AD-A065 135 RHODE ISLAND UNIV KINGSTON DEPT OF ELECTRICAL ENGIN--ETC F/G 9/5
SYNTHESIS OF DIGITAL FILTER STRUCTURES WITH LOW ROUNDOFF NOISE --ETC(U)
DEC 78 L B JACKSON, A G LINDGREN, Y KIM AFOSR-76-3057
UNCLASSIFIED AFOSR-TR-79-0036 NL

END
DATE FILMED
4 -79
DDC

LEVEL II

SYNTHESIS OF DIGITAL FILTER STRUCTURES WITH LOW ROUNDOFF NOISE AND COEFFICIENT SENSITIVITY

FINAL REPORT

Grant: AFOSR-76-3057

Period Covered: 1 June 1976 to 30 November 1978

Report Date: 15 December 1978

Submitted by:

Leland B. Jackson, Co-Principal Investigator
Allen G. Lindgren, Co-Principal Investigator
Department of Electrical Engineering
University of Rhode Island
Kingston, Rhode Island 02881

Submitted to:

Directorate of Mathematical and Information Sciences
Air Force Office of Scientific Research (NM)
Bolling Air Force Base, Washington, D.C. 20332

ABSTRACT

Optimal synthesis procedures for second-order state-space digital filters have been developed in the sense of minimum output roundoff noise with L_2 scaling. It has been demonstrated that these procedures are also nearly optimal for L_∞ scaling. The coefficient sensitivities have been shown to be closely related to the roundoff-noise components, and hence the optimal designs also have low sensitivity properties. The limit-cycle behavior of state-space structures has been investigated for rounding and for magnitude truncation. It was shown that rounding often leads to large autonomous limit cylces; while with magnitude truncation, it is possible to avoid limit cycles altogether. These results are presented in the two papers attached to this report.

PUBLICATIONS

1. L. B. Jackson, A. G. Lindgren, and Y. Kim, "Synthesis of State-Space Digital Filters with Low Roundoff Noise and Coefficient Sensitivity", Proc. 1977 IEEE Int'l. Symp. on Circuits and Systems, Phoenix, Arizona, April 1977, pp. 41-44.

2. L. B. Jackson, "Limit Cycles in State-Space Structures for Digital Filters", IEEE Trans. on Circuits and Systems, Vol. 26, No. 1, Jan. 1979 (attached).

3. L. B. Jackson, A. G. Lindgren, and Y. Kim, "Optimal Synthesis of Second-Order State-Space Structures for Digital Filters", IEEE Trans. on Circuits and Systems, Vol. 26, No. 3, Mar. 1979 (attached).

4. Y. Kim, "State-Space Structures for Digital Filters," Ph.D. Thesis, Univ. of Rhode Island, 1979.

OPTIMAL SYNTHESIS OF SECOND-ORDER STATE-SPACE STRUCTURES FOR DIGITAL FILTERS

Leland B. Jackson, Allen G. Lindgren, and Young Kim

University of Rhode Island
Kingston, Rhode Island 02881

Abstract

Sufficient conditions are derived for a second-order state-space digital filter with L_2 scaling to be optimal with respect to output round-off noise; and from these, a simple synthesis procedure is developed. Parallel-form designs produced by this method are equivalent to the block-optimal designs of Mullis and Roberts. The corresponding cascade-form designs are not equivalent, but they are shown by example to be quite close in performance. It is also shown that the coefficient sensitivities of this structure are closely related to its noise performance. Hence, the optimal design has low coefficient sensitivity properties, and any other low-sensitivity design is a good candidate for near-optimal noise performance. The uniform-grid structure of Rader and Gold is an interesting and useful case in point.

This research was sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant. No. AFOSR-76-3057.

Introduction

The synthesis of IIR digital filter structures with low roundoff noise based on state-space formulation has been introduced by Hwang [1-4] and Mullis and Roberts [5-7]. These structures are especially effective in reducing the output noise level from narrow-bandwidth filters. However, the general technique leads to structures having many more multipliers than the canonical structures (i.e., N^2 more, where N is the filter order). Recognizing that this would often make these structures impractical, Mullis and Roberts have proposed that the state-space structural form be used only for the individual sections of the cascade or parallel forms and that these structures then be optimized to produce "block-optimal" cascade or parallel forms. These block-optimal structures have only about twice the number of multipliers as the canonical structures, but greatly decreased noise levels are still achieved for narrow-bandwidth filters.

The synthesis procedure of Mullis and Roberts involves the calculation of "block second-order modes" for the filter, which are obtained from the eigenvalues of certain matrices. For the block-optimal parallel form, the optimization of the overall network is equivalent to optimizing each of the parallel sections separately. For the block-optimal cascade form, however, the choice of pairing and ordering for the poles and zeros of the filter influences the resulting optimized design, although the pairing and ordering choice itself is not optimized. In either case, the synthesis procedure is relatively complicated, and simpler procedures for optimal or near-optimal design are desirable.

