TABLE VIII
NOISE GAIN IN DECIBELS OF NEAR OPTIMAL ORDERINGS, 24TH-ORDER NONRECURSIVE FILTER, EACH WITH 10 RANDOM STARTS

Interchanging all pairs	Interchanging pairs 2 apart	Interchanging pairs 3 apart
13.01	13.01	13.03
13.16	13.16	13.51
14.42	14.72	15.04
14.09	15.88	16.34
13.23	14.33	14.55
13.46	13.64	15.84
12.96	14.13	15.27

only a modest amount of computer time even for very high-order filters.

These two advantages make this method more desirable than existing methods which require either expertise or significant computation time for moderately high-order filters.

The examples presented in Sections II and III also point out the importance of using an optimization procedure as the "noise gain" for a bad assignment may be up to 100 000 times higher than that for a good assignment. Since the noise gain is directly related to the number of bits used in hardware implementation of such filters, the undesriable consequences of a bad assignment are clear.

Note: The authors will gladly send to persons interested listings of the programs that have been referred to in this paper.

REFERENCES

[1] J. B. Knowles and R. Edwards, "Effects of a finite-word-length computer in a sampled-data feedback system," *Proc. Inst. Elec. Eng.* (London), vol. 112, pp. 1197–1207, June 1965.

[2] L. B. Jackson, "Roundoff-noise analysis for fixed-point digital filters realized in cascade or parallel form," *IEEE Trans. Audio Electroacoust. (Special Issue on Digital Filtering)*, vol. AU-18, pp. 107–122, June 1970.

[3] W. S. Lee, "Optimization of digital filters for low roundoff noise," in *Proc. 1973 Int. Symp. Circuit Theory*, Toronto, Ont., Canada, pp. 381–383.

[4] S. Y. Hwang, "On optimization of cascade fixed-point digital filters," *IEEE Trans. Circuits and Systems* (Lett.), vol. CAS-21, pp. 163–166, Jan. 1974.

[5] E. Leuder, "Minimizing the roundoff noise of digital filters by dynamic programming," presented at the 1974 Arden House Workshop on Digital Signal Processing, Harriman, N.Y., Jan. 14–17, 1974.

[6] S. Reiter and G. Sherman, "Discrete optimizing," *J. Soc. Ind. Appl. Math.*, vol. 13, pp. 864–878, Sept. 1963.

[7] S. Lin, "Computer solutions of the traveling salesman problem," *Bell Syst. Tech. J.*, vol. 44, pp. 2245–2270, Dec. 1965.

[8] K. Steiglitz and P. Weiner, "Some improved algorithms for computer solution of the traveling salesman problem," in *Proc. 6th Annu. Allerton Conf. Circuits and Systems Theory*, 1968, pp. 814–821.

[9] K. J. Astrom, E. I. Jury, and R. G. Agniel, "A numerical method for the evaluation of complex integrals," *IEEE Trans. Automat. Contr.* (Short Papers), vol. AC-13, pp. 468–471, Aug. 1970.

[10] L. B. Jackson, "An analysis of roundoff noise in digital filters," Sc.D. dissertation, Dep. Elec. Eng., Stevens Inst. Technol., Hoboken, N.J., 1969.

[11] A. Peled and B. Liu, "A new hardware realization of digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 456–462, Dec. 1974.

[12] D. S. K. Chan and L. R. Rabiner, "Theory of roundoff noise in cascade realizations of finite impulse response digital filters," *Bell Syst. Tech. J.*, vol. 52, pp. 329–345, Mar. 1973.

[13] —, "An algorithm for minimizing roundoff noise in cascade realizations of finite impulse response digital filters," *Bell Syst. Tech. J.*, vol. 52, pp. 347–385, Mar. 1973.

[14] J. McClellan, T. W. Parks, and L. R. Rabiner, "A computer program for designing optimum FIR linear phase digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 506–526, Dec. 1973.

Roundoff Noise Characteristics of a Class of Orthogonal Polynomial Structures

JOHN D. MARKEL, MEMBER, IEEE, AND AUGUSTINE H. GRAY, JR., MEMBER, IEEE

Abstract—**The roundoff noise characteristics of three digital filter structures derived from the theory of orthogonal polynomials are studied and compared to the two standard forms, the direct and parallel forms. Both theoretical and experimental results are presented. It is shown that in terms of roundoff noise, the worst of the orthogonal filter forms (the two-multiplier lattice form) is superior to the direct form, and that the best (the normalized or ladder form) is nearly equal to or superior to the parallel form. The normalized form is shown to be vastly superior to the parallel form when filters with clustered poles are designed.**

Manuscript received June 8, 1974; revised December 17, 1974. This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Office of Naval Research under Contract N00014-73-0221.

J. D. Markel is with the Speech Communications Research Laboratory, Inc., Santa Barbara, Calif. 93109.

A. H. Gray, Jr. is with the Speech Communications Research Laboratory, Inc., Santa Barbara, Calif. 93109 and the Department of Electrical Engineering and Computer Science, University of California, Santa Barbara, Calif. 93106.

I. INTRODUCTION

IT is a well-known fact that mathematically equivalent digital filter structures may produce very different results when implemented with finite word length (FWL) arithmetic.[1] The differences are due to the arithmetic roundoff errors, truncation of filter coefficient accuracy, and input signal

[1] For the purposes of this study FWL arithmetic will be used to mean 2's complement fixed-point arithmetic.

quantization. For nontrivial filters, the roundoff errors are by far the most important contributor to the differences between the ideal (infinite word length) and FWL implementation.

Jackson and others [1]-[5] have developed the theory of roundoff noise effects in digital filters and shown very accurate predictions of the roundoff noise characteristics for a wide variety of filter types (low-pass, bandpass, etc.) under the assumption of rounding arithmetic. The basic filter structures (direct, parallel, and cascade) have been studied in detail. Jackson's conclusion [1] was that a parallel structure exists that is usually superior to the best ordered configuration of a cascade structure in the sense of minimizing the $\mathcal{L}_p$ norm $\|N(\theta)\|_p$ of the roundoff noise spectrum $N(\theta)$ where

$$\|N(\theta)\|_p = \left[\frac{1}{2\pi}\int_0^{2\pi} |N(\theta)|^p d\theta\right]^{1/p}. \tag{1}$$

θ is the normalized frequency $2\pi f/F_s$, F_s is the sampling frequency, and f is the unnormalized frequency.

Considerable research has been applied to the problem of finding alternate structures to minimize $\|N(\theta)\|_p$. A particular class of structures that has gained recent interest is the two input-two output form, generally implemented as a digital ladder or lattice filter [6]-[11]. It has been suggested that for at least one ladder form, the coefficient sensitivity was several bits less than for a cascade implementation [12]. Effects of roundoff noise based upon total filter implementation in a FWL format have not appeared in the literature.

The purpose of this paper is to present results from a study of the roundoff noise characteristics of a subclass of two input-two output structures, namely digital filter structures derived from the theory of orthogonal polynomials. For comparative purposes the commonly accepted worst and best of the standard forms, namely the direct and the 1-P parallel form [3], are included. This study will focus upon the L_2 norm ($p = 2$) measure for roundoff noise computation due to its mathematical tractability and physical significance. Effects of input quantization and coefficient quantization only are not considered here.

A. Results

1) Theoretical roundoff noise equations have been derived for the orthogonal forms and shown to be in excellent agreement with results obtained by simulation with rounding arithmetic.

2) In terms of roundoff noise characteristics due to FWL implementation, the worst of the orthogonal forms is in general the two-multiplier form [11]. Under certain conditions the one-multiplier forms with constant sign parameters are worse than the two-multiplier form. The two-multiplier form appears to have similar (although always superior) FWL characteristics to the direct form.

3) The optimal one-multiplier form [11] is superior to the two-multiplier and constant one-multiplier forms in general. In one series of elliptic low-pass filter designs, it mimicked the features of the parallel form with at most 6 dB or 1 bit of degradation over a wide range of bandwidths.

4) A newly developed orthogonal polynomial structure, the normalized form [13], has been shown to have nearly equal to or superior FWL characteristics to all of the other polynomial or standard forms. It appears to have vastly superior FWL characteristics over any other known filter structure for the implementation of tightly clustered poles. In a particular digital bandpass filter example the roundoff noise for the normalized form was shown to be 61 dB, or about 10 bits less than that of the parallel form.

A very important property of these orthogonal polynomial structures is that they are completely general, i.e., any stable direct form digital filter can be efficiently transformed using simple recursive relations [14].

B. Approach

Although a large number of structures can be derived from the theory of orthogonal polynomials, three particular structures will be concentrated upon in this paper. The first structure (referred to as the two-multiplier form) is of interest because it is the most direct derivation from the theory. Manipulation of this form leads to a one-multiplier form which has the property that it is canonic in multiplies and delays, and allows for a "sign-parameter" introduction for finding optimal scaling within the basic structure. Further manipulation leads to the "normalized" form which has the important property that the norms at every node in the all-pole portion are precisely unity. For comparative purposes, the direct form and parallel form will also be used to implement each of the filters.

Both theoretical and experimental procedures will be used for determining the FWL characteristics of these filters. The parameter of major interest here will be the normalized noise roundoff variance or noise figure

$$\nu = \sigma_n^2/\sigma_e^2 \tag{2}$$

where σ_n^2 is the output roundoff noise variance defined by

$$\sigma_n^2 = \frac{1}{2\pi}\int_0^{2\pi} |N(\theta)|^2 d\theta \tag{3}$$

with $|N(\theta)|^2$ being the output roundoff noise spectrum. The term σ_e^2 defines the variance of the quantization step size b, and is given by

$$\sigma_e^2 = \frac{1}{12}2^{-2b} \tag{4}$$

where $\beta = b + 1$ is the fixed-point computer word length with b bits for magnitude and one bit for sign.

