FIR Filter Structures Having Low Sensitivity and Roundoff Noise

ANIL MAHANTA, MEMBER, IEEE, RAMESH C. AGARWAL, MEMBER, IEEE, AND SUHASH C. DUTTA ROY

***Absrraer*—A class of structures for FIR fiiters is presented, which exhibits reduced coefficient sensitivity and superior roundoff noise properties as compared to the direct form realization. It is shown that using fixed-point arithmetic, these structures achieve the same accuracy and about the, same roundoff noise as those obtained in the floating-point implementation. The optimum structure to achieve the minimum roundoff noise can be found; in most cases na optimum results are easily obtained by simple permutations and combinations of the impulse response coefficients. While a xerial form realization of these structures requires a certain amount of software complexity, a parallel form, on the other hand, does not require additional complexity.**

I. INTRODUCTION

THE most commonly used FIR filter structures are the direct and the cascade forms, of which, the former is easier to implement because of its simplicity. However, the direct form is generally more sensitive to effects of coefficient quantization in fixed-point implementation, because of large dynamic range of the coefficients. The cascade form, on the other hand, results in reduced dynamic range and hence decreased sensitivity, but the realization is more complicated since it involves scaling of the coefficients and proper ordering of the sections to avoid overflow and minimize roundoff noise.

This paper presents a class of structures for FIR filters, which offers an attractive solution to the finite word length problems. These structures, to be designated as "nested structures" (NS), are easily derived by nesting of the transfer function polynomial *H(z)*. Due to nesting and subsequent scaling, the dynamic range is reduced considerably; at the same time, the round-off noise also decreases since most of the noise gets attenuated as it propagates towards the output.

11. COEFFICIENT SENSITIVITY

Error Bounds

Consider a length N FIR filter with the transfer function

$$H(z) = \sum_{n=o}^{N-1} a_n z^{-n}. \tag{1}$$

Instead of writing the summation in the natural order, let it be arbitrarily permuted, i.e., let

$$H(z) = \sum_{n=o}^{N-1} a_{p_n} z^{-p_n} \tag{2}$$

where p_n's are the permuted elements of the set (0, 1, $\cdots$, N- 1}. We rewrite (2) in the form

$$H(z) = a_{p_l}\left(z^{-p_0} + \frac{a_{p_1}}{a_{p_0}}\left(z^{-p_1} + \frac{a_{p_2}}{a_{p_1}} \cdot \left(z^{-p_2} + \cdots + \frac{a_{p_{N-1}}}{a_{p_{N-2}}} z^{-p_{N-1}}\right)\cdots\right)\right) \tag{3a}$$

$$= b_0\,(z^{-p_0} + b_1(z^{-p_1} + b_2 \cdot (z^{-p_2} + \cdots + b_{N-1} z^{-p_{N-1}})\cdots) \tag{3b}$$

where

$$b_0 = a_{p_0}, \quad K = \frac{a_{p_n}}{a_{Pn-l}} \text{ apn}, \quad n = 1 \text{ to } N\text{- } 1 \tag{4}$$

so that

$$a_{p_n} = \prod_{k=0}^{n} b_k. \tag{5}$$

When b_n's are rounded, the realized filter will have an effective a_{Pn} given by

$$(a_{p_n})_{\text{eff}} \triangleq \bar{a}_{p_n} = \prod_{k=0}^{n} b_k^* \tag{6}$$

where

$$b_k^* = (b_k)_r \tag{7}$$

with "r" standing for the rounding operation.

When the quantized b_n's are obtained via (4) and (7), it can be shown that the relative error in a_{Pn}, given by $E_n la_{p,}{}_n$, where

$$E_n = \bar{a}_{p_n} - a_{p_n} \tag{8}$$

tends to grow with n, due to the cumulative errors in b_o through $b_{q\text{-}l}$. Therefore we redefine b_n's as

$$b_o = \text{«B}_0 \tag{9a}$$

$$K = <\text{V}^{\text{a}}\text{Vi} = {}^{a}P_n \Big/ \prod_{k=0}^{n-l} b_k^*, \quad n = \backslash\ \text{to} N\text{- } 1 \tag{9b}$$

where

Manuscript received July 10, 1980; revised February 26, 1982.

