leq k|x|$, which yield significant results in all other cases. TCT quantization is a common occurence in digital filter implementation and has been studied to some extent. Direct form digital filters with TCT quantization have been analyzed. The existence areas for limit cycles in the parameter plane and the maximum amplitude of these limit cycles have been studied in [2] and [3], respectively. In [1], the global asymptotic stability (g.a.s) regions have been found for second-order direct form filters implemented in processors with single length accumulators. Some general properties of limit cycles due to TCT quantization were established in [4]. More recently, g.a.s. regions for double length accumulator digital filters were obtained in [5], [6] and [7]. In these references, both direct form and normal form filters were considered. The g.a.s. regions found in these papers are restricted to rather small regions in the parameter plane. The motivation for this paper is to use block realizations so that any linearly stable (stable without quantizer(s)) normal form filter implemented with TCT quantization would be free of limit cycles.

The idea of using block realization to suppress limit cycles in fixed-point digital filters was suggested in [8], [9], and then quantitatively analyzed in [10]. It was shown in 1980, [10] that block filters realized in normal form do not have any overflow limit cycles. Also, with suitably long block lengths, roundoff limit cycles can be prevented. However, that paper does not give any results on TCT limit cycles in normal form block filters. The reason being that the g.a.s. regions for normal form filters under TCT quantization were not established until very recently.

In this paper, it is shown that limit cycles due to TCT quantization can be suppressed in block normal form filters. The organization of the paper is as follows. In section 2, normal form digital filters are briefly described along with the TCT quantizer characteristics. Some exisiting results on TCT quantization are also stated. The main results of the paper are presented in section 3. Some illustrative examples are given in section 4. Section 5 is the conclusion.

2. NORMAL FORM FILTERS UNDER TCT

Consider a digital filter with state and output equations

$$\begin{bmatrix} x(k+1) \\ y(k) \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x(k) \\ u(k) \end{bmatrix} \qquad (1)$$

where $u(k)$, $y(k)$ are scalars representing the input and output respectively, $x(k)$ is $n \times 1$ state vector, A, B, C, D are $n \times n$, $n \times 1$, $1 \times n$, and 1×1 matrices, respectively. With zero input and quantization, the filter satisfies the state equation

$$x(n+1) = Q\{Ax(n)\}$$

where $Q\{.\}$ represents the TCT quantization operation. For a $n \times 1$ vector x, the quantizer is assumed to operate on each element independently of the others, that is

$$Q\{x\} = [Q\{x_1\}, Q\{x_2\}, \cdots Q\{x_n\}]^T$$

*This work was supported in part by a grant from the Colorado Advanced Software Institute and by Mobile Data Systems, Ltd., Boulder, Colorado.

where ^{T} denotes the transpose. For a scalar v, the TCT quantizer satisfies

$$0 \leq v - Q\{v\} < q$$

where q is the quantization step size. Without loss of generality, q will be assumed to be unity. Then $Q\{v\}$ is the maximum integer not greater than v, that is, the floor of v.

Normal form filters are those for which A is a normal matrix, that is, $AA^T = A^TA$. These realizations are attractive because they are free of limit cycles due to overflow and MT quantization as long as they are linearly stable. However, these filters are still susceptible to limit cycles due to TCT quantization. The following theorems (established in [5]) give the g.a.s. regions for normal form filters with TCT quantization.

Theorem 1 *Consider the following first-order system*

$$x(k+1) = Q\{\lambda x(n)\}$$

where $Q\{.\}$ represents TCT quantization. If $-1 < \lambda < 0$, then $x(n) \to 0$, as $n \to \infty$.

Theorem 2 *Consider a second-order coupled-form digital filter*

$$x(n+1) = Q\{Ax(n)\}$$

where $A = \begin{bmatrix} \sigma & \omega \\ -\omega & \alpha \end{bmatrix}$, $x(n) = \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix}$ *and $Q\{.\}$ represents TCT quantization. If $|\sigma| + |\omega| < 1$ and $\sigma < 0$ then $x(n) \to 0$, as $n \to \infty$.*

The above conditions are very restrictive in the class of filters that can be designed. The g.a.s. region given by Theorem 2 is shown in Fig. 1. In [6] and [7], the g.a.s. region has been extended by finding some bounds on the limit cycles and then using an exhuastive search. The extended region is given in Fig. 2.

