MINIMUM ROUND-OFF NOISE SECOND-ORDER DIGITAL FILTER WITH PRACTICAL COMPLEXITY CONSTRAINTS

Kung Yao*
Hughes Aircraft Company
Canoga Park, CA 91303

ABSTRACT

It is known that, for a specified second-order digital filter transfer function, various realizations with finite precision arithmetic can yield significantly different round-off noises. For high performance communication and radar signal processing applications, the need for low round-off noise is clear. The minimum round-off noise n-th order digital filter of Mullis-Roberts generally requires $(n+1)^2$ multipliers. Most practical systems, however, desire to use a low number of multipliers. In this paper, we consider the minimum round-off noise second-order digital filter realization under the practical complexity constraints of using only four multipliers, two delays, and four two-input adders, The optimum constraint filter has the same complexity as the know canonic direct-form realization, yet its round-off noise can be significantly smaller for low-frequency rejection filtering applications. Some numerical results are presented.

INTRODUCTION

In many communication and radar signal processing problems, a desired overall digital filter transfer function is specified. If this filter is of even order n , then often, for practical reasons, it is implemented as cascades of n/2 second-order filters. Thus, with this cascade assumption, the optimum overall filter design problem reduces to that of the optimum second-order filter design problem. It is well known that, for a specified second-order digital filter transfer function, various realizations with finite precision arithmetic can yield significantly different round-off noises [1, p.153].

Consider a second-order digital filter with transfer function

$$H(z) = \frac{Y(z)}{U(z)} = \frac{1 + b_1 z^{-1} + b_2 z^{-2}}{1 - a_1 z^{-1} - a_2 z^{-2}} = \frac{(z-\alpha)(z-\bar{\alpha})}{(z-\beta)(z-\bar{\beta})} \quad . \tag{1}$$

* The author is with the Department of System Science, University of California, Los Angeles.
The work at UCLA is partially supported by the Electronics Program of the Office of Naval Research.

If all the filter coefficients $\{b_1,b_2,a_1,a_2\}$ are real-valued, then the zeros $\{\alpha,\bar{\alpha}\}$ and poles $\{\beta,\bar{\beta}\}$ form complex conjugate pairs. A commonly encountered topological realization of (1) is the canonic direct form II given in Fig.1. In the next section, a multiplicative round-off quantization noise model for this realization is discussed. Then, minimum round-off noise realizations with and without complexity constraints are considered.

ROW-OFF NOISE MODEL

In a digital filter that uses fixed-point arithmetic, additions introduce no error when no overflow occurs. However, the multiplication of two words of B_1 and B_2 bits generally yields a new word of $B_3 = (B_1+B_2)$ bits. If B_3 is greater than the allowed processor wordlength B , then some word-reducing operation such as rounding or truncation must be used. In this paper, we will use only rounding operations.

While the multiplication-rounding of two given finite-length words is a deterministic non-linear operation, complete deterministic analysis of all rounding operations in a processor is essentially too complicated for practical consideration. Thus, a simpler random linear model is used to replace the finite-word multiplication and rounding operation by an infinite-precision multiplication followed by an additive random round-off noise. Thus, the linear random round-off noise model, as applied to Fig.1, is shown in Fig. 1'.

Clearly, there are many possible ways to model the round-off noise e of the multiplier-rounding operations. The simplest models that have been used include the assumptions [1, p.415; p.310] :

1. The round-off noise e_i of the i-th multiplier-rounding operation is a zero-mean uniformly distributed random variable on [-q/2,q/2] of variance $\sigma_{e_i}^2 = q^2/12$, where $q = 2^{-(B_i-1)}$, and B_i is the i-th processor wordlength including the sign bit.

2. The noise e_i , modeled as a function of time, is a zero-mean wide-sense stationary white random sequence with uniform spectral density of $\sigma_{e_i}^2$ on $[-\pi,\pi)$.