We have derived a set of sufficient conditions for a second-order state-space structure to be optimal, and from these conditions, a simple synthesis procedure is produced. Parallel designs produced by this procedure are block-optimal; but because the sections are individually synthesized (except for scaling), the resulting cascade designs are not block-optimal. Design examples show, however, that the difference in performance between our technique and that of Mullis and Roberts for the cascade form is quite small.

We have also shown that the state-variable noise terms are closely related to the sensitivities of the network coefficients. Hence, the designs with optimal noise performance also have low coefficient sensitivities. Alternatively, any design with low-sensitivity properties is a good candidate for near-optimal noise performance. In particular, an alternate synthesis procedure leading to good (but not optimum) noise performance is to force the state transition matrix A to be anti-symmetric, which corresponds to the uniform-grid structure of Rader and Gold [12].

Noise Minimization

A flow diagram of the second-order state-space structure is shown in Fig. 1, corresponding to the state equations

$$\bar{x}(n+1) = A\bar{x}(n) + bu(n) + \bar{e}(n)$$
$$y(n) = c\bar{x}(n) + du(n) + e_3(n). \qquad (1)$$

The roundoff errors $e_j(n)$, $j = 1,2,3$, are assumed to be uncorrelated over n and j, and hence the output noise power is given by

$$\sigma^2 = k\,\sigma_o^2 \sum_{j=1}^{3} \|G'_j\|_2^2 \qquad (2)$$

where $G'_j(z)$ is the (scaled) frequency response from $e_j(n)$ to $y(n)$, $\|\cdot\|_p$ denotes the L_p norm, σ_o^2 is the variance of the noise from a single rounding operation, and k is the number of round-off errors included in each $e_j(n)$ [9]. In general, k = 3 if rounding is performed after multiplication (before summation); while k = 1 if rounding is performed after summation. Scaling is assumed to be of the form

$$\|F'_i\|_p = 1 \qquad (3)$$

where $F'_i(z)$ is the (scaled) frequency response from u(n) to $x_i(n)$, and most commonly, p = 2 or p = ∞ [9]. We analyze and optimize only the L_2 scaling case here although the resulting networks can be readily rescaled to realize L_∞ scaling.

Given a filter design (A,b,c,d) with unscaled and unoptimized responses $F_i(z)$ and $G_j(z)$, we consider the class of designs defined by $(A',b',c',d) = (T^{-1}AT,\ T^{-1}b, cT, d)$. Defining the response vectors $F^t(z) = \left(F_1(z), F_2(z)\right)$ and $G^t(z) = \left(G_1(z), G_2(z)\right)$ and similarly for the scaled vectors $F'(z)$ and $G'(z)$, it is readily shown that

$$F'(z) = [zI - A']^{-1} b' = T^{-1}F(z) \qquad (4)$$

and

$$G'(z) = [zI - A'^t]^{-1} c'^t = T^t G(z). \qquad (5)$$

(Note that $G'_3(z) = 1$ independently of T, and hence it cannot affect the noise minimization).

Mullis and Roberts [5-7] have shown that necessary and sufficient conditions for minimum output noise variance σ^2 subject to L_2 scaling are

$$W' = DK'D \tag{6}$$

and

$$K'_{ii}W'_{ii} = K'_{jj}W'_{jj} \text{ for all } i,j \tag{7}$$

where

$$K' = \sum_{k=0}^{\infty} (A'^k b') (A'^k b')^t \tag{8}$$

$$W' = \sum_{k=0}^{\infty} (c' A'^k)^t (c' A'^k) \tag{9}$$

and D is a diagonal matrix. Definitions (8) and (9) may be expressed in the frequency domain as

$$K' = \oint F'(z) F'^t(z^{-1}) z^{-1} dz \tag{10}$$

and

$$W' = \oint G'(z) G'^t(z^{-1}) z^{-1} dz. \tag{11}$$

Hence, the scaling constraint in (3) implies that

$$K'_{ii} = \|F'_i\|_2^2 = 1 \text{ for all } i, \tag{12}$$

and from (7) we then have

$$W'_{ii} = W'_{jj} \text{ for all } i,j \tag{13}$$

or

$$\|G_i\|_2^2 = \|G'_j\|_2^2 \text{ for all } i,j. \tag{14}$$

That is, the optimal networks is characterized by having equal noise contributions from each error source.

Equations (12) and (13) show that (6) is satisfied if, and only if, $D = \rho I$; and thus an alternate condition which is both necessary and sufficient for optimality (after scaling) is simply

$$W' = \rho^2 K' . \tag{15}$$

In the second-order case, we note that (15) is not changed by writing it as

$$W' = \rho^2 MK'M \tag{16}$$

where

$$M = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

because W' and K' are symmetric matrices with equal diagonal elements. From (8) and (9), it is readily shown that (16) and thus (15) are satisfied by a network of the form

$$A'^t = MA'M \tag{17a}$$

and

$$c'^t = \rho Mb' ; \tag{17b}$$

and for complex-conjugate poles, (17) can always be realized with real-valued coefficient matrices A', b', and c'. After exploring

this case further, we will consider the case of real-valued poles in a second-order filter.