The theoretical procedure will follow Jackson's approach except that the complete filter will be scaled to insure against overflow in a fixed-point (fractional) representation.

After presenting the theoretical derivations and experimental procedures in more detail, a number of digital low-pass and bandpass filter examples will be discussed.

II. Theoretical Procedures

A. Roundoff Noise Analysis for Direct and Parallel Forms

Presentation of fixed-point roundoff noise comparisons among the orthogonal polynomial filter forms and the stan-

dard forms requires two modifications to Jackson's results. First, in general, the filter coefficients must be scaled along with the input, and second, the tap gains (coefficients used for implementing the filter zeros) must be scaled so that they are as large as possible while lying within the fixed-point format and insuring against overflow at the output.

Direct Form: The direct form implementation is given by

$$G(z) = P(z)/A(z) \tag{5}$$

where

$$P(z) = \sum_{i=0}^{M} p_i z^{-1} \tag{6}$$

and

$$A(z) = \sum_{i=0}^{M} a_i z^{-1}, \text{ with } a_0 = 1. \tag{7}$$

To insure against overflow in the internal computations of the all-pole portion of the filter for a unit sample[2] input (or a scaled norm of unity in Jackson's work [1]), the input must be scaled down by[3]

$$\|F_u\| = \|1/A(z)\| \tag{8}$$

where $F_u(z)$ is the transfer function from the input x_n to the output of the all-pole portion u_n. A recursive procedure for evaluating $\|F_u\|$ is given elsewhere [11], [14].

In general, the denominator coefficient values are bounded only by the binomial distribution. Since this is generally a rather poor bound (except in tightly clustered pole situations), scale factors s_i are obtained from the coefficients as

$$s_i = [\log_2 \{|a_i|, \quad i = 1, 2, \cdots, M\}] \tag{9}$$

where $[x]$ is the largest integer less than x. The scaled, fractional coefficients are then given by

$$a_i' = a_i 2^{-s_i} \quad i = 1, 2, \cdots, M. \tag{10}$$

In the direct form, the scaling parameters s_i are applied to the all-pole portion. Only a constant scaling of the tap parameter set $\{p_i\}$ is possible since the norms measured from x_n to each node all equal $\|F_u\|$. The scaled tap parameters are given by

$$v_i' = p_i/w \tag{11}$$

where w is computed using the scaling algorithm presented elsewhere [14]. The scaled implementation of the direct form filter is shown in Fig. 1. As long as the output of the input summer is bounded by unity, overflows on partial sums do not affect the final result in 2's complement arithmetic. The multiplication of u_{n-i} and a_i' must therefore precede the shift factor 2^{s_i} since the shifting may result in overflow of a partial sum. The scaled output y_n' must be multiplied by the factor $\|F_u\|w$ to obtain the correct overall filter gain. The roundoff

[2] More precisely, for β-bit 2's complement arithmetic, using fractional notation, input is $1 - 2^{-(\beta-1)}$

[3] For notational simplicity, the subscript in the norm $\|\cdot\|_p$ is dropped since only the $p = 2$ case is considered here.

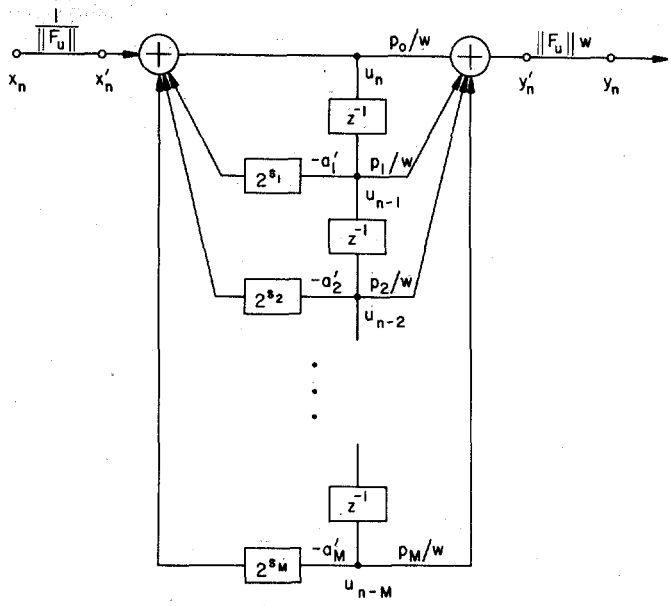

Fig. 1. Fixed-point implementation for direct form.

noise figure (2) can be written by inspection of Fig. 1 as

$$\nu = \|F_u\|^2 \left\{ \left(1 + \sum_{i=1}^{M} 2^{2s_i} \right) \|F_y\|^2 + (M+1)w^2 \right\} \tag{12}$$

where $\|F_y\|^2$ is the norm square of the total filter $G(z)$ or, equivalently, the sum of the squares of the unit sample response of $G(z)$.

Each of the multiplications by a_i' results in an equivalent independent noise source at the input to $G'(z)$ (the scaled version of $G(z)$ having variance $2^{2s_i}\sigma_e^2$). The output variance contribution is obtained by multiplying this term by the norm square of the scaled filter $G'(z)$, $\|F_y\|^2/w^2$ times the gain squared from y_n' to y_n, $\|F_u\|^2 w^2$. The input scaling from x_n to x_n' results in an output contribution of σ_e^2 times $\|F_y\|^2/w^2$ times $\|F_u\|w^2$. The output noise contribution due to each of the $M + 1$ tap parameter multiplications is simply σ_e^2 times $\|F_u\|^2 w^2$.

Parallel Form: By performing a partial fraction expansion, (5) can be equivalently rewritten for M even and nonrepeated poles as

$$G(z) = v_{00} + \sum_{l=1}^{M} c_l/(1 - z_l z^{-1})$$

$$= v_{00} + \sum_{l=1}^{M/2} \frac{v_{l0} + v_{l1} z^{-1}}{1 + \beta_{l1} z^{-1} + \beta_{l2} z^{-2}} \tag{13}$$

where

$$v_{l0} = 2 \text{ Re } (c_l)$$
$$v_{l1} = -2 \text{ Re } (c_l z_l^*)$$
$$\beta_{l1} = -2 \text{ Re } (z_l)$$
$$\beta_{l2} = |z_l|^2$$

and Re $(\cdot)$ denotes the real part of $(\cdot)$. If $\|F_u^l\|$ is used to denote the norm of the all-pole portion of the lth second-order section, then the step-down procedure described elsewhere [11], [14] can be used to perform the evaluations

$$\|F_u\|^2 = \alpha_{l0} \tag{14}$$

and

$$\|F_y^l\|^2 = \left\| \frac{v_{l0} + v_{l1}\, z^{-1}}{1 + \beta_{l1} z^{-1} + \beta_{l2} z^{-2}} \right\|^2 \tag{15}$$

as

$$\|F_y^l\|^2 = (v_{l0} - v_{l1}\, k_{l0})^2\, \alpha_{l0} + v_{l1}^2\, \alpha_{l1}, \tag{16}$$

where

$$\begin{aligned} k_{l0} &= \beta_{l1}/(1 + k_{l1}) \\ k_{l1} &= \beta_{l2} \\ \alpha_{l0} &= \alpha_{l1}/(1 - k_{l0}^2) \\ \alpha_{l1} &= 1/(1 - k_{l1}^2). \end{aligned}$$

To insure against overflow within the all-pole portion of filter section l, the section input must be scaled down by $\|F_u\|$ for a unit sample input. Since $|\beta_{l2}| < 1$ and $|\beta_{l1}| < 2$ for all l, the scaled coefficients will be defined by

$$\begin{aligned} \beta'_{l2} &= \beta_{l2} \\ \beta'_{l1} &= \beta_{l1}/2. \end{aligned} \tag{17}$$

Although in some instances the scaling of β_{l1} by 2 is unnecessary, this procedure will be followed throughout for simplicity. For the cases of most interest later, namely clustered pole conditions, the scaling is necessary. Since $\|F_u^l\|$ will generally be different for every section, the reciprocal factor must be included within the scaled taps as

$$\begin{aligned} v'_{l0} &= v_{l0}\|F_u^l\|/w \\ v'_{l1} &= v_{l1}\|F_u^l\|/w \end{aligned}$$

and

$$v'_{00} = v_{00}/w \tag{18}$$

for $l = 1, 2, \cdots, M/2$, where w is computed by defining the maximum tap parameter value $v_{\max}$ as

$$v_{\max} = \max\{V_{00}, V_{l0}, V_{l1}, l = 1, 2, \cdots, M/2\} \tag{19}$$

in the fixed-point scaling algorithm [14]. A block diagram of the fixed-point implementation of the parallel form is shown in Fig. 2. The roundoff noise figure ν can be obtained from this figure as

$$\nu = 6\left\{\sum_{l=1}^{M/2} \|F_y^l\|^2\, \|F_u\|^2\right\} + (M+1)w^2. \tag{20}$$

If the filter order is odd, or if any pole pairs are repeated, this development must be slightly modified.

Orthogonal Polynomial Forms: By making use of Mason's gain formula for flowgraphs [15], [16] the roundoff noise figure γ for the three-orthogonal polynomial structures can be derived. The results are given below.