3. Any two different noise sources e_i and e_j are uncorrelated for all times.

4. Each noise source e_i is uncorrelated with the input data sequence.

Let H(z) be the transfer function of the filter given by Fig. 1'. Then, the total round-off noise variance at the output is given by

$$\sigma_T^2 = \sum_{i=1}^{2} \sigma_{b_i}^2 + \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(e^{i\theta})|^2 d\theta \sum_{i=1}^{2} \sigma_{a_i}^2 \quad . \tag{2}$$

In (2), if any a_i or b_i is an integer, then the corresponding $\sigma_{a_i}^2$ or $\sigma_{b_i}^2$ is zero. Furthermore, if all rounding operations are done to B bits, and all multipliers are non-integers, then (2) reduces to

$$\sigma_T^2 = \sigma^2\{2 + \frac{2}{2\pi}\int_{-\pi}^{\pi} |H(e^{i\theta})|^2 d\theta\} \quad , \tag{3}$$

where

$$\sigma^2 = q^2/12 \quad , \quad q = 2^{-(B-1)} \quad . \tag{4}$$

From (2) or (3), it can be seen that the effect of the output round-off noise depends not only on the processor wordlengths B_i or B through $\sigma_{a_i}^2$ and $\sigma_{b_i}^2$, but also on the transfer function H(z) . Specifically, the effect of $\sigma_{a_i}^2$ can be reduced considerably if the transfer function attenuates over large parts of the frequency bandwidth. This basic property is important in the consideration of forthcoming minimum round-off noise digital filters.

OPTIMUM DIGITAL FILTER

In this section, some relevant results on the optimum digital filters without complexity constraints are summarized. As in this entire paper, the criterion of optimality is in the sense of minimum total round-off noise. This problem was originally formulated by Kaiser. [3] and studied in greater detail by Jackson [4],[5]. In recent years, much work has been done on this problem. Mullis and Roberts [6] have formulated a quite complete theory on the analysis and design of a minimum round-off noise n-th order digital filter. In their theory, fixed-point arithmetic is used and all input signals are assumed to be white random sequences. By using Jackson's ℓ^2 scaling rule, the probability of overflow is restricted to be sufficiently small, so that the digital filter can be assumed to be a linear system. Then the output round-off noise is evaluated in terms of internal multiplication-rounding noises via linear state-variable methods.

Upon coordinate transformations of the internal states of the filter by similarity transformations, maximum utilization of the dynamic range of the internal states and minimum output round-off noise is realized. For both equal and unequal state wordlength filters, remarkably compact minimum output round-off noise variance expressions were obtained. Explicit evaluation of these expressions is of the order of complexity of simultaneous diagonalization of two nxn positive-definite matrices which are, in turn, solutions of Liapunov matrix equations expressed in terms of the state-variable matrices of the transfer function.

For certain applications, such as narrow bandwidth low-pass filtering, the new Mullis-Roberts filters can yield output round-off noise variances many orders of magnitude better than known standard forms. Unfortunately, the complexity of these new optimum filters grows with the order of the filters. Specifically, an n-th order optimum filter generally needs $(n+1)^2$ multipliers. Thus, an optimum second-order filter (n=2) requires 9 multiplications.

In the practical realization of such digital filters, either by dedicated hardware or by software in some programmable signal processors, a large number of multiplications is generally objectionable. This can be due to large multiplicative CPU time requirements and/or to the large number of multiplier coefficient memory storage requirements. In the next section, optimum second-order filters subject to practical complexity constraints are presented.

OPTIMUM CONSTRAINED COMPLEXITY SECOND-ORDER FILTER

In order to motivate the general discussions on optimum constrained complexity second-order filters, let us reconsider the canonic direct form II realization in Fig. 1 and its roundoff noise model in Fig. 1'. The total output round-off noise variance σ_T^2 is given in general by Eq.(2) and for equal processor wordlength by Eq.(3). Suppose H(z) is a narrowband low-frequency rejection filter on $[-\theta_o,\theta_o]$ with sharp transition regions. This means

$$|H(e^{i\theta})| \simeq 0 \quad , \quad |\theta| \leq \theta_o << \pi \; ; \qquad |H(e^{i\theta})| \simeq 1 \quad , \quad 0 < \theta_o < |\theta| \leq \pi \; ;$$

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{i\theta})|^2 d\theta \simeq 1 \; . \tag{5}$$