In terms of the elements of (A', b', c', d), (17) states simply that

$$a'_{11} = a'_{22}$$

and (18)

$$\frac{b'_1}{b'_2} = \frac{c'_2}{c'_1} .$$

This network is readily synthesized from an arbitrary (A,b,c,d) as follows: One implication of (17) is that if the transpose of the optimal network is formed and the states x_1 and x_2 are interchanged, the resulting network is identical with the original optimal network except for scaling. But if we form the network $(MA^tM, Mc^t/2, b^tM, d/2)$ and place it in parallel with the network $(A, b/2, c, d/2)$, we produce an overall network with the above property and the same transfer function. Therefore, we can synthesize the optimal network simply by merging these two parallel networks into a single network $(\hat{A}, \hat{b}, \hat{c}, d)$ and then scaling it via

$$T = \begin{bmatrix} \|\bar{F}_1\|_2 & 0 \\ 0 & \|\hat{F}_2\|_2 \end{bmatrix} . \tag{19}$$

The above synthesis technique is particularly straightforward when one starts with the transfer function $H(z)$ expressed as

$$H(z) = d + \frac{\gamma_2 z^{-2} + \gamma_1 z^{-1}}{\beta_2 z^{-2} + \beta_1 z^{-1} + 1} \tag{20}$$

and implemented, for example, by

$$A = \begin{pmatrix} -\beta_1 & -\beta_2 \\ 1 & 0 \end{pmatrix} \qquad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{21}$$

$$c = (\gamma_1 \ \gamma_2).$$

The coefficients of $(\hat{A},\hat{b},\hat{c},\hat{d})$ are then

$$\hat{a}_{11} = \hat{a}_{22} = -\beta_1/2$$

$$\hat{b}_1 = \tfrac{1}{2}(1+\gamma_2) \qquad \hat{b}_2 = \tfrac{1}{2}\gamma_1 \tag{22}$$

$$\hat{c}_1 = \frac{\gamma_1}{1+\gamma_2} \qquad \hat{c}_2 = 1$$

$$\hat{\alpha}_{12} = \gamma_1^{-2}(1+\gamma_2)\left[(\gamma_2 - \tfrac{1}{2}\beta_1\gamma_1) \pm \sqrt{\gamma_2^2 - \gamma_1\gamma_2\beta_1 + \gamma_1^2\beta_2}\,\right]$$

$$\hat{\alpha}_{21} = (1+\gamma_2)^{-1}\left[(\gamma_2 - \tfrac{1}{2}\beta_1\gamma_1) \mp \sqrt{\gamma_2^2 - \gamma_1\gamma_2\beta_1 + \gamma_1^2\beta_2}\,\right].$$

The expression under the radical in $\hat{a}_{12}$ and $\hat{a}_{21}$ is always positive for complex-conjugate poles, and hence the coefficients in (22) are all real-valued in that case. Note that the coefficients in (22) do indeed satisfy (18). If the network $(\hat{A},\hat{b},\hat{c},\hat{d})$ is then scaled via (19), the resulting second-order network is optimal for L_2 scaling. The network may be scaled instead via

$$T = \begin{pmatrix} \|\hat{F}_1\|_\infty & 0 \\ 0 & \|\hat{F}_2\|_\infty \end{pmatrix}$$

to realize L_∞ scaling, if desired. The resulting network is not the optimal L_∞ network, in general, but our initial experiments with this case indicate it is very nearly optimal.

Turning briefly to the case of real-valued poles (p_1 and p_2), we consider an $\hat{A}$ of the form

$$\hat{A} = \begin{pmatrix} a_1 & a_2 \\ a_2 & a_1 \end{pmatrix} \tag{23}$$

where we then have

$$a_1 = \frac{1}{2}(p_1 + p_2), \qquad a_2 = \pm \frac{1}{2}(p_1 - p_2).$$

Note that (23) still satisfies (17a) as well as $\hat{A}^t = \hat{A}$. The appropriate relationship between $\hat{b}$ and $\hat{c}$ depends on the location of the zero for $H(z)-d$. In particular, $\hat{b}$ and $\hat{c}$ can satisfy either (17b) or

$$\hat{c}^t = \rho \hat{b} \tag{24}$$

in order to satisfy the optimality condition in (16) or (15), respectively. Scaling of a network satisfying (23) and (17b) is performed as before via (19); while scaling for (23) and (24) is accomplished via an orthogonal transformation as described in the next section in order to preserve the symmetry of A'.

Coefficient Sensitivities

The minimization of the roundoff noise also implies low sensitivities for the network coefficients. To show this, we note first that

$$\frac{\partial H(z)}{\partial b'_i} = G'_i(z)$$

and (25)

$$\frac{\partial H(z)}{\partial c'_j} = F'_j(z)$$

where $H(z)$ is the overall transfer function of the second-order section. But in the optimal network we have minimized the sum of the L_2 norms of these sensitivities, subject to the scaling constraint in (3). We then define the sensitivities

$$S'_{ij}(z) = \frac{\partial H(z)}{\partial a'_{ij}} = G'_i(z) \cdot F'_j(z) \tag{26}$$

for the feedback coefficients in A'. In [8] it was shown that

$$\|S'_{ij}\|_1 \leq \|G'_i\|_2 \, \|F'_j\|_2, \tag{27}$$

and hence, from (3) and (27), we have

$$\|S'_{ij}\|_1 \leq \|G'_i\|_2. \tag{28}$$

Therefore, the upper bounds on the four sensitivity norms in (28) have also been jointly minimized. Note, in particular, from (14) and (28) that the upper bounds on all four $\|S'_{ij}\|_1$ are the same.