Two-multiplier form:

$$\gamma_T = w^2\, \|\bar{F}\|^2\, \{M + 1 + \sum_{m=0}^{M} (\|V_{m+1}^-\|^2 + \|V_{m+1}^+\|^2)\}. \tag{21a}$$

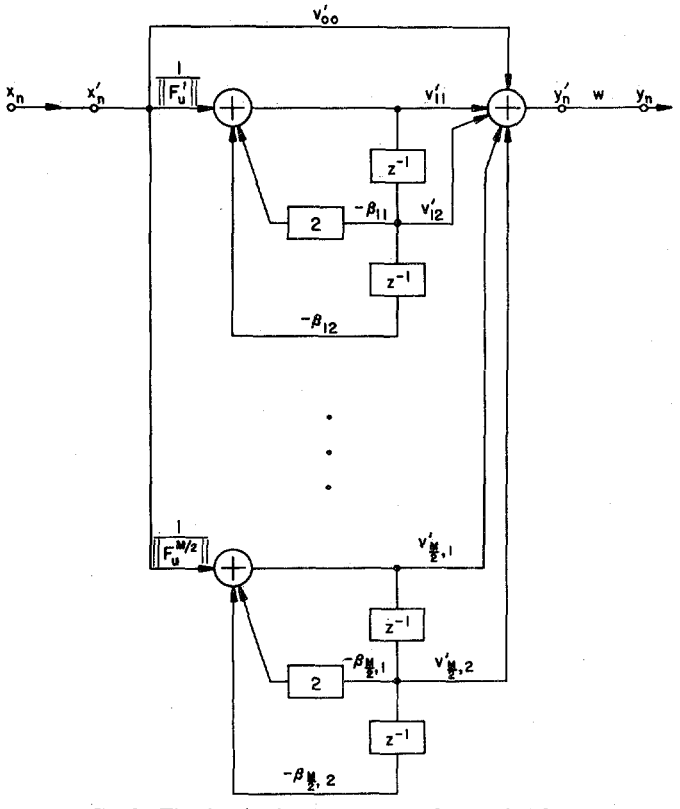

Fig. 2. Fixed-point implementation for parallel form.

One-multiplier form:

$$\gamma_O = w^2\, \|\bar{F}\|^2\, \{M + 1 + \sum_{m=0}^{M-1} (\|V_{m+1}^- + \epsilon_m V_m^+\|^2)\}. \tag{21b}$$

Normalized form:

$$\gamma_N = w^2\, \|\bar{F}\|^2\, \{M + 1 + 2\sum_{m=0}^{M-1} (\|V_{m+1}^-\|^2 + \|V_m^+\|^2)\} \tag{21c}$$

where $\|\bar{F}\|$ defines the maximum norm within the filter, V_m^+ defines the transfer function from node m in the upper path of the filter (denoted by m^+) to the output of the scaled filter, and V_m^- defines the transfer function from node m in the lower feedback path of the filter (denoted by m^-) to the output of the scaled filter.

Due to the number of feedback paths in these structures it is a nontrivial task to compute V_m^+ and V_m^-. First the concept of equivalent sources will be applied to pull the noise sources out of each internal filter section. Next a right- and left-transfer function will be introduced for obtaining node to node transfer functions. Mason's gain formula for flowgraphs is then applied to compute the transfer functions from each noise source to the filter output.

Equivalent Sources: In the two-multiplier model there are two multiplications per section. If these multiplications insert independent noise terms into the section m (whose input–output relations are denoted as $G_m(z)$), the effect is equivalent

to inserting noise sources outside the block at the nodes $(m+1)^-$ and $(m+1)^+$, as shown in Fig. 3(a).

In the one-multiplier model the single internal noise source appears at both outputs of the section. The same effect can be realized by taking a single noise source η_1, and inserting it into the nodes m^+ and $(m+1)^-$, provided the noise inserted into the nodes is negated for one of the nodes when the sign parameter for that block is negative, as indicated in Fig. 3(b).

In the normalized form, there are four multipliers and hence four noise sources within each block. The same effect can be realized with external noise sources if two independent sources are fed into each of the output nodes m^+ and $(m+1)^-$, as shown in Fig. 3(c).

The noise analysis can thus proceed by treating the filter sections as having no internal noise sources. All noise sources are introduced externally at the appropriate nodes. The next step is to obtain node-to-node transfer functions so that the effect of the noise sources from each external node to the output can be evaluated. The internal structure of the sections labeled $G_m(z)$ for each of the three forms is shown elsewhere [13], [14].

Left- and Right-Transfer Functions: In order to obtain node-to-node transfer functions, equivalent branches are found to represent right and left portions of the filter. This procedure is used to simplify the flowgraphs that will be obtained later. The filter is physically separated at the nodes n^- and n^+, so that sections $m = 0, 1, \cdots, n-1$ form the right portion and sections $m = n, n+1, \cdots, M-1$ form the left portion. Starting with the right portion, a right-transfer function[4] R_n from an input at node n^+ and an output at n^- is defined. In a similar manner a left-transfer function, L_n, is defined from an input at node n^- to an output at node n^+.

By definition, $R_0 = 1$ accounts for the boundary condition to the right of the 0th block, and $L_M = 0$ accounts for the lack of feedback to the left of the Mth block. If the z transform values at the nodes m^+ and m^- are defined by X_m^+ and X_m^-, respectively, then one can express all of the orthogonal polynomial filter forms by the equations

$$[\pi_m/\pi_{m+1}]\, X_{m+1}^+ = X_m^+ + k_m z^{-1} X_m^- \tag{22a}$$

$$[\pi_m/\pi_{m+1}]\, X_{m+1}^- = k_m X_m^+ + z^{-1} X_m^- \tag{22b}$$

where the π_m are the pi-parameters which are a function of the different filter structures as follows:

$$\pi_m = 1 \qquad \text{two-multiplier} \tag{23a}$$

$$\pi_m/\pi_{m+1} = 1 + \epsilon_m k_m \qquad \text{one-multiplier} \tag{23b}$$

$$\pi_m = 1/(1-k_m^2)^{1/2} \qquad \text{normalized form.} \tag{23c}$$

A flowgraph representation of section m based upon (22), is shown in Fig. 4.

In terms of the node values from Fig. 4, the right-transfer function R_m is

$$R_m = X_m^-/X_m^+. \tag{24}$$

[4]For notational simplicity, the explicit relationship to transfer functions as z transforms will be omitted except where necessary. For example, $R_n = R_n(z)$ and $A_m = A_m(z)$.

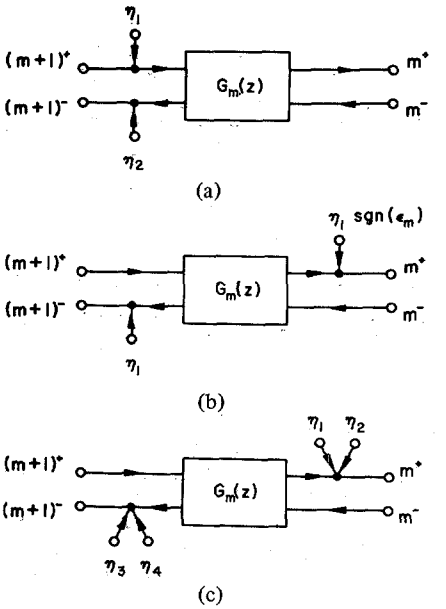

Fig. 3. Equivalent noise models: (a) two-multiplier, (b) one-multiplier, (c) normalized form.

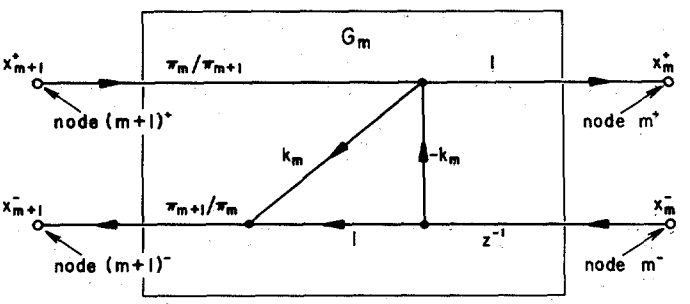

Fig. 4. Flowgraph representation of section m for the orthogonal polynomial structures.

Dividing (22b) by (22a) and substituting (24) results in a recursive evaluation for R_n as

$$R_{m+1} = \frac{k_m + z^{-1} R_m}{1 + k_m z^{-1} R_m} \tag{25}$$

where $m = 0, 1, \cdots, n-1$ and $R_0 = 1$. The left-transfer function L_m is defined in terms of the node variables as

$$L_m = X_m^+/X_m^-. \tag{26}$$

Dividing (22a) by (22b), substituting (26), and solving for L_m, results in a recursive evaluation for L_n as

$$L_m = -z^{-1} \frac{k_m - L_{m+1}}{1 - k_m L_{m+1}} \tag{27}$$

where $m = M-1, M-2, \cdots, n$ and $L_M = 0$. It is known from previous work that if X_m^+ is the input to the m section filter $1/A_m$ at node m^+, the output seen at node m^- is given by $zB_m X_m^+/A_m$ where A_l and B_l, $l = 0, 1, \cdots, M$ are each orthogonal polynomials [11]. Therefore, in terms of the orthogonal polynomial relationships B_m and A_m,

$$R_m = zB_m/A_m \tag{28a}$$

with $A_0 = zB_0 = 1$ so that $R_0 = 1$. The left-transfer function is defined as the polynomial relation

$$L_m = -P_m/Q_m \tag{28b}$$

with $P_M = 0$ so that $L_M = 0$. The denominator at $m = M$ is arbitrarily set to $Q_M = 1$. Substitution of (28a) into (25) gives the known recursive relations [11]

$$zB_{m+1} = k_m A_m + B_m \tag{29}$$

$$A_{m+1} = A_m + k_m B_m \tag{30}$$

for $m = 0, 1, \cdots, n-1$. A similar substitution of (28b) into (27) gives the results

$$zP_m = k_m Q_{m+1} + P_{m+1} \tag{31}$$

$$Q_m = Q_{m+1} + k_m P_{m+1} \tag{32}$$

for $m = M-1, M-2, \cdots, n$.