3. STABILITY OF BLOCK NORM FORM FILTERS WITH TCT

Consider a digital filter with the state and output equations given in (1). Instead of processing a signal input $u(k)$ to obtain signal output $y(k)$, we process the input sequence in blocks of length L. That is, the state vector is updated every L samples by using input and output buffers. Writing out (1) for $k+2, k+3, \cdots, k+L$ gives the following:

$$\begin{bmatrix} x'(k+L) \\ y(k) \\ y(k+1) \\ \vdots \\ y(k+L-1) \end{bmatrix} = \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} \begin{bmatrix} x'(k) \\ u(k) \\ u(k+1) \\ \vdots \\ u(k+L-1) \end{bmatrix} \quad (2)$$

where $A' = A^L, B' = [A^{L-1}B, A^{L-2}B, \cdots B], C' = [C, CA, \cdots CA^{L-1}]^T$, and

$$D' = \begin{bmatrix} D & 0 & \cdots & 0 \\ CB & D & \cdots & 0 \\ & \vdots & & \\ CA^{L-2}B & CA^{L-3}B & \cdots & D \end{bmatrix}.$$

The above is the well known block form structure [8]. We now present a new block structure that extends the TCT g.a.s. region of Fig. 2.

Let $L = 2$ and A be a second-order normal form realization, that is

$$A = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}. \quad (3)$$

Define a diagonal matrix E such that

$$E_{ii} = -sgn\{A_{ii}\} = -sgn\{\sigma\} \quad (4)$$

where $sgn\{.\}$ denotes the signum operation. Also define

$$A_- = A \times E. \quad (5)$$

It is easy to see that A_- is negative in the diagonal and that $A_-A_- = A^2$. Please note that the above also holds for first-order systems, that is, with A being a scalar.

Let us now realize a first- or second-order normal form filter in **modified block form** as shown in Fig. 3, where the block length is $L = 2$. This structure is essentially the same as a standard block realization, except that instead of updating $x(k+2) = Q\{A^2x(k)\}$, we perform

$$x(k+2) = Q\{A_-x(k+1)\} = Q\{A_-Q\{A_-x(k)\}\}. \quad (6)$$

The main theorems of the paper are now presented.

Theorem 3 *Consider a first-order digital filter with system matrix $A = \lambda$. If $|\lambda| < 1$, then the block filter realized as in Fig. 3 with TCT quantization, is globally asymptotically stable.*

Theorem 4 *Consider a second-order normal form digital filter with system matrix* $A = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}$ *implemented in block form as in Fig. 3 with TCT quantization. If $|\sigma|+|\omega| < 1$ then the filter is globally asymptotically stable.*

Proof: The proofs of Theorems 3 and 4 are straight forward. $x(k+2)$ in (6) represents the samples of the state variable in normal form realization with system matrix A_-, which is globally asymptotically stable based on Theorems 1 and 2. Therefore the realization of Fig. 3 is globally asymptotically stable. ♣

The results of Theorems 3 and 4 are significant because as illustrated in *Fig.*4, the g.a.s region has been extended to the right hand side of parameter plane. That is, the g.a.s. region has been doubled with a block length of only 2.

The g.a.s. region in Fig. 4 can actually be extended to the entire unit circle by increasing the block length. Since the eigenvalues of the block system matrix are $\lambda' = \lambda^L$, it is possible to move the eigenvalues further inside by increasing L. A suitable search algorithm for finding the minimum block length is given in Fig. 5, which we now explain in detail. We start with a given second order filter with eigenvalues $\sigma \pm j\omega$. If the eigenvalues are in the g.a.s. region of Fig. 4 ($|\sigma| + |\omega| < 1$), then we determine if they are in the g.a.s. region of Fig. 1 ($\sigma < 0$). If so, we implement in standard normal form. If not, we implement in modified

block form with block length $L = 2$ as in Fig. 3. On the other hand, if the g.a.s. region of Fig. 4 is not satisfied, then we know that we need a block length $L > 2$. Next, in Part A of Fig. 5, L is increased until the eigenvalues are in the g.a.s. region of Fig. 4. This block size is denoted by L_0. At this point, we can implement the modified block form as in Fig. 3 with a block size of $2L_0$ and be guaranteed of stability. However, it may be possible to have a block size less than $2L_0$ and still have stability. Part B of Fig. 5 is devoted to search for this minimal block size. If $\sigma_k < 0$, then we can implement as a standard block system with $L = k$. Otherwise, we increase the block size at each increment of k and check if $\sigma_k < 0$. If so, we can implement as a standard block system with $L = k$. But if we end up increasing the block size to greater than $2L_0$, then we resort to implementation as a modified block form with block size $2L_0$. In other words, the maximum block size needed is $2L_0$, and the algorithm of Fig. 5 guarantees an implementation with a minimal block size.