Thus, Eq.(2) becomes

$$\sigma_T^2 \simeq \sum_{i=1}^{2} (\sigma_{a_i}^2 + \sigma_{b_i}^2) \quad , \tag{6}$$

and Eq.(3) becomes

$$\sigma_T^2 \simeq 4\sigma^2 \quad . \tag{7}$$

Now, suppose we per-form a “long-hand” division of the numerator by the denominator in Eq.(l). Then,

$$H(z) = 1 + \frac{c_1 z^{-1} + c_2 z^{-2}}{1 - a_1 z^{-1} - a_2 z^{-2}} = 1 + H_1(z) \; , \tag{8}$$

$$c_1 = b_1 + a_1 \quad , \qquad c_2 = b_2 + a_2 \quad .$$

The modified canonic form given by Eq.(8) can be realized in Fig. 2. The corresponding round-off noise model is given in Fig. 2'. Then, the total output round-off noise variance in general is given by

$$\sigma_T^2 = \sum_{i=1}^{2} \sigma_{c_i}^2 + \frac{1}{2\pi} \int_{-\pi}^{\pi} |H_1(e^{i\theta})|^2 d\theta \sum_{i=1}^{2} \sigma_{a_i}^2 \quad , \tag{9}$$

and for equal processor wordlength, is given by

$$\sigma_T^2 = \sigma^2 \{2 + \frac{2}{2\pi} \int_{-\pi}^{\pi} |H_1(e^{i\theta})|^2 d\theta\} \quad . \tag{10}$$

For the case of the narrow-band low-frequency rejection filter given in Eq.(5), sharp transition regions imply that the poles are near the zeros, and $c_1 \simeq c \simeq 0$. Thus, $H_1(e^{i\theta}) \simeq 0$ for $|\theta| \geq \theta_0$. Then,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |H_1(e^{i\theta})|^2 d\theta \simeq 0 \quad .$$

Thus, Eq.(9) becomes

$$\sigma_T^2 \simeq \sum_{i=1}^{2} \sigma_{c_i}^2 \quad , \tag{11}$$

and Eq.(10) becomes

$$\sigma_T^2 \simeq 2\sigma^2 \quad . \tag{12}$$

By comparing (6) to (11), we see that the round-off noise introduced by $\sigma_{a_i}^2$ has been filtered by $H_1(z)$ in the modified canonic form and has not been filtered out by H(z) in the canonic direct form II. For the equal processor wordlength case, by comparing (7) to (12), we see that the modified form has a 50% reduction in round-off noise compared to the canonic direct form II.

In this narrow-band low-frequency rejection filter example, if originally H(z) in Eq. (1) was restricted to an elliptic digital filter, then $b_2 = 1$ and $\sigma_{b_2}^2 = 0$. Thus σ_T^2 for the canonic direct form II corresponding to (6) and (7) becomes, respectively,

$$\sigma_T^2 \simeq \sigma_{b_1}^2 + \sigma_{a_1}^2 + \sigma_{a_2}^2 \quad , \tag{13}$$

and

$$\sigma_T^2 \simeq 3\sigma^2 \quad , \tag{14}$$

while σ_T^2 for the modified form remains that of (11) and (12). By comparing (12) to (14), we see that, for the elliptic filter, the modified form has a 33% reduction in round-off noise compared to the canonic direct form II.

At the more fundamental level, the important point to note is that the canonic direct form II realization given in Fig.1 has the same complexity as that of the modified canonic form in Fig.2 . In each case, we use four multipliers, two unit delays, and four two-input adders. The significance of needing four multipliers instead of nine multipliers, as in the Mullis-Roberts case, is clear for practical implementation.

In the light of the above observations and examples, it is meaningful to find the minimum round-off noise filter subject to a practical constraint of four multipliers. Szczupak and Mitra [7] have shown that, under the restriction of four multipliers, two unit delays, four two-input adders, no products of multipliers appear in the transfer function expression, there are only 15 possible different topological realizations. These realizations are given in Figs.3 4, and 5 of this paper and correspond to those given in the same Figs. 3, 4, and 5 in [7]. The basis for the classification of all these realizations into three different figures depends on the way in which the multipliers are extracted. This rather technical detail need not concern us here.