It can be noted from the recursion relations that both A_n and zB_n must be nth-order polynomials. The leading coefficient of A_n, the coefficient of z^0, is unity. Both Q_n and zP_n will be polynomials of order $M-(n+1)$ (except for $n = M$ in which case $Q_M = 1$ and $zP_M = 0$). The leading coefficient of Q_n is also unity.

The above relations can be utilized to recursively compute the left- and right-transfer functions L_n and R_n.

Denominator Determinant: The procedure for obtaining the transfer function between two nodes, where the nodes in question are at μ^+, μ^- and l^+, l^-, assuming $\mu \geqslant l$, is now presented. All sections to the right of the l^+, l^- nodes will be replaced by the right-transfer function R_l and those to the left of the μ^+, μ^- nodes replaced by the single left-transfer function L_μ. The net result is illustrated in Fig. 5. The noise-free filter sections are defined by (22) and shown as a flowgraph form in Fig. 4.

In order to express the transfer functions between any of the nodes using Mason's gain formula, it is necessary to evaluate both a numerator and a denominator term. The denominator can be expressed in terms of the loops of the flowgraph, and is often called the denominator determinant, for it is the determinant that results if the equations represented by the flowgraph are solved using Cramer's rule. The flowgraph approach to evaluating this determinant is to take one minus the sum of all the loop gains plus the sum of the products of all of the loop gain pairs for nontouching loops, etc. By combining Fig. 4 with Fig. 5 it is seen that each of the loop gains will consist of either a constant times a power of z^{-1}, or a constant times a power of z^{-1} times R_l or L_l or both. Thus each term of the denominator determinant must be either a polynomial, or a polynomial divided by A_l, Q_μ, or $A_l Q_\mu$ and thus the final determinant $\Delta_{l,\mu}$ for Fig. 5 is of the form

$$\Delta_{l,\mu} = [\text{polynomial}]/A_l Q_\mu, \tag{33}$$

where the polynomial in the numerator has a leading coefficient of unity since the leading coefficients of the polynomials A_l and Q_μ are each unity.

The zeros of $\Delta_{l,\mu}$ must equal the poles of the system, which are the roots of A_M, and thus, the numerator polynomial must itself equal A_M. As a result one obtains the important relation

$$\Delta_{l,\mu} = A_M/[A_l Q_\mu]. \tag{34}$$

Transfer Functions: To obtain any of the transfer functions

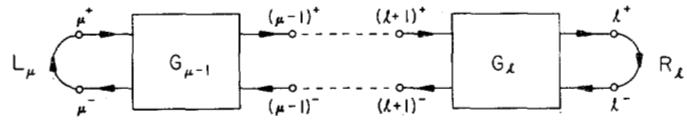

Fig. 5. Flowgraph representation of the filter $1/A(z)$ with a right- and left-transfer function.

between two nodes μ^+, μ^- and l^+, l^-, one needs only the numerator expression, for the denominator determinant is given by (34). The numerator is obtained by finding direct paths between the nodes (of which there is only one as seen from Figs. 4 and 5) multiplied by the modified determinant of Mason's flowgraph procedure. The modified determinant is formed in the same manner as the determinant; however it excludes all loops which touch the path. Since all loops touch the paths between the nodes in question, all modified determinants equal unity. Thus all that is needed to obtain the appropriate transfer functions is to find the direct path gains, and then divide by $\Delta_{l,\mu}$ of (34).

Since the final output of the filter is obtained only from the lower nodes m^-, $m = 0, 1, \cdots, M$, it is only necessary to compute all possible transfer functions to these nodes. Notationally $T(\mu^-, l^-)$ defines, for example, the transfer function from the left lower node at μ^- to the right lower node l^-. Application of Mason's gain formula from Figs. 4 and 5 gives the following results by inspection:

1)

$$T(l^+, l^-) = R_l/\Delta_{l,l} \tag{35a}$$

$$T(l^-, l^-) = 1/\Delta_{l,l}. \tag{35b}$$

2)

$$T(\mu^+, l^-) = \left[\prod_{m=l}^{\mu-1} \pi_m/\pi_{m+1}\right] R_l/\Delta_{\mu,l}$$

$$= \frac{\pi_l z B_l A_l Q_\mu}{\pi_\mu A_l A_M}$$

$$= \frac{\pi_l z B_l Q_\mu}{\pi_\mu A_M}. \tag{36a}$$

$$T(\mu^-, l^-) = L_\mu T(\mu^+, l^-)$$

$$= \frac{-\pi_l z B_l P_\mu}{\pi_\mu A_M}. \tag{36b}$$

3)

$$T(l^-, \mu^-) = \left[\prod_{m=l}^{\mu-1} z^{-1}(1-k_m^2)\,\pi_m/\pi_{m+1}\right]/\Delta_{l,\mu}$$

$$= \pi_l z^{-(l-\mu)}\left[\prod_{m=\mu}^{l-1}(1-k_m^2)\right]\frac{A_\mu Q_l}{\pi_\mu A_M}. \tag{37a}$$

$$T(l^+, \mu^-) = R_l T(l^-, \mu^-)$$

$$= \pi_l z^{-(l-\mu)}\left[\prod_{m=\mu}^{l-1}(1-k_m^2)\right]\frac{z B_\mu Q_l}{\pi_\mu A_M}. \tag{37b}$$

The transfer function V_m^+ from a particular node, m^+, to the output node requires computing the individual transfer functions from m^+ to each of the lower nodes multiplied by the appropriate tap gain and then summing, i.e.,

$$V_m^+ = \sum_{i=0}^{M} \hat{v}_i \, T(m^+, i^-). \tag{38a}$$

Similarly, for the node m^-,

$$V_m^- = \sum_{i=0}^{M} \hat{v}_i \, T(m^-, i^-). \tag{38b}$$

Numerically, the transfer functions $T(m^+, i^-)$ and $T(m^-, i^-)$ are evaluated from 1) if $m = i$, 2) if $m > i$, and 3) if $m < i$, with the proper signs. Now, by inspection of Fig. 3(a), single noise sources exist at nodes $(m+1)^+$ and $(m+1)^-$, $m = 0, 1, \cdots, M-1$, with noise contributions given by $\sigma_e^2 \|V_{m+1}^+\|^2$ and $\sigma_e^2 \|V_{m+1}^-\|^2$, respectively. A final multiplication of each term by $w^2 \|\bar{F}\|^2$ gives the unscaled noise components for comparing to the original noiseless filter. Each tap gain multiplication also contributes noise in the amount of $\sigma_e^2 w^2 \|\bar{F}\|^2$ so that the final noise output γ_T for the two-multiplier forms, normalized by σ_e^2, results in (21a).

Since the same noise term η_1 appears at both nodes m^+ and $(m+1)^-$ in the one-multiplier form of Fig. 3(b), the norm of the combined transfer function $V_{m+1}^- + \epsilon_m V_m^+$ must be computed instead of the norms of the individual terms. Thus (21b) is obtained by summing all contributions and multiplying by the overall gain $w^2 \|\bar{F}\|^2$. From Fig. 3(c), γ_N for the normalized form can be written as (21c). The factors of two are due to the two independent noise sources at each node.

III. Experimental Procedures

Simulation programs were written in Fortran for implementing two's complement fixed-point arithmetic with both rounding and truncation. As long as the final output of a series of summations or subtractions lies within the computer word length, overflow for partial sums is allowable because of the modulo feature of two's complement arithmetic. Division operations do not occur in the recursive implementations of the orthogonal or standard filter forms. Thus only the fixed-point multiplication operation has to be simulated.

Let i_c define the β-bit result obtained by multiplying the two β-bit integers i_a and i_b. Also let $f_x(\cdot)$ define the "float-to-fix" operation which produces the largest integer that is less than the argument, and $f_l(\cdot)$ define the Fortran integer to floating-point operation. Simulation of β-bit two's complement fixed-point multiplication is then performed by

$$i_c = f_x(c/\omega) \tag{39a}$$

for truncation arithmetic, and

$$i_c = f_x(c/\omega + 0.5 \operatorname{sgn}(c)) \tag{39b}$$

for rounding arithmetic, where

$$c = f_l(i_a) f_l(i_b) \tag{39c}$$

$$\omega = 2^{\beta-1} \tag{39d}$$

and

$$\operatorname{sgn}(c) = \begin{cases} 1 & c \geq 0 \\ 0 & c < 0. \end{cases} \tag{39e}$$

The fixed-point multiplication simulation will be exact if the mantissa in the floating-point representation equals or exceeds 2β-bits. Generally speaking, if the computer uses 1.5β-bits for the mantissa (e.g., 24-bit floating-point mantissa for a 16-bit integer word length) exact results will be obtained for nearly all multiplication and division operations. Maximum error will occur for β-bit operations on two numbers each having magnitudes near $2^{\beta-1}$. In most practical cases, exact results will be obtained by using double-precision floating-point operations for $f_l(\cdot)$. As a limitation to this procedure, β cannot exceed the integer word length of the particular Fortran compiler. A 36-bit word length PDP-10 computer was used for the simulation. Use of double-precision arithmetic allowed approximately 16 digits of precision or exact simulation results for β up to 26 bits.

Each of the five filter structures (one-multiplier, two-multiplier, normalized, direct, and parallel) was implemented as Fortran subroutines using function subprograms for performing the fixed-point multiplication described above, e.g., IC = MUL (IA, IB). The implementation equations for the orthogonal filter forms are presented elsewhere [14].