4. EXAMPLES

In this section, some examples are given to illustrate the foregoing results.

Example 1: Let $\sigma = 0.8, \omega = 0.1, x(0) = [-2, -1]'$. The eigenvalues satisfy the g.a.s. region of Fig. 4 but not of Fig. 1. When this system is implemented with TCT quantization in standard (non block) form, it yields the follwing period-one limit cycle: $[-2, -1]'$, $[-2, -1]'$, $\cdots$.

With the modified block normal form realization ($L = 2$), the system yields $x(1) = [1, 1]'$, $x(2) = [-1, -1]'$, $x(3) = [0, 0]' \Rightarrow$ stable.

Example 2: Let $\sigma = 0.8, \omega = -0.5, x(0) = [2, -1]'$. The eigenvalues do not satisfy the g.a.s. region of Fig. 4 and therfore we need a block size $L > 2$. Standard realization yields a limit cycle of period 10.

The algorithm of Fig. 5 yields $L = 3$ with standard block realization. With this structure, the state vector gives $x(3) = [0, 1]'$, $x(6) = [-1, -1]'$, $x(9) = [0, -1]'$, $x(12) = [0, 0]' \Rightarrow$ stable.

5. CONCLUSION

Normal form digital filters have been investigated for stability under TCT quantization. It is shown that if a block structure is used, then with a suitable block size, all limit cycles can be suppressed as long as the filter is linearly stable. An algorithm is presented that determines for a given filter, what kind of structure is required, namely, non-block, standard block, or modified block. The algorithm also determines the minimum block size required.

REFERENCES

[1] A. Lepschy, G.A. Mian and U. Viaro, " Effects of quantization in second order fixed-point digital filters with two's complement truncation quantizers," IEEE Trans. on Circuits and Systems, vol. CAS-35, pp.461-466, Apr. 1988.

[2] T. Thong and B. Liu, "Limit cycles in the combinatorial implementation of digital filters," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. ASSP-24, pp. 248-256, June 1976.

[3] D.C. Munson, J.H. Strickland, and T.P. Walker, "Maximum amplitude zero-input limit cycles in digital filters," IEEE Trans. on Circuits and Systems, vol. CAS-31, pp. 266-275, Mar. 1984.

[4] T. Bose, D.P. Brown, "Limit Cycles in zero-input digital filters due to two's complement quantization," IEEE Trans. on Circuits and Systems, vol. CAS-37, pp.568-571, Apr. 1990.

[5] T. Bose and M.-Q. Chen, "Stability of digital filters implemented with two's complement truncation quantization," IEEE Trans. on Signal Processing, vol.40 pp.24-31, Jan. 1992.

[6] T. Bose, M.-Q. Chen and F. Brammer, "Stability of normal form digital filters with two's complement quantization," IEEE Intl. Conf. on Acous., Speech, and Signal Processing, pp. 1889-1892, May 1991.

[7] K. Premaratne, E.C. Kulasekere, P.H. Bauer, and L.J. Leclerc, "An exhaustive search algorithm for checking limit cycle behavior of digital filters," IEEE Trans. on Signal Processing, vol. SP-44, pp. 2405-2412, Oct. 1996.

[8] C.S. Burrus, "Block implementation of digital filters," IEEE Trans. Circuit Theory, vol. CT-18, pp. 697-701, Nov. 1971.

[9] S.K. Mitra and R. Gnanasekaran, "Block implementation of recursive digital filters - new structures and properties," IEEE Trans. Circuits Syst., vol. CAS-25, pp. 200-207, Apr. 1978.

[10] C.W. Barnes, and S. Shinnaka, "Finite word effects in block-state realizations of fixed-point digital filters," IEEE Trans. on Circuits and Systems, vol. CAS-27, pp. 345-349, May 1980.

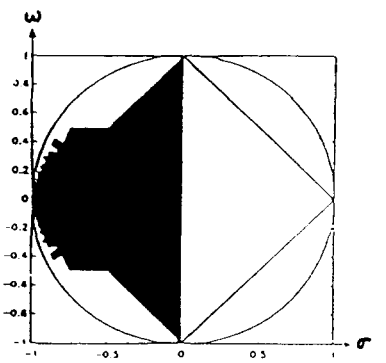

Fig. 1 G.A.S Region for TCT Quantizer Given by [6] and [7].