Once the topological connections of these filters are given, then the transfer functions H(z) can be obtained readily in terms of the coefficients all α_1 α_2, α_3 and α_4. A summary of H(z) is given in Table I below and has also appeared as Table I in [7].

For any specified filter transfer function H(z), the round-off noise model discussed above can be applied to the 15 realizations. In general, for arbitrary transfer function, it is not possible to conclude the optimality of any one realization from theoretical considerations. In practice, for a specified H(z) , we need to perform the evaluation of the total output roundoff noise variances σ_T^2 for all 15 realizations and then choose the one with the minimum noise.

For the specific cases of low-frequency rejection filters, with equal processor wordlengths when the rejection bandwidth $[0,\theta_0]$ becomes arbitrarily small, simple explicit results can be obtained.

TABLE I:

Figures	*Transfer Function H(z)*
3a; 3c	$\dfrac{1 + (\alpha_1 - \alpha_2)z^{-1} + (\alpha_2 - \alpha_4)z^{-2}}{1 - \alpha_3 z^{-1} - \alpha_4 z^{-2}}$
3b	$\dfrac{1 + \alpha_1 z^{-1} + \alpha_2 z^{-2}}{1 - \alpha_3 z^{-1} - \alpha_4 z^{-2}}$
3d	$\dfrac{1 + \alpha_1 z^{-1} + (\alpha_2 - \alpha_4)z^{-2}}{1 - \alpha_3 z^{-1} - \alpha_4 z^{-2}}$
3e; 3f	$\dfrac{1 + (\alpha_1 - \alpha_3)z^{-1} + \alpha_2 z^{-2}}{1 - \alpha_3 z^{-1} - \alpha_4 z^{-2}}$
4a; 4b	$\dfrac{1 - \alpha_3 z^{-1} + (\alpha_2 - \alpha_4)z^{-2}}{1 - (\alpha_1 + \alpha_3)z^{-1} - \alpha_4 z^{-2}}$
4c; 4e	$\dfrac{1 - \alpha_3 z^{-1} + \alpha_2 z^{-2}}{1 - (\alpha_1 + \alpha_3)z^{-1} - \alpha_4 z^{-2}}$
4d; 4g	$\dfrac{1 + (\alpha_2 - \alpha_3)z^{-1} - \alpha_4 z^{-2}}{1 - \alpha_3 z^{-1} - (\alpha_1 + \alpha_4)z^{-2}}$
4f	$\dfrac{1 + \alpha_2 z^{-1} - \alpha_1 z^{-2}}{1 - \alpha_3 z^{-1} - (\alpha_1 + \alpha_4)z^{-2}}$
5a; 5b	$\dfrac{1 - \alpha_3 z^{-1} - \alpha_4 z^{-2}}{1 - (\alpha_1 + \alpha_3)z^{-1} - (\alpha_2 + \alpha_4)z^{-2}}$

THEOREM 1. Consider a narrow-band low-frequency rejection second-order digital filter H(z) denoted by Eq.(1), where the normalized total round-off noise variance ratio σ_T^2/σ^2 , for equal processor wordlengths, is obtained based on the model given in the second section of this paper. Column 1 of Table II below shows the general case of the σ_T^2/σ^2 ratio as the rejection bandwidth parameter θ_o which approaches zero with $-2 \neq b_1 \rightarrow -2$, $1 \neq b_2 \rightarrow 1$, $2 \neq a_1 \rightarrow 2$, and $-1 \neq a_2 \rightarrow -1$. Column 2 shows the general elliptic filter results with the assumptions of $-2 \neq b_1 \rightarrow -2$, $b_2 = 1$, $2 \neq a_1 \rightarrow 2$, and $-1 \neq a_2 \rightarrow -1$. Column 3 represents the special elliptic filter results with assumptions of $b_1 = -2$, $b_2 = 1$, $2 \neq a_1 \rightarrow 2$, and $-1 \neq a_2 \rightarrow -1$/