All scaling procedures are based upon unity norm scaling which implies fractional representation for storage of coefficient values and all numerical operations. The simulation is performed by transforming the fractional coefficients to β-bit integers via

$$i_a = f_x(\omega a) \tag{40a}$$

for coefficient truncation and

$$i_a = f_x(\omega a + 0.5 \operatorname{sgn}(a)) \tag{40b}$$

for coefficient rounding, where $|a| < 1$. After simulating the filter structure, it is driven by a pseudorandom number generator having a uniform distribution and sufficiently low amplitude to insure high probability against filter overflow, i.e., to insure against the input, output, or any multiplier input exceeding $2^{\beta-1}$ in magnitude. The same input is applied to a floating-point filter simulation to obtain the "exact" output y_n. The fixed-point filter simulation results in a β-bit integer output $i_y(n)$ for $n = 0, 1, \cdots$. Since y_n defines the "exact" output, $\hat{y}_n$ defines the fixed-point simulation estimate where

$$\hat{y}_n = w\|F_u\| f_l(i_y(n))/\omega, \tag{41}$$

and $w\|F_u\|$ is a (possibly nonfractional) multiplicative factor obtained from the FWL scaling algorithm. The roundoff noise estimate $\hat{\sigma}_n^2$ is obtained from

$$\hat{\sigma}_n^2 = \frac{1}{n_2 - n_1 + 1} \sum_{n=n_1}^{n_2} (\epsilon_n - \bar{\epsilon})^2 \tag{42a}$$

where

$$\epsilon_n = y_n - \hat{y}_n \tag{42b}$$

defines the output roundoff error at index n, and

$$\bar{\epsilon} = \frac{1}{n_2 - n_1 + 1} \sum_{n=n_1}^{n_2} \epsilon_n \tag{42c}$$

defines the mean of the data in the interval n_1 to n_2. Therefore, the normalized roundoff noise estimate $\hat{\gamma}$ is from (2), (4), (39d), and (42a)

$$\hat{\gamma} = 12\,\omega^2\,\hat{\sigma}_n^2. \tag{43}$$

The lower limit n_1 is chosen sufficiently large so that the initial condition response of the filter has decayed sufficiently. The upper limit $n_2 > n_1$ is chosen sufficiently large so that reasonable statistics are obtained. With the exception of clustered pole cases, reasonable values are $n_1 = 128$ and $n_2 = 512$. For clustered pole conditions, n_1 must be increased to an index value sufficiently high to include minimal transient effects.

IV. FWL Characteristics

A. Roundoff Noise Comparison of Orthogonal and Standard Forms

A number of elliptic digital filters have been designed and analyzed based upon the theoretical equations for γ. The general results can be illustrated by way of three comparisons.

1) Roundoff Noise as a Function of Bandwidth: In the first comparison, low-pass digital elliptic filters were designed with the following conditions: $M = 6$, $F_s = 1$, 0.2 dB in-band ripple, -50.0 dB stopband ripple, maximum in-band gain of unity, and variable cutoff frequency or passband edge f_p. Roundoff noise was studied as a function of the particular filter structure and percentage bandwidth, with all other conditions held constant. The results for the five filter structures, each implementing six different conditions as a function of the normalized bandwidth $B = 2f/F_s$ are shown in Fig. 6 where $0 \leqslant B \leqslant 1$. The direct form (indicated by the lines connecting to "*D*") is inferior to all the other filters over the complete range of B. The graph has a rather steep minimum for $B = 0.50$ and clearly shows the well-known effect that pole clustering (B near zero) causes the direct form to be very undesirable. With respect to the minimum direct form noise factor γ at $B = 0.50$, at $B = 0.0625$ the filter structure is degraded by approximately 90 dB or 15 bits. It is seen that wide bandwidth also causes a similarly degrading effect on the direct form due to pole clustering near $B = 1$.

The two-multiplier lattice form roundoff noise characteristics (indicated by lines connecting to "2") are similar to the direct form characteristics in that with respect to the minimum γ near $B = 0.5$, substantial increase occurs for B near zero or unity. It is superior by a minimum of 5 dB at the least critical point ($B = 0.5$) with an increasing advantage as B moves nearer to $B = 0$ or 1. For $B = 0.9375$ it is superior to the direct form by over 30 dB.

The one-multiplier structure with optimal sign parameter choices (indicated by lines connecting to "0") is vastly superior to either the direct or lattice form for the most critical regions of B near zero and unity. This structure is judged to be significantly more robust since for $0.0625 \leqslant B \leqslant 0.9375$ a maximum variation of approximately 20 dB or 3.3 bits is computed.

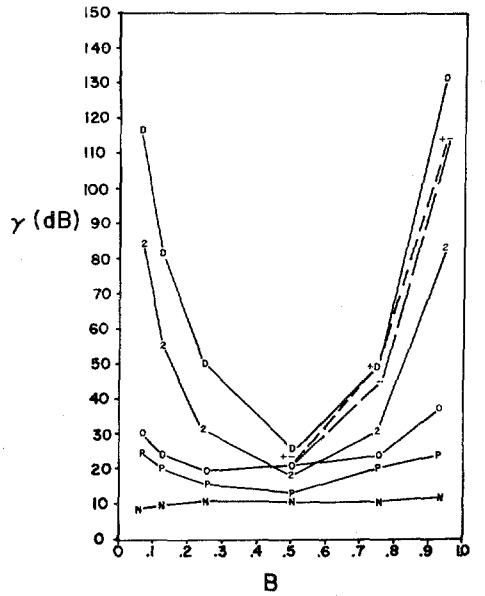

Fig. 6. Normalized roundoff noise ν for low-pass filters as a function of normalized bandwidth.

The one-multiplier form with constant sign parameters ("+" for positive sign and "-" for negative sign on graphs) has the interesting property that its noise properties are essentially identical to the optimal one-multiplier form out to $B = 0.5$ at which point it begins to track closely to the direct form. For $B > 0.5$, the two-multiplier form is significantly better than the one-multiplier form with constant sign parameters.

The parallel form (indicated by lines connecting to "*P*") has similar properties for $0.0625 \leqslant B < 0.4$, being superior by at most 5 dB. For $B > 0.5$ the advantage of the parallel form increases to about 15 dB near $B = 1$.

The normalized form (indicated by lines connecting to "*N*") is superior to the other forms for all conditions. It is remarkably robust as B decreases. In fact γ is lowest at $B = 0.0625$. Over the range of $0.0625 < B \leqslant 0.9375$ there is at most a 3 dB variation in γ. From $0.0625 \leqslant B < 0.5$ the maximum variation in γ is 1 dB as compared to 13 dB for the parallel form and 90 dB for the direct form.

For very narrow bandwidths or for very wide bandwidths, the poles will tend to cluster about the z plane points $z = +1$ and -1, respectively. For bandwidths near the quarter sampling frequency $F_s/4$, minimal clustering occurs so the concave nature of the γ curves should be expected. What was unexpected was the almost perfect symmetry about $B = 0.5$ for the forms including the one-multiplier structures.

The lack of symmetry for the constant sign parameter cases is expected because the dynamic range of the individual norms will be small when the signs of the k-parameters alternate (poles near $z = +1$) and large when the signs are identical (poles near $z = -1$). As a result, the norm values of the filter will monotonically increase or decrease with the possibility of a dynamic range even greater than the two multiplier.

2) Roundoff Noise as a Function of Filter Order: The purpose of this analysis was to determine how roundoff noise is

affected by increasing the order of a particular low-pass filter design, with all other parameters held constant. The conditions were: passband ripple = 0.2 dB, sampling frequency F_s = 1.0, passband edge $f_p = 0.1$, stopband edge $f_s = 0.105$, and maximum in-band gain of unity. As M is increased, the stopband attenuation is increased substantially, as governed by the elliptic filter design equations [5]. For example, as M increases from 2 to 10 the stopband attenuation increases from 0.696 to 76.96 dB. The results are shown in Fig. 7.

The roundoff noise for the direct form is again worse than that of all other structures. The two-multiplier form results in about the same roundoff noise as the direct form for a filter design of two orders higher. The optimal sign choice for the one-multiplier form shows only a moderate increase in roundoff noise as M is increased. The constant sign parameter forms are shown to generate substantially more noise than the optimal form for $M > 6$. The parallel and normalized forms appear to have very similar noise characteristics for this example with the normalized form being superior by about 5 dB.

3) Bandpass Filter Analysis: The third comparison was based upon the design of bandpass filters with varying center frequencies. The purpose was to see whether any substantially different roundoff noise characteristics occur with respect to the low-pass filter designs. The bandpass filter parameters are as follows: $M = 12$, passband ripple = 0.2 dB, sampling frequency $F_s = 1$, stopband attenuation of -50 dB, bandwidth = 0.05 with variable center frequency f_o (based upon the arithmetic mean) and maximum in-band gain of unity.

The theoretical noise calculations for the 42 different filters are summarized in Fig. 8. The results are consistent with the low-pass filter comparisons in the sense that the direct form is inferior to the two-multiplier form which in turn is inferior to the optimal one-multiplier structure.

The normalized form and the parallel form are essentially equivalent in noise characteristics for this bandpass case. The most striking difference between the bandpass and low-pass analysis is that whereas a maximum variation of 15 dB exists between the different structures in the low-pass case, near B = 0.5, an 84 dB difference exists in the bandpass case. All forms with the exception of the constant parameter one-multiplier forms have perfectly symmetric noise characteristics about f_o = 0.5. In contrast to the low-pass case, the roundoff noise for the constant parameter one-multiplier implementation of the bandpass filters monotonically increases with increasing center frequency.

For bandpass filters the roundoff noise symmetry is theoretically predictable for all but the one-multiplier forms since replacing z by $-z$ in any filter simply reflects the effect about $F_s/4$ in the frequency domain. In terms of the k-parameters and tap parameters, replacing z by $-z$ effects only a sign change in the even subscript k-parameters. The tap parameters remain unchanged. The noise analysis results thus remain unchanged for the two-multiplier and normalized form. The one-multiplier form with constant sign parameter follows the same pattern as for the low-pass filters. If optimal sign parameters are used the results are again symmetric since changes in the signs of the k-parameters are compensated for by simply changing the sign parameters.