Alternatively, of course, (28) provides lower bounds on the noise contributions from $e_1(n)$ and $e_2(n)$ in terms of the sensitivities of the coefficients in A'. If these bounds are reasonably tight, as experience has shown them to be [8], then low-sensitivity networks should also provide low roundoff noise. One network having uniformly low sensitivity over the entire z plane, at least in terms of the grid of possible pole locations with quantized coefficients, was noted ten years ago by Rader and Gold [12]. This network is produced simply by forcing A to be anti-symmetric, i.e.,

$$\begin{aligned} a_{11} &= a_{22} = \text{Re[poles]} \\ -a_{12} &= a_{21} = \text{Im[poles]} \end{aligned} \tag{29}$$

where the poles must be complex conjugates. The coefficient vectors b and c are not uniquely specified by (29), and there are two distinctly useful ways to generate b and c. We note first that (29) does satisfy the first condition in (18); and thus if b and c are chosen to satisfy the second condition in (18) as well, the anti-symmetric network and the optimal network are related by a simple diagonal scaling transformation of the form of (19). This constitutes, in fact, another simple technique for synthesizing the optimal structure when the poles are complex. The uniform-grid property is not, however, preserved in the optimal network, in general.

To preserve the uniform-grid property of the anti-symmetric network, one instead seeks a transformation T which alters b (and c) to satisfy (3), but does not change A. The appropriate transformation is an orthogonal transformation times a scaling constant [15], i.e.,

$$T = \mu \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{30}$$

where the "rotation" angle θ and the scaling constant μ are determined such that $\|F_1'\|_2 = \|F_2'\|_2 = 1$. In particular, from (4) and (30) we find that (with $\mu = 1$)

$$\|F_1'\|_2^2 = \cos^2\theta \, \|F_1\|_2^2 + \sin^2\theta \, \|F_2\|_2^2 - 2\sin\theta\cos\theta \, (F_1, F_2) \tag{31}$$

$$\|F_2'\|_2^2 = \sin^2\theta \, \|F_1\|_2^2 + \cos^2\theta \, \|F_2\|_2^2 + 2\sin\theta\cos\theta \, (F_1, F_2) \tag{32}$$

where

$$(F_1, F_2) = \left(\frac{1}{2\pi j}\right) \oint F_1(z)\, F_2(z^{-1}) z^{-1} dz.$$

Subtracting (31) and (32), we obtain

$$\tan 2\theta = \frac{\|F_1\|_2^2 - \|F_2\|_2^2}{2(F_1, F_2)} \tag{33}$$

from which to solve for θ. The constant μ is then readily obtained as $1/\|F_i'\|_2$. The contour integrals in (33) are rapidly computed as described in [14]. The resulting anti-symmetric design is not optimal, but as shown in the examples of the next section, the noise performance is only about 1-3db worse than the optimal design.

Examples

Five filter design examples are presented in Tables 1 and 2, where the noise gains σ^2/σ_o^2 are given in dB. The first three designs are from [13], and the last two are described by Mullis and Roberts [6]. In Table 1, the canonical (1P), anti-symmetric, and block-optimal designs in parallel form are compared. The block-optimal designs were obtained from (22) and (19). The anti-symmetric designs satisfy (3) and (29), and hence are suboptimal. Note that the performance of the block-optimal design is best in all 5 cases, with the most significant improvement coming in the narrowband case, as expected. Note also that the anti-symmetric designs are close to the block-optimal designs in performance.

The corresponding results for the cascade-form designs are presented in Table 2. An additional column is included here because, as discussed in the Introduction, our "section-optimal" designs resulting from (22) and (19) are not quite the same as

the block-optimal designs of Mullis and Roberts. The functions $\hat{F}_j(\omega)$ in (19) include the transfer functions of all preceding sections so that (3) is satisfied exactly. This form of scaling has also been applied here to the block-optimal designs.

(No 9)

No attempt has been made to optimize the section ordering. As the table shows, the results for the section-optimal and block-optimal designs are quite close and are significantly better than those for the canonical structure in all but one case, where they are comparable. The results for the anti-symmetric designs are also near-optimal, but differ by 2-2.5db in the Butterworth cases.

It should be noted that it is possible for the canonical forms to have lower noise gain than the optimal forms, although this did not happen in these examples, because the second-order state-space structure has almost twice the number of multipliers of the canonical structure. However, this rarely happens, and when it does, the advantage is small.