FIGURES	1 GENERAL FILTER COEFFICIENTS	2 ELLIPTIC FILTER COEFFICIENTS	3 SPECIAL ELLIPTIC FILTER COEFFS.	4 SPECIFIC REJECT ELLIPTIC FILTER
3a	2	2	2	2.18
3b	4	3	2	3.18
3c	2	2	2	2.18
3d	3	3	2	3.18
3e	3	2	2	2.18
3f	3	2	2	2.18
4a	2	2	2	2.27
4b	2	2	2	2.27
4c	3	2	2	2.26
4d	2	2	2	2.18
4e	3	2	2	2.26
4f	3	3	2	3.18
4g	2	2	2	2.18
5a	2	2	2	2.26
5b	2	2	2	2.26

TABLE II. Normalized total output round-off noise variance ratios σ_T^2/σ^2 of equal processor wordlengths for various cases of narrowband low-frequency rejection fitters.

Several comments can be made on the results presented in Table II. The σ_T^2/σ^2 ratios in Columns 1, 2 and 3 can take values of 2, 3 and 4 . In Column 4, results are given for a specific second-order elliptic digital filter when $H(\theta)$ has 1 dB ripple in the pass-band of $[0.028\pi,\pi]$, and has a rejection of greater than -39 dB on $[0,0.004\pi]$. The filter coefficients for this specific filter are $b_1 = -1.99999$, $b_2 = 1$, $a_1 = 1.91016$, and $a_2 = -.91699$. For this example, a minimum σ_T^2/σ^2 ratio of 2.18 is obtained. From all the results in Table II, it seems that the realizations given by 3a and 3c are optimum in the sense of minimum round-off noise generation for low-frequency rejection purposes. It is interesting to note that realization 3a is indeed the modified canonic form presented in Fig. 2. The canonic direct form II, which is realization 3b, is not optimum for low-frequency rejection filtering purposes.

CONCLUSION

In this paper, we studied the minimum round-off noise second-order digital filtering problem under the practical complexity constraints of four multipliers, two delays, and four two-input adders. For purposes of narrow low-frequency rejection filtering, explicit optimum realizations are obtained. Of course, all the results obtained here are based on the

simple linear random model where all round-off errors are uncorrelated with the data. In principle, these assumptions lead to optimistic round-off noise variances. Considerable simulations have been done on these 15 realizations. While the simulated round-off noise variances are indeed larger than that evaluated from the simple analytical model, the relative ordering of the advantages of the realizations appears to be still preserved. That is, for low-frequency rejection filtering applications, simulation results still indicate realizations 3a and 3c to be optimum. More detailed results on minimum round-off noise digital filter under practical constraints will be presented in later publications.

ACKNOWLEDGMENTS

The author wishes to thank Mr. T. Brukiewa and Mr. C. Stirman of HAC for technical discussions and support of this work. The author also wishes to thank Mr. F. Pollara-Bozzola of UCIA for technical discussions and computation of the results in Column 4 of Table II.

REFERENCES

1. Oppenheim, A.V., and R.W. Schafer, *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

2. Rabiner, L.R., and B. Gold, *Theory and Application of Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

3. Kaiser, J.F., "Some Practical Considerations in the Realization of Linear Digital Filters," *Proc.3rd Ann. Allerton Conf. on Circuits and System Theory*, 1965, pp.621-633.

4. Jackson, L.B., "On the Interaction of Round-Off Noise and Dynamic Range in Digital Filters," *Bell System Tech.J., 49* (1970) pp. 159-184.

5. Jackson, L.B., "Round-Off Noise Analysis for Fixed-Point Digital Filters Realized in Cascade or Parallel Form," *IEEE Trans. Audio and Vectroacoustics, AU-18* (1970) pp. 107-122.

6. Mullis, C.T., and R.A. Roberts, "Synthesis of Minimum Round-Off Noise Fixed Point Digital Filters," *IEEE Trans. on Circ. and Syst., CAS-23* (1976) pp. 551-562.

7. Szczyupak, J., and S.K. Mitra, "Digital Filter Realization Using Successive Multiplier-Extraction Approach," *IEEE Trans.Acoust., Speech, Sig.Proc., ASSP-23* (1975) pp.235-239.