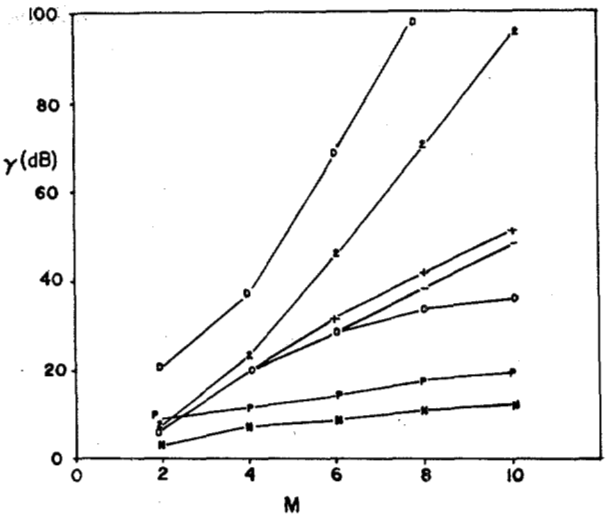

Fig. 7. Normalized roundoff noise ν as a function of filter order.

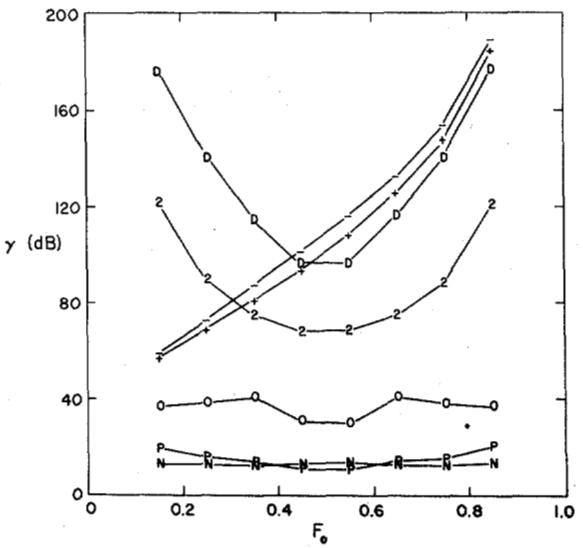

Fig. 8. Normalized roundoff noise ν for bandpass filters as a function of normalized center frequency F_O.

B. Clustered Pole Implementations

In the previous section, it was shown that for the five filter structures studied, varying degrees of increasing roundoff noise are predicted as the bandwidth of the low-pass filters either decreased or increased close to the half sampling frequency. In this section, theoretical and experimental results are presented in more detail for specific narrow bandwidth or clustered pole pair filters.

1) Narrow-Band Low-Pass Filter Implementation: As the bandwidth of a low-pass or high-pass filter is decreased, the poles tend to cluster near the points $1 + j0$ or $-1 + j0$, respectively in the z plane. The denominator of the direct form is then expected to have coefficients close to the binomial coefficients

$$(1 \pm z^{-1})^M = \sum_{m=0}^{M} \binom{M}{m} (\pm 1)^m z^{-m}. \tag{44}$$

The numerator and denominator coefficients of a 6th-order narrow-band digital elliptic low-pass filter

$$G(z) = \frac{P(z)}{A(z)} = \frac{\sum_{i=0}^{6} p_i z^{-i}}{\sum_{i=0}^{6} a_i z^{-i}} \tag{45}$$

for example, are given in Table I along with the binomial coefficients of $(1 - z^{-1})^6$. With a normalized sampling frequency $F_s = 1$, the filter design parameters were $f_p = 0.03125$ (passband frequency), $f_s = 0.0390625$ (stopband frequency), and $\epsilon_p = 0.2$ dB (passband ripple). The filter was designed so that 0 dB was the upper limit of the frequency response. The stopband attenuation was computed from the above conditions as -46.68 dB. It is precisely for the clustered pole condition that Kaiser [17] has shown the high sensitivity of the poles to quantizing errors in the polynomial coefficients.

The k-parameters k_m, $m = 0, 1, \cdots, M-1$, alpha parameters α_m and tap parameters v_m, $m = 0, 1, \cdots, M$, recursively obtained [14] from $G(z)$ are shown in Table II. Also included are the scaled tap gains to be used for the fixed-point implementation. Several interesting characteristics of the k-parameters are immediately noticed. 1) All k-parameters are fractional (as must be true for roots of $A(z)$ within the unit circle) and in addition they satisfy $\frac{1}{2} < |k_i| < 1$ for all i, so that optimum fixed-point scaling is automatically obtained, 2) the k-parameter signs alternate in correspondence with the direct form polynomial coefficients as sgn (k_i) = sgn (a_{i+1}) $i = 0, 1, \cdots, M$, 3) the k-parameter magnitudes are all very close to unity except for the first k-parameter computed from the recursion $k_5 = a_6$. In addition, whereas $p_i = p_{M-i}$ with sign alterations for $i = 1, 2, 3$, it is seen that starting from $v_6 = p_6$, the tap parameters are all positive with almost monotonic reduction in amplitude down to v_o. Although we have not been able to prove the sign correspondence under general conditions it has been observed for a large number of filter designs including high-pass, band reject forms, etc., in both narrow- and wide-band designs.

The observed relation between the k-parameters and filter coefficient values is easily shown for the tightly clustered pole-pair conditions. The recursion relation between $A(z) = A_M(z)$ and the k-parameters is [11]

$$A_{m+1}(z) = A_m(z) + k_m z^{-(m+1)} A_m(1/z) \tag{46}$$

for $m = 0, 1, \cdots, M-1$ with $A_o(z) = 1$. For $k_m = 1$ or $k_m = (-1)^m$ it is easily verified by substitution that this recursion reduces to an algorithm for generation of the binomial coefficients via Pascal's triangle, i.e.,

$$A_M(z) = (1 - z^{-1})^M = \sum_{m=0}^{M} (-1)^m \binom{M}{m} z^{-m} \tag{47a}$$

for $k_m = (-1)^m$, and

$$A_M(z) = (1 + z^{-1})^M = \sum_{m=0}^{M} \binom{M}{m} z^{-m} \tag{47b}$$

for $k_m = 1$, $m = 0, 1, \cdots, M-1$. Recalling that the norm at node m for the two-multiplier form [11] is given by $\|F_m\| = \sqrt{\alpha_m}$ and $\alpha_o > \alpha_1 > \cdots > \alpha_M = 1$, where

$$\alpha_{m+1} = \alpha_m (1 - k_m^2), \tag{48}$$

so that

$$\alpha_o = 1 \Big/ \prod_{m=0}^{M} (1 - k_m^2), \tag{49}$$

the difficulty in fixed-point implementation using the direct or two-multiplier form is immediately seen. The clustering of poles near the boundary $z = +1$ in the z plane causes the k-parameters to approach their bounds of unity magnitude and therefore, the reciprocal of the products of the quantities $(1 - k_m^2) \ll 1$ results in a very large maximum norm square $\|F_u\|^2 = \|F_o\|^2 = \alpha_o$. If a unit sample is applied at the input section, it will be increased to 1.58×10^9 in energy at the filter output. Thus, to insure against overflow at the output the input for both structures must be scaled down by the norm $\|F_u\| = \sqrt{\alpha_o} = 39\,773$. Therefore neither the direct form nor the two-multiplier lattice structure can even be implemented with 16-bit fixed-point signed arithmetic since all input samples would be scaled to zero!

Assuming greater than 16-bit arithmetic is available, the tap parameters for the two-multiplier form are scaled in accordance with the procedure described elsewhere [14] giving the results shown in the last two columns of Table II. To insure that the scaled norms into each tap multiplier exceed $\frac{1}{2}$ but remain less than unity, a maximum of 10 shifting operations at each of two nodes is necessary. The final scaled tap parameters are scaled almost optimally in the sense that they are bounded by unity and are all greater than 0.44. With the scaled unit sample input x'_n the fixed-point filter $G'(z)$ and its output y'_n are assured against overflow, assuming the parameters k_m, s_m, and v_m of Table II. For comparison with an exact implementation of $A(z)$, the scale factor $\|F_u\| w = 0.2502871$ must be applied at the output of the fixed-point implementation. It should be noted that the second-order two-multiplier implementation could be applied in a cascade form with substantial benefit in the same manner as second-order factors of the direct form. Due to the necessary coefficient scaling of 2 for cascade forms, it is conjectured that second-order cascaded forms based upon the orthogonal polynomial design will have considerably less roundoff noise when implemented in fixed-point arithmetic. This approach, however, will not be pursued here.

A vast improvement over the two-multiplier form is possible with the one-multiplier forms. The optimal sign choices are all positive due to the fact that the k-parameters alternate in sign and the sign of $k_{\max} = k_1$ is positive. The sign alternations on the k-parameters and the fact that the largest negative valued

TABLE I
COEFFICIENT LISTING FOR DIRECT FORM AND BINOMIAL EXPANSION $(1 - z^{-1})^6$

m	a_m	p_m	Binomial Coefficient
0	1.0000000	0.0047079	1
1	-5.6526064	-0.0251014	-6
2	13.3817570	0.0584417	15
3	-16.9792460	-0.0760820	-20
4	12.1764710	0.0584417	15
5	-4.6789191	-0.0251014	-6
6	0.7525573	0.0047079	1

TABLE II
COEFFICIENT LISTING FOR TWO-MULTIPLIER FORM INCLUDING FIXED-POINT SCALING FOR TAP GAINS

m	k_m	α_m	ν_m	s_m	ν'_m
0	-0.9849726	0.1581891 $E+10$	0.0000047	1	0.7535095
1	0.9970941	0.4718617 $E+08$	0.0000115	4	0.4561157
2	-0.9932416	0.2738374 $E+06$	0.0002673	64	0.6636658
3	0.9920536	0.3688878 $E+04$	0.0004605	128	0.5716858
4	-0.9800562	0.5839384 $E+02$	0.0028653	1024	0.4446411
5	0.7525573	0.2305968 $E+01$	0.0015103	512	0.4687195
6	–	0.1000000 $E+01$	0.0047079	1024	0.7305603

k-parameter is nearly equal in magnitude to the largest positive valued k-parameter cause the norms to simply appear as approximately reversed versions of the optimal or positive sign parameter norms. The norms $\|F_m\|$ of the orthogonal polynomial structures computed for node m are shown on a logarithmic scale in Fig. 9. The norms are normalized by the maximum norm within each structure. The positive, negative, and optimal sign parameter results are indicated by the ϵ^+, ϵ^-, and ϵ_{opt} curves, respectively. The multiplier node results for the positive sign and the minus sign parameters are denoted by the dashed lines. For this case, the optimal and negative sign curves are identical.