Conclusions

We have determined sufficient conditions for a second-order digital filter in state-variable form with L_2 scaling to be optimal with respect to output roundoff noise and have described two simple techniques to synthesize the optimal network. In addition, we have shown that L_1 or L_2 norms of the coefficient sensitivities of the optimal network (or bounds on these sensitivities) are also minimized, and hence this network provides both low roundoff noise and coefficient sensitivity. These second-order structures can be combined to form "section-optimal" parallel or

cascade forms which are the same as or close to, respectively, the "block-optimal" parallel or cascade forms of Mullis and Roberts.

The optimal network is within a simple diagonal transformation of the anti-symmetric network of Rader and Gold, which has the nice property of a uniform grid of possible pole locations in the z plane with quantized coefficients. If the anti-symmetric network is scaled so as to preserve the anti-symmetric property, we find that the resulting noise performance is close to (within typically 1-3db of) that of the optimal network. In either case, the noise is relatively constant (or actually decreases) as the bandwidth of the filter is reduced, in contrast with the performance of canonical structures.

Recently, it has also been shown that the anti-symmetric ("normal") form [16,17] and our optimal form [17] for second-order sections cannot sustain autonomous overflow oscillations. Hence, parallel or cascade structures comprised of such sections have the additional desirable property that autonomous overflow oscillations cannot occur.

Acknowledgments

The authors wish to thank C.T. Mullis and R.A. Roberts for providing us with a copy of their computer program for the design of block-optimal digital filters.

The helpful suggestions of the reviewers are also appreciated.

References

1. S. Y. Hwang, "Roundoff Noise in State-Space Digital Filtering: A General Analysis", IEEE Trans., Vol. ASSP-24, pp. 256-262, June 1976.

2. _____, "Dynamic Range Constraint in State-Space Digital Filtering", IEEE Trans., Vol. ASSP-23, pp. 591-593, Dec. 1975.

3. _____, "Minimum Unit Noise in the State-Space Digital Filtering", Proc. IEEE ISCAS, pp. 352-355, April 1976.

4. _____, "Roundoff-Noise Minimization in State-Space Digital Filtering", Proc. IEEE ICASSP, pp. 498-500, April 1976.

5. C. T. Mullis and R. A. Roberts, "Synthesis of Minimum Roundoff Noise Fixed Point Digital Filters", IEEE Trans. Vol. CAS-23, pp. 551-562, Sept. 1976.

6. _____, "Roundoff Noise in Digital Filters: Frequency Transformation and Invariants", IEEE Trans., Vol. ASSP-24, pp. 538-550, Dec. 1976.

7. _____, "Filter Structures which Minimize Roundoff Noise in Fixed Point Digital Filters", Proc. IEEE ICASSP, pp. 505-508, April 1976.

8. L.B. Jackson, "Roundoff Noise Bounds Derived from Coefficient Sensitivities for Digital Filters", IEEE Trans., Vol. CAS-23, pp. 481-485, Aug. 1976.

9. _____, "On the Interaction of Roundoff Noise and Dynamic Range in Digital Filters", B.S.T.J., Vol. 49, pp. 159-184, Feb. 1970.

10. _____, "Lower Bounds on the Roundoff Noise from Digital Filters in Cascade or Parallel Form", Proc. IEEE ISCAS, pp. 638-641, April 1976.

11. _____, "Roundoff-Noise Analysis for Fixed-Point Digital Filters Realized in Cascade or Parallel Form", IEEE Trans. Vol. AU-18, pp. 107-122, June 1970.

12. C. M. Rader and B. Gold, "Effects of Parameter Quantization on the Poles of a Digital Filter", Proc. IEEE, Vol. 55, pp. 688-689, May 1967.

13. L. B. Jackson, "An Analysis of Roundoff Noise in Digital Filters", Sc.D. Thesis, Stevens Inst. of Tech., Hoboken, N.J., 1969.

14. 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., Vol. ASSP-22, pp. 326-329, October 1974.

15. C. T. Mullis, private communication.

16. C. W. Barnes and A. T. Fam, "Minimum Norm Recursive Digital Filters that are Free of Overflow Limit Cycles", IEEE Trans., Vol. CAS-24, pp. 569-574, Oct. 1977.

17. W. L. Mills, C. T. Mullis and R. A. Roberts, "Digital Filter Realizations with Overflow Oscillations", Proc. 1978 IEEE ICASSP, pp. 71-74, April 1978.

Table 1. σ^2/σ_o^2 IN dB FOR PARALLEL FORM DESIGN

Filter	Canonical	Anti-Symm.	Block-Opt
Cheb-II BRF, N=6	13.8	11.0	10.9
Cheb-I LPF, N=10	19.2	14.0	13.8
Elliptic LPF, N=10	16.8	13.7	13.5
Butterwth LPF, N=6 $f_c=0.25f_s$	14.2	14.6	13.4
Butterwth LPF, N=6 $f_c=0.025f_s$	27.0	14.8	13.4