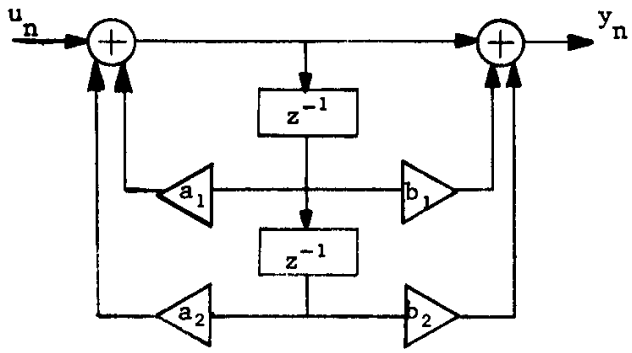

FIG. 1. Canonic Direct Form II Realization.

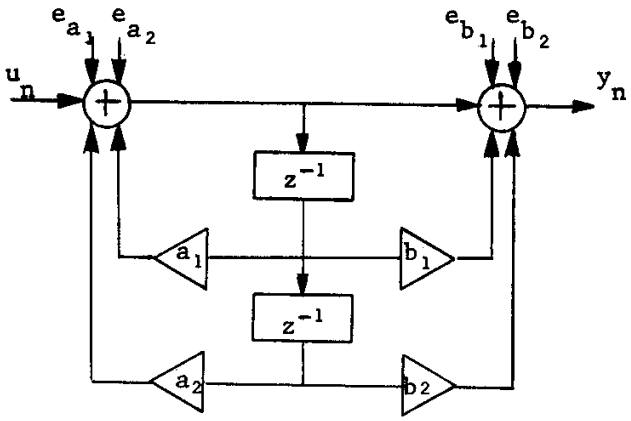

FIG.1'. Round-off Noise Model for Figure 1.

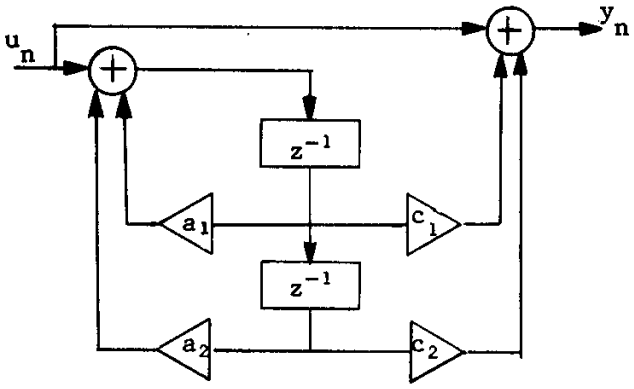

FIG.2. Modified Canonic Form Realization.

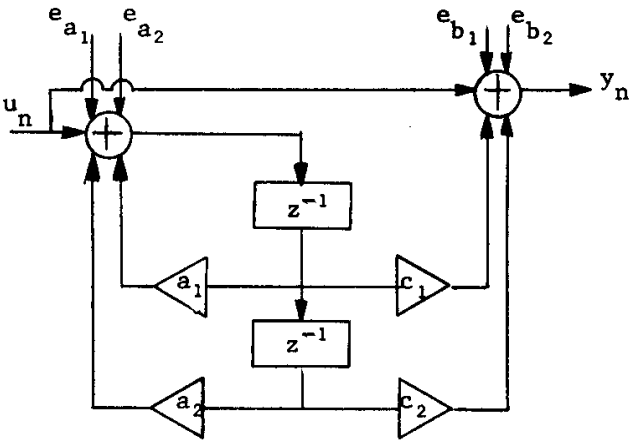

FIG.2'. Round-Off Noise Model for Figure 2.