With the desire of having signal values at all nodes as large as possible without overflowing, this graph would indicate that the one-multiplier model should be far superior to the two-multiplier form indicated by "2 MUL," and that within the one-multiplier class, the various sign parameter choices should lead to rather similar results. Furthermore, since the normalized form was developed to have unity norm at each node by definition, it should then have superior FWL characteristics over the one- and two-multiplier forms (and the direct form since the norm at each node of the all-pole portion will be $\sqrt{\alpha_o}$). Both theoretical and experimental results confirm this expectation. The theoretical noise figure ν_{theo} and experimental noise figure ν_{exp} for each of the filter structures is shown in Table III, including the parallel form. In addition, the necessary input scaling factor $1/\|\bar{F}\|$ (the reciprocal of the maximum norm for the all-pole portion of the filter) for FWL implementation is shown. The experimental and theoretical results are in extremely close agreement with the possible exception of a 2.4 dB difference in the two-multiplier case. The experimental results were based upon a simulation with $\beta = 24$.

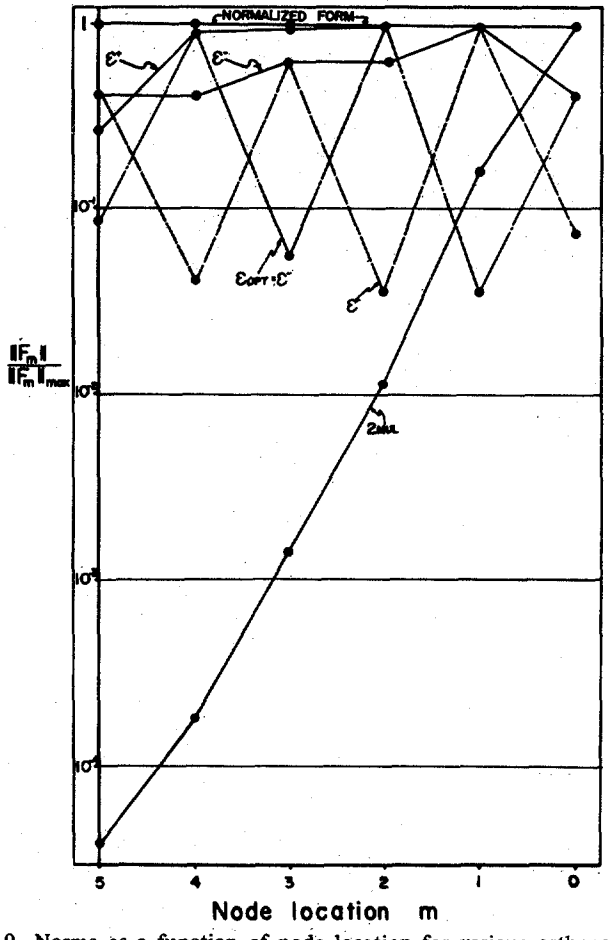

Fig. 9. Norms as a function of node location for various orthogonal polynomial filters.

The direct form is again seen to be inferior to all the structures by at least 26 dB. The normalized form is found to be superior with a maximum theoretical noise figure of 8.95 dB. The parallel form has about 15 dB worse noise figure than the optimal one-multiplier form. A large noise figure difference of 55 dB exists between the optimal one-multiplier and two-multiplier structure. Thus, either the optimal one-multiplier or normalized forms are acceptable candidates for clustered pole filter implementations while the two-multiplier form would have to be considered generally unacceptable.

2) Bandpass Filters with Clustered Poles: The vast superiority of the normalized form over the parallel form for FWL

TABLE III
THEORETICAL AND EXPERIMENTAL ROUNDOFF NOISE CALCULATIONS FOR CLUSTERED POLE EXAMPLE

Structure	ν_{theo} (Decibels)	ν_{exp} (Decibels)	$1/\|\|\overline{F}\|\|$
Direct form	112.19	111.53	$2.531\ 10^{-5}$
Two Mult.	85.68	83.25	$2.531\ 10^{-5}$
One Mult. (Opt.)	30.12	30.13	$1.4818\ 10^{-1}$
Parallel	24.17	24.47	$4.076\ 10^{-2}$
Normalized	8.95	8.61	1.000

implementation of filters with highly clustered poles is illustrated in this section by way of several bandpass filter examples. A series of eight 12th-order digital elliptic bandpass filters were designed and implemented in both parallel and normalized structures. The design parameters were 0.2 dB passband ripple, maximum gain of 0 dB, stopband attenuation of -50 dB, bandwidth of 50 Hz, sampling frequency of 10 kHz, and center frequency $f_0(\text{kHz}) = 0.5l + 0.75$, $l = 1, 2, \cdots, 6$.

These filters present extremely difficult implementation characteristics. For example, the log-magnitude spectrum of the $l = 1$ case filter unit sample response is shown in Fig. 10. Viewed over the total 5-kHz range, the filter response appears to be almost an impulse in frequency with a base line at -50 dB (the lack of ripples in the stopband is due to the limited frequency domain resolution). The double-precision direct form coefficients are presented along with the k-parameters, tap parameters, and alpha parameters (norm-squared values for the two-multiplier lattice form) in Table IV. The direct form scaling norm for this filter is the rather enormous value $\sqrt{\alpha_o} = 0.649 \times 10^9$. The k-parameters are seen to alternate between -0.7 and $1.0 - \epsilon$ where ϵ is a very small positive number. This property is based upon the poles of $1/A(z)$ having nearly identical values.

A theoretical noise figure comparison for each of the six parallel and normalized structures is presented in Fig. 11. The significant advantage of the normalized form for the case of highly clustered poles is clearly shown. The noise figure is nearly constant versus frequency at 12 dB over the total frequency range shown. The parallel form, under clustered pole conditions tends to show similar characteristics to the two-multiplier and the direct form, namely, monotonic increases in γ to the right and left of $f = F_s/4$, as illustrated in previous sections. For the $l = 1$ case, the normalized form is superior to the parallel form by approximately 61 dB or 10 bits.

CONCLUSIONS

The roundoff noise characteristics of three structures developed from the theory of orthogonal polynomials have been presented in this paper. These structures are referred to as a two-multiplier, one-multiplier, and normalized form. A complete theoretical roundoff noise analysis of the filters, when scaled for fixed-point implementation, was presented along with experimental results which verify the theory. The orthogonal polynomial structures were compared to two standard structures, the direct and parallel form.

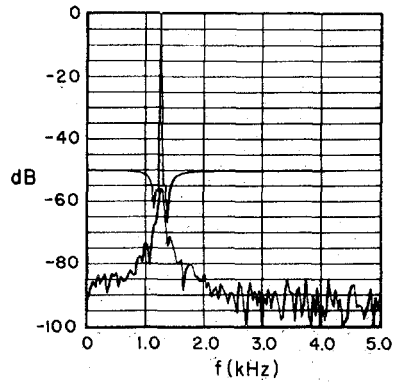

Fig. 10. Frequency response of case $l = 1$.

The theoretical noise analysis was developed in terms of recursive relations based upon orthogonal polynomial properties and Mason's flow graph theory.

It can generally be stated that the direct form is inferior to all the other structures studied. The two-multiplier lattice form (having two multiplications and two additions per section) has similar properties to the direct form, only its roundoff noise is smaller. The lattice form degrades badly in clustered pole conditions. The optimal one-multiplier form (having one multiplication and three additions per section) appears to match more closely the characteristics of the parallel form over a number of different conditions with generally larger roundoff noise.

The third, the normalized structure (having four multiplications and two additions per section), appears to have remarkably robust roundoff noise characteristics. For moderately clustered pole-pair filters, the normalized and parallel forms seem to have similar characteristics, with the normalized form having lower overall roundoff noise. For tightly clustered conditions, the normalized form is vastly superior to the parallel form. Even though this structure appears to be computationally expensive, each stage can be shown to be precisely implemented as a single complex multiplication [13], an elementary operation on some signal processing machines.