Table 2. σ^2/σ_o^2 IN dB FOR CASCADE FORM DESIGNS

Filter	Canonical	Anti-Sym.	Sect-Opt	Block-Opt
Cheb-II BRF, N=6	21.0	10.5	10.5	10.5
Cheb I LPF, N=10	29.9	24.2	24.5	24.1
Elliptic LPF, N=10	21.5	16.9	16.8	16.5
Butterwth LPF, N=6 $f_c=0.25f_s$	9.2	11.0	9.1	9.0
Butterwth LPF, N=6 $f_c=0.025f_s$	18.7	10.3	7.9	7.7

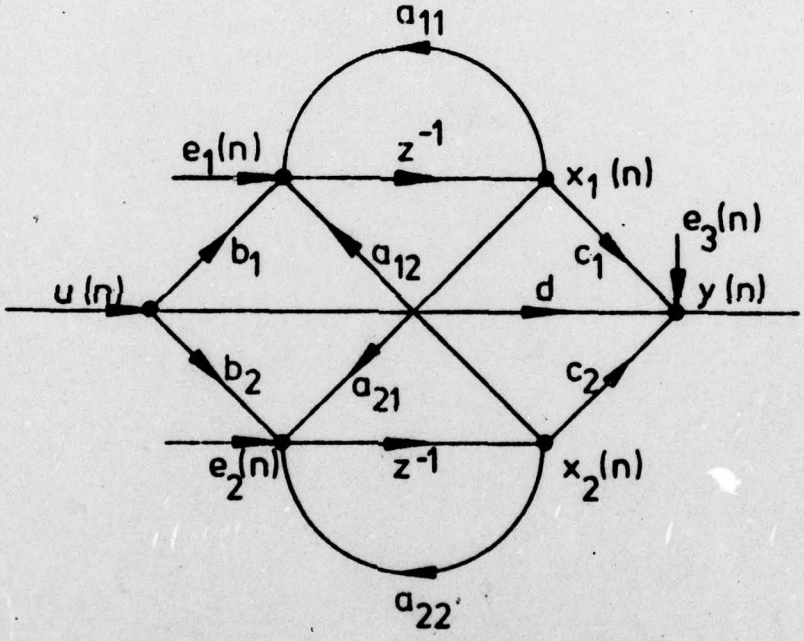

Fig. 1. Second-order state-space structure showing roundoff noise sources.

(Correspondence Item)

Limit Cycles in State-Space Structures for Digital Filters

Leland B. Jackson
University of Rhode Island
Kingston, Rhode Island 02881

ABSTRACT

It is shown that large quantization limit cycles are possible in state-space structures for digital filters with rounding, as opposed to magnitude truncation where such limit cycles can be avoided. The coupled-loop structure is considered specifically. The maximum limit-cycle amplitudes with rounding are obtained analytically for poles near the imaginary (or real) z-plane axis and by simulation for other regions of the z-plane.

This research was supported by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant No. AFOSR-76-3057

Introduction

State-space structures for digital filters have been described by Mullis and Roberts [1-3], Hwang [4-7], and Jackson, Lindgren and Kim [8-9]. They have shown that these structures can possess very low roundoff noise and coefficient sensitivity at the expense of increased computation over canonical structures. Barnes and Fam [10] have also shown that these structures can be designed to eliminate the possibility of overflow oscillations. Specifically, the second-order "coupled-loop" structure of Rader and Gold [11], which has an anti-symmetric system matrix A, is free of autonomous overflow oscillations; and Jackson, Lindgren and Kim have shown that the roundoff-noise performance of the coupled-loop structure is close to optimal.

A key remaining question concerning state-space structures is their limit-cycle behavior due to internal data quantization. Although not noted by Barnes and Fam, filters satisfying their conditions for the absence of overflow oscillations will also be free of zero-input quantization limit cycles if magnitude truncation is used. However, if rounding is employed, limit cycles can occur which are substantially larger than those in the corresponding canonical structures, as we show in this correspondence. Hence, in using state-space structures for digital filters, one may well choose to employ magnitude truncation in spite of its increased noise variance and noise/signal correlation as compared with rounding.

Rounding Limit Cycles

We will investigate the zero-input limit-cycle behavior of the second-order coupled-loop structure with rounding as being indicative of the more general (and more complicated) second-order case. Second-order structures are of particular interest because they are the basic building-blocks of various optimal or near-optimal cascade and parallel forms [1-3,8-9].

With zero input and rounding, the filter obeys the state equation

$$\underline{x}(n+1) = [A\underline{x}(n)]_R \qquad (1)$$

where $\underline{x}(n)$ is the state vector at time n, A is the system matrix, and $[\cdot]_R$ denotes the rounding operation. The second-order coupled-loop case is depicted in Fig. 1, and the canonical case in Fig. 2. We assume rounding after each multiplication. The complex-conjugate poles of this filter are simply a+jb.

A simple illustration of the large limit cycles possible with the coupled- loop structure is provided by setting a = 0. Then the effective-value model of Jackson [12] shows that we can have limit cycles with amplitudes as large as

$$K = \left[\frac{0.5}{1-b} \right]_I \qquad (2)$$

where $[\cdot]_I$ denotes the integer part. The corresponding expression for the canonical structure is

$$K = \left[\frac{0.5}{1-b^2} \right]_I \approx \left[\frac{0.25}{1-b} \right]_I \qquad (3)$$

because the second-order coefficient in that case is $\beta_2 = b^2$. Hence, the maximum limit-cycle amplitude in the coupled-loop case is approximately twice that in the canonical case for poles on or near the

imaginary axis. A similar analysis can be made for the coupled-loop structure with poles on or near the real axis (i.e., $b \approx 0$).