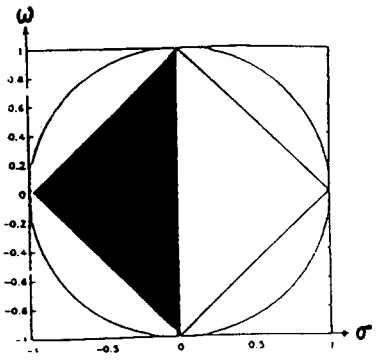

Fig. 2 G.A.S Region for TCT Quantizer Given by Theorem 2.

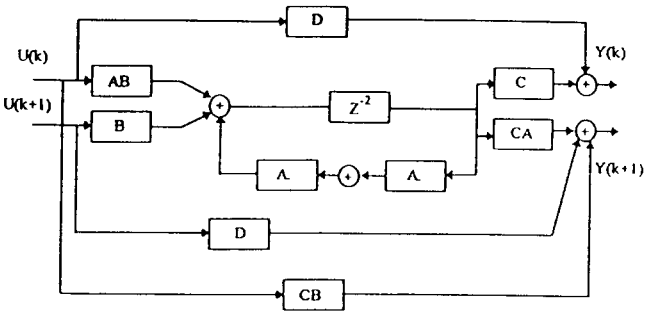

Fig. 3 Modified Block Structure.

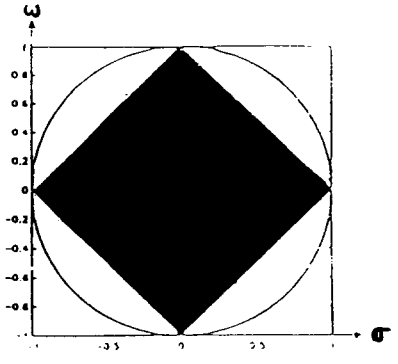

Fig. 4 G.A.S Region for TCT Quantizer Given by Theorem 4.

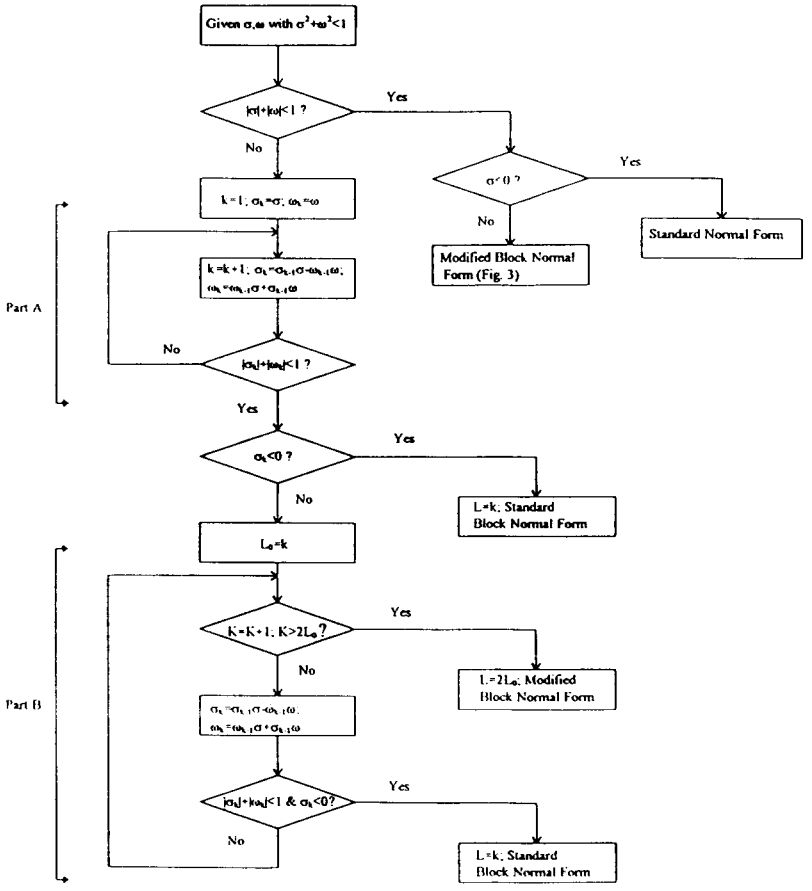

Fig. 5 Search Algorithm for Finding the minimum Block Size