There are several modifications and extensions possible to the results presented here. For example, considerably better roundoff noise results for the one-multiplier and two-multiplier forms are possible by simply inserting binary shifts (scaling by powers of two) between adjacent sections. Since the norms at the nodes are known [11], scaling factors can be computed to increase the norms toward unity, thus improving roundoff

TABLE IV
PARAMETER LISTINGS FOR CLUSTERED POLE BANDPASS FILTER EXAMPLE

i	a_i	p_i
0	0.10000 00000 00000 D + 01	0.31086 20533 01988 D - 02
1	-0.84529 51876 65357 D + 01	-0.26333 92951 75400 D - 01
2	0.35725 53888 44417 D + 02	0.11156 90817 82533 D + 00
3	-0.97864 35972 54877 D + 02	-0.30646 12979 58780 D + 00
4	0.19203 68436 25898 D + 03	0.60318 74426 87684 D + 00
5	-0.28303 37826 85337 D + 03	-0.89197 99578 96173 D + 00
6	0.32056 44430 01992 D + 03	0.10139 44793 16550 D + 01
7	-0.28089 27896 66911 D + 03	-0.89197 99578 96173 D + 00
8	0.18914 25233 22361 D + 03	0.60318 74426 87684 D + 00
9	-0.95660 23916 98399 D + 02	-0.30646 12979 58780 D + 00
10	0.34656 76009 84371 D + 02	0.11156 90817 82533 D + 00
11	-0.81380 40444 81358 D + 01	-0.26333 92951 75399 D - 01
12	0.95546 29607 88787 D + 00	0.31086 20533 01988 D - 02

m	k_{m-1}	v_m	α_m
0	-	0.92464 66030 01427 D - 13	0.42170 E + 20
1	-0.70906 26919 71048 D + 00	0.19212 15222 30053 D - 10	0.20968 E + 20
2	0.99964 94678 57360 D + 00	-0.33950 52108 81388 D - 10	0.14698 E + 17
3	-0.70584 79764 44104 D + 00	-0.11952 34062 40397 D - 09	0.73749 E + 16
4	0.99991 43518 50219 D + 00	-0.42857 86472 68260 D - 07	0.12632 E + 13
5	-0.70705 58761 09997 D + 00	0.21688 25105 11948 D - 09	0.63171 E + 12
6	0.99982 85102 22952 D + 00	0.16884 23952 63443 D - 06	0.21665 E + 09
7	-0.70709 16745 51414 D + 00	-0.12769 23129 00162 D - 07	0.10833 E + 09
8	-0.99979 51230 16348 D + 00	-0.17590 85835 99515 D - 06	0.44383 E + 05
9	-0.70727 02736 87968 D + 00	-0.40979 60836 44792 D - 04	0.22181 E + 05
10	0.99948 26156 85041 D + 00	0.69316 56096 68874 D - 04	0.22947 E + 02
11	-0.70682 79059 12227 D + 00	-0.56909 74914 57393 D - 04	0.11482 E + 02
12	0.95546 29607 88787 D + 00	0.31086 20533 01988 D - 02	0.10000 E + 01

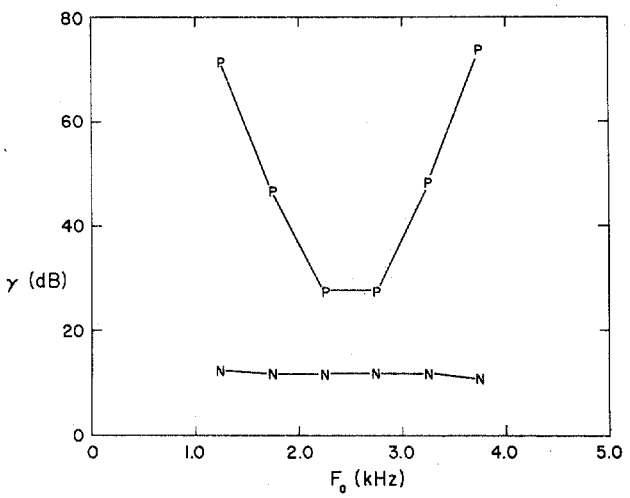

Fig. 11. Normalized roundoff noise ν as a function of center frequency F_O.

noise characteristics. The direct filter form could also be implemented in a modified cascade or parallel form where the individual sections are implemented as orthogonal polynomial filter sections. The roundoff noise properties of these alternate forms have not been studied.

REFERENCES

[1] L. B. Jackson, "An analysis of roundoff noise in digital filters," Sc.D. dissertation, Dep. Elec. Eng., Stevens Inst. Technol., Hoboken, N.J., 1969.

[2] —, "On the interaction of roundoff noise and dynamic range in digital filters," *Bell Syst. Tech. J.*, vol. 49, pp. 158–184, 1970.

[3] —, "Roundoff-noise analysis for fixed-point digital filters realized in cascade or parallel form," *IEEE Trans. Audio Electroacoust.* (*Special Issue on Digital Filtering*), vol. AU-18, pp. 107–122, June 1970.

[4] A. V. Oppenheim and C. J. Weinstein, "Effects of finite register length in digital filtering and the fast Fourier transform," *Proc. IEEE*, vol. 60, pp. 957–976, Aug. 1972.

[5] B. Gold and C. M. Rader, *Digital Processing of Signals.* New York: McGraw-Hill, 1969.

[6] J. L. Kelly, Jr. and C. Lochbaum, "Speech synthesis," in *Proc. Stockholm Speech Communications Seminar*, R.I.T., Stockholm, Sweden, Sept. 1962.

[7] F. Itakura and S. Saito, "Digital filtering techniques for speech analysis and synthesis," presented at the 7th Int. Congr. Acoust., Paper 2SC-1, Budapest, Hungary, 1971.

[8] A. Fettweis, "Some principles of designing digital filters imitating classical filter structures," *IEEE Trans. Circuit Theory* (Corresp.), vol. CT-18, pp. 314–316, Mar. 1971.

[9] S. K. Mitra and R. J. Sherwood, "Canonic realizations of digital filters using the continued fraction expansion," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 185–194, Aug. 1972.

[10] —, "Digital ladder networks," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 30–36, Feb. 1973.

[11] A. H. Gray, Jr. and J. D. Markel, "Digital lattice and ladder filter synthesis," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 491–500, Dec. 1973.

[12] R. Crochiere, "Digital ladder filter structures and coefficient sensitivity," Res. Lab. Electron., Mass. Inst. Technol., Cambridge, Rep. 103, Oct. 15, 1971.

[13] A. H. Gray, Jr. and J. D. Markel, "A normalized digital filter structure," *IEEE Trans. Acoust., Speech, Signal Processing* (*Special Issue on 1974 Arden House Workshop on Digital Signal Processing*), vol. ASSP-23, pp. 268–277, June 1975.

[14] J. D. Markel and A. H. Gray, Jr., "Fixed-point implementation algorithms for a class of orthogonal polynomial filter structures," this issue, pp. 486–494.
[15] S. J. Mason, "Feedback theory: Some properties of signal flow graphs," *Proc. IRE*, vol. 41, pp. 1144–1156, Sept. 1953.
[16] —, "Feedback theory–further properties of signal flow graphs," *Proc. IRE*, vol. 44, pp. 920–926, July 1956.
[17] J. F. Kaiser, "Some practical considerations in the realization of linear digital filters," in *Proc. 3rd Allerton Conf. Circuit and Systems Theory*, pp. 621–633, 1965.

Fixed-Point Implementation Algorithms for a Class of Orthogonal Polynomial Filter Structures

JOHN D. MARKEL, MEMBER, IEEE, AND AUGUSTINE H. GRAY, JR., MEMBER, IEEE

Abstract–**In previous papers it has been demonstrated that a class of digital filter structures derived from the theory of orthogonal polynomials has practical utility in digital signal processing problems. In this paper the specific algorithms necessary to synthesize and scale these structures in fixed-point arithmetic (based upon L_2 norm scaling) are presented and implemented as Fortran programs.**

I. Introduction

It has been shown recently that a new class of digital filter structures based upon a theory of orthogonal polynomials has important properties relevant to a number of signal processing problems. These structures are completely general in the sense that any direct form can be transformed into any of the orthogonal polynomial structures [1]-[3].

Important applications of these structures are in 1) linear prediction speech analysis/synthesis where the filter coefficients most desirable for transmission are precisely those used for implementing the filter, 2) general or special purpose computer implementation where high accuracy must be maintained with limited fixed-point word lengths, and 3) in one form, for implementing clustered pole filters.

The "two-multiplier structure" is obtained recursively from a direct form rational structure. It has been shown to have similar characteristics to the direct form but always with superior roundoff noise characteristics. A "one-multiplier" structure is obtained by generalizing the "two-multiplier" equations and then performing a simple substitution. The resultant filter is canonic in multipliers and delays and in one form (referred to as the "optimal one-multiplier structure") has generally much improved noise characteristics over the "two-multiplier." A further transformation has resulted in an orthonormal polynomial form referred to as the "normalized structure." This structure has been shown to be superior in its roundoff noise characteristics (for both rounding and truncation arithmetic) to any of the other orthogonal polynomial forms or the direct, or parallel, form. In fact, its superiority increases as the poles of the filter become more clustered (the situation in which all other forms are severely degraded).

Due to the relatively new introduction of these structures [1], [2], and their possible appearance of complexity, it is believed that algorithms for implementing these structures starting from the standard direct form would be quite useful for those interested in either studying properties of the filters or implementing particular filters on a computer with fixed-point arithmetic. The structures to be presented are designed very efficiently. As long as the direct form is stable with single precision floating-point arithmetic, no double precision operations are necessary. All parameters are automatically scaled to a fractional format with a final (possibly nonfractional) output gain factor for referencing the results to an exact implementation. Computer programs based upon the algorithms will be given along with several examples of the programs.

Manuscript received August 2, 1974; revised March 28, 1975. This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Office of Naval Research under Contract N00014-73-C-0221.

J. D. Markel is with the Speech Communications Research Laboratory, Inc., Santa Barbara, Calif. 93109.

A. H. Gray, Jr. is with the Speech Communications Research Laboratory, Inc., Santa Barbara, Calif. 93109 and the Department of Electrical Engineering and Computer Science, University of California, Santa Barbara, Calif. 93106.

II. Algorithms

A. Direct to Two-Multiplier Lattice Structure

The direct form filter is defined by

$$G(z) = P_M(z)/A_M(z) \tag{1}$$

where

$$P_m(z) = \sum_{i=0}^{m} p_{m,i}\, z^{-i} \tag{2}$$

and

$$A_m(z) = \sum_{i=0}^{m} a_{m,i}\, z^{-i} \tag{3}$$

with $a_{m,o} = 1$.

If