The nature of the more general situation for the coupled-loop structure is illustrated in Fig. 3, which shows in one quadrant of the z plane the corresponding maximum limit-cycle amplitudes as determined by computer simulation for radii up to about 0.9. Mirror-image symmetry holds about both the real and imaginary axes, as well as about $\pm 45^{\circ}$ diagonals through the origin. Hence, the amplitude for poles with given radius near the real axis is the same as near the imaginary axis, as opposed to the canonical case where much larger amplitudes are produced near the real axis [12].

The worst case for the coupled-loop structure occurs on the 45° diagonals, i.e., for frequencies of $\omega = \pi/4$ and $3\pi/4$. Table 1 shows that the radii for K = 1,2,3 in the canonical case are comparable to those for K = 3,6,10, respectively, in the coupled-loop case, and hence the maximum limit-cycle amplitudes for the coupled-loop structure are about triple those for the canonical structure at these frequencies.

Conclusions

The advantages of "well-designed" state-space structures for digital filters over canonical structures (at the expense of increased computation) include reduced roundoff noise and coefficient sensitivity, freedom from overflow oscillations, and freedom from quantization limit cycles if magnitude truncation is employed. If rounding is employed, however, large quantization limit cycles can occur in the state-space structures. The coupled-loop

structure was selected for investigation because it is particularly simple to design and analyze and because it possesses the advantages cited above.

References

1. C. T. Mullis and R. A. Roberts, "Synthesis of Minimum Roundoff Noise Fixed Point Digital Filters", IEEE Trans. Vol. CAS-23, pp. 551-562, Sept. 1976.

2. _____, "Roundoff Noise in Digital Filters: Frequency Transformation and Invariants", IEEE Trans., Vol. ASSP-24, pp. 538-550, Dec. 1976.

3. _____, "Filter Structures which Minimize Roundoff Noise in Fixed Point Digital Filters", Proc. IEEE ICASSP, pp. 505-508, April 1976.

4. S. Y. Hwang, "Roundoff Noise in State-Space Digital Filtering: A General Analysis", IEEE Trans., Vol. ASSP-24, pp. 256-262, June 1976.

5. _____, "Dynamic Range Constraint in State-Space Digital Filtering", IEEE Trans., Vol. ASSP-23, pp. 591-593, Dec. 1975.

6. _____, "Minimum Unit Noise in the State-Space Digital Filtering", Proc. IEEE ISCAS, pp. 352-355, April 1976.

7. _____, "Roundoff-Noise Minimization in State-Space Digital Filtering", Proc. IEEE ICASSP, pp. 498-500, April 1976.

8. L.B. Jackson, A.G. Lindgren and Y. Kim, "Synthesis of State-Space Digital Filters with Low Roundoff Noise and Coefficient Sensitivity", Proc. 1977 IEEE ISCAS, pp. 41-44, Apr. 1977.

9. _____, "Optimal Synthesis of Second-Order State-Space Structures for Digital Filters", submitted to the IEEE Trans. on Circuits and Systems.

10. C. W. Barnes and A. T. Fam, "Minimum Norm Recursive Digital Filters that are Free of Overflow Limit Cycles", IEEE Trans. Vol. CAS-24, pp. 569-574, Oct. 1977.

11. C. M. Rader and B. Gold, "Effects of Parameter Quantization on the Poles of a Digital Filter", Proc. IEEE, Vol. 55, pp. 688-689, May 1967.

12. L. B. Jackson, "An Analysis of Limit Cycles due to Multiplication Rounding in Recursive Digital (Sub) Filters", Proc. 7th Allerton Conf. on Circuits and Systems, pp. 69-78, Oct. 1969.

Canonical Structure		Coupled-Loop Structure	
Amplitude	Min. Radius	Amplitude	Min. Radius
K = 1	0.707	K = 3	0.707
K = 2	0.868	K = 6	0.884
K = 3	0.913	K = 10	0.919

Table 1. Minimum Radii for Given Limit-Cycle Amplitudes at $\omega = \pi/4$ and $3\pi/4$.