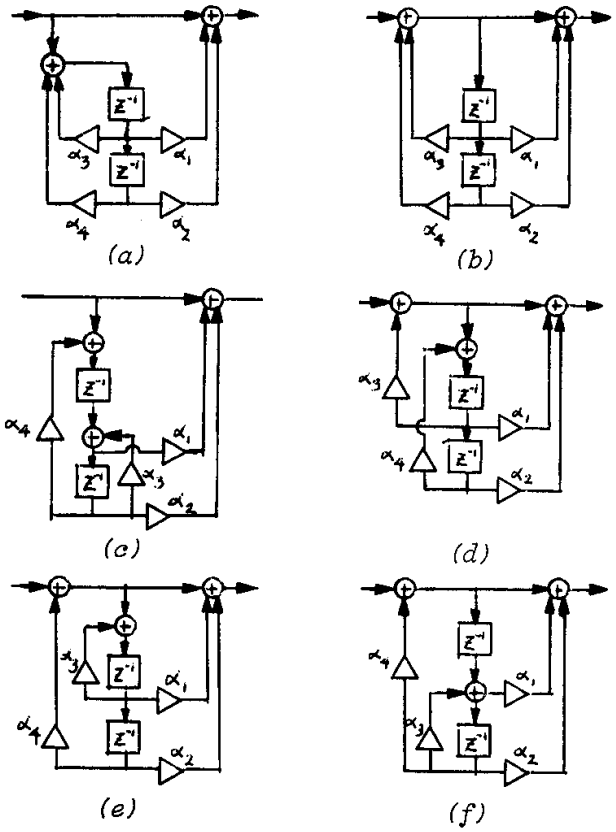

FIG. 3. Constrained Complexity Second-Order Filter Realizations -- (Part 1).

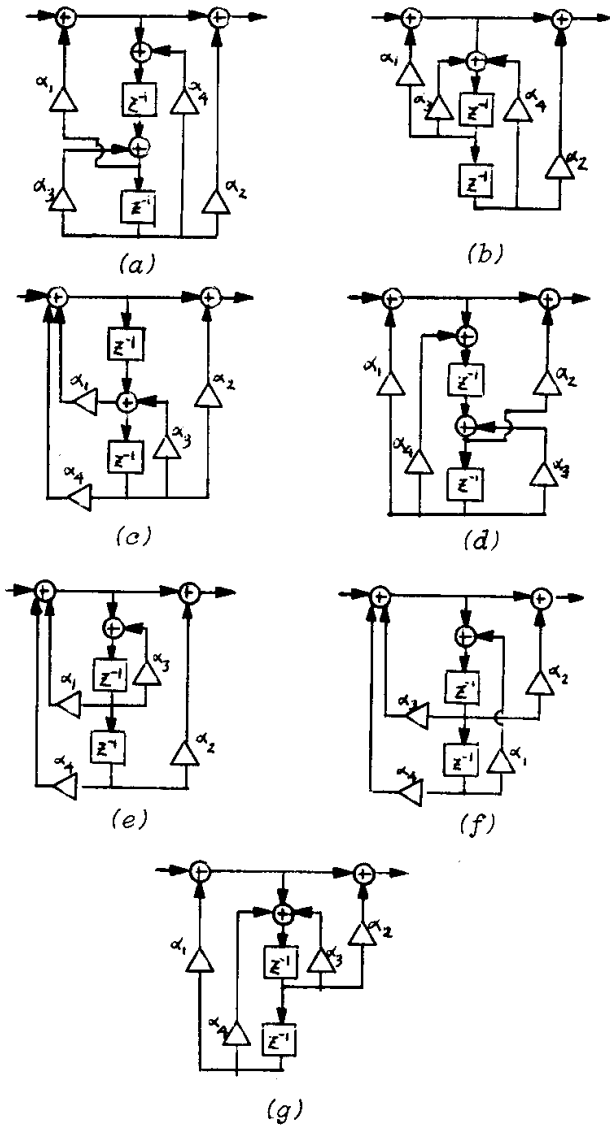

FIG.4. Constrained Complexity Second-Order Filter Realizations -- (Part 2).

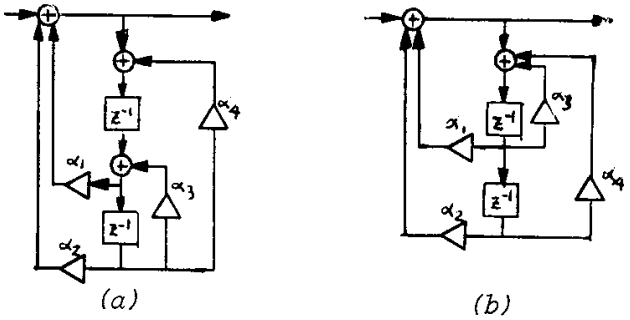

FIG. 5. Constrained Complexity Second-Order Filter Realizations --(Part 3).