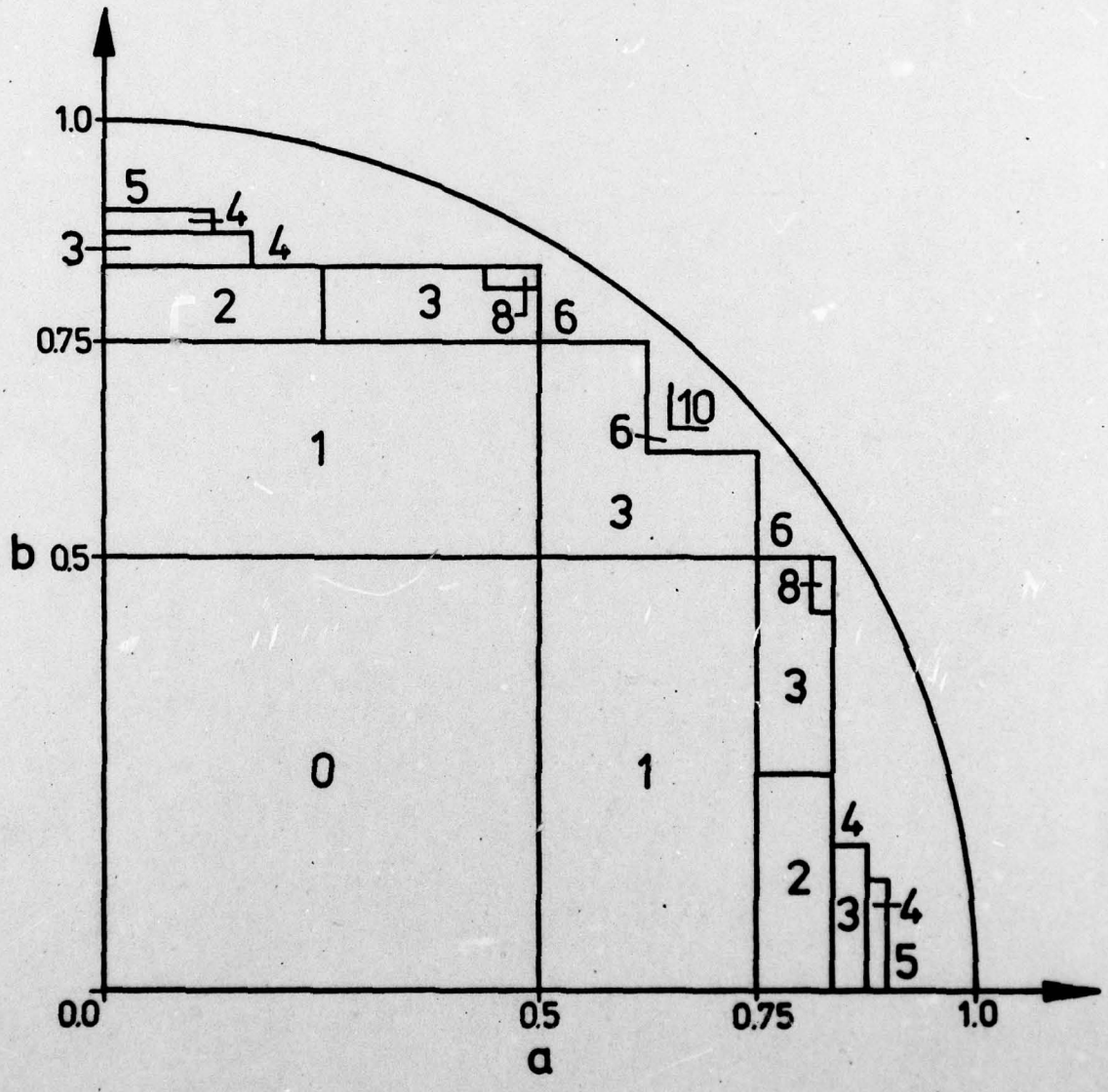

Fig. 3 - First quadrant of z plane showing maximum limit-cycle amplitudes for coupled-loop structure except very near the unit circle.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER (18) AFOSR-TR- 79-0036 (11) 2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) (6) SYNTHESIS OF DIGITAL FILTER STRUCTURES WITH LOW ROUNDOFF NOISE AND COEFFICIENT SENSITIVITY	5. TYPE OF REPORT & PERIOD COVERED (9) Final rept. 1 Jun 76-30 Nov 78
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) Leland B. Jackson and Allen G. Lindgren Young Kim	8. CONTRACT OR GRANT NUMBER(s) (15) AFOSR-76-3057 New
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Rhode Island Department of Electrical Engineering Kingston, R.I. 02881	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F (16) 2304/A6 (17) A6
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research/NM Bolling AFB, Washington, DC 20332 (11)	12. REPORT DATE 15 December 1978
	13. NUMBER OF PAGES 29
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) (12) 30p	15. SECURITY CLASS. (of this report) UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Digital Filters, Roundoff Noise, Coefficient Sensitivity, State-Space, Limit Cycles, Quantization

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Optimal synthesis procedures for second-order state-space digital filters have been developed in the sense of minimum output roundoff noise with L_2 scaling. It has been demonstrated that these procedures are also nearly optimal for L_∞ scaling. The coefficient sensitivities have been shown to be closely related to the roundoff-noise components, and hence the optimal designs also have low sensitivity properties. The limit-cycle behavior of state-space structures has been investigated for rounding and for magnitude truncation. It was shown that rounding often leads to large autonomous limit cycles; (continued on back)

DD FORM 1 JAN 73 1473 305 570 UNCLASSIFIED

20. Abstract continued.

while with magnitude truncation, it is possible to avoid limit cycles altogether. These results are presented in the two papers attached to this report.