A. Mahanta and S. C. Dutta Roy are with the Department of Electrical Engineering, Indian Institute of Technology, New Delhi 110 016, India.

R. C. Agarwal was with the Centre for Applied Research in Electronics, Indian Institute of Technology, New Delhi 110 016, India. He is now with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598.

$$v_n = (v_n)_r = v_n + \epsilon_n \tag{10}$$

where e_n is the rounding error ($|e_n| < 2/2$, $Q = 2^{-t}$ = quantization step, t = number of bits). The effective a_{pn} now becomes

$$= a_{p_n} + \bar{a}_{p_{n-1}} \epsilon_n. \tag{11}$$

Therefore

$$E_n = \bar{a}_{p_{n-1}} \epsilon_n, \qquad n = 0 \text{ to } N, \text{ with } \bar{a}_{p-1} \triangleq 1. \tag{12}$$

Let $H(e^{j\omega})$ and $H^*(e^{j\omega})$, respectively, be the frequency response of (3) before and after quantization. Then the error in the frequency response is

$$E(e^{j\omega}) = H^*(e^{j\omega}) - H(e^{j\omega}) = \sum_{n=0}^{N-1} E_n e^{-j\omega Pn}. \tag{13}$$

Therefore the bound on the frequency response error is

$$|E(e^{j\omega})| \leq \sum_{n=0}^{N-1} |E_n| \leq \sum_{n=0}^{N-1} |\bar{a}_{p_{n-1}} \epsilon_n|$$

$$\leq \frac{Q}{2} \left(1 + \sum_{n=1}^{N-1} |\bar{a}_{p_{n-1}}|\right) < \frac{Q}{2} \cdot (2) \tag{14}$$

where we have assumed

$$\sum_{n=1}^{N-1} |\bar{a}_{p_{n-1}}| < 1$$

in order that the output is bounded by 1. Also, the variance, for any w, is

$$\overline{|E(e^{j\omega})|^2} = \sum_{n=0}^{N-1} \overline{\epsilon_n^2} = \frac{Q^2}{12} \left(1 + \sum_{n=1}^{N-1} \bar{a}_{p_{n-1}}^2\right) < \frac{Q^2}{12}(2). \tag{15}$$

The corresponding results for the direct form are [1]

$$|E(e^{j\omega})|_{DF} \leq NQ/2 \tag{16}$$

$$\overline{|E(e^{j\omega})|_{DF}^2} = NQ^2/12. \tag{17}$$

These bounds indicate that the nested structure will have reduced coefficient sensitivity as compared to the direct form.

It should be noted that the absolute upper bounds (14) and (16), derived under worst case conditions are overly pessimistic. On the other hand, Gersho *et al.* [2] have pointed out that the statistical bounds (15) [and hence (17)] are not high probability upper bounds for the maximum of $|E(e^{j\omega})|$ over all w. They have shown that the high probability upper bound for large N is

$$\max_{\omega} |E(e^{j\omega})| \leq \frac{Q}{2} (N \log N)^{1/2}. \tag{18}$$

Similar bound may be derived for the nested structure also; however, it is clear from (14) that even the worst case bound in NS realization will be less than that given by (18).

Coefficient Scaling

In an actual hardware realization we shall represent b_n as

$$b_n = c_n B_n; \qquad B_n = 2^m, \quad m = 0, +1, +2, \ldots \tag{19}$$

where

$$|c_n| < 1. \tag{20}$$

Hence from (11), the effective u_{Pn} is now obtained as

$$u_n = d_{Pn} = b_n a_{Pn-1} + B_n c_n^* \tag{21}$$

where

$$c_n^* = (c_n)_r = c_n + \epsilon_n'; \qquad |\epsilon_n'| \leq Q/2. \tag{22}$$

Therefore

$$E_n = u_{Pn} - u_{Pn} = B_n (c_n^* - c_n) = \bar{a}_{Pn-1} B_n \epsilon_n' = a_{Pn} \epsilon_n' / c_n \tag{23}$$

where the last expression is obtained using (9b). Equation (23) implies that

$$|E_n| \leq |a_{Pn}| Q. \tag{24}$$

This is precisely the result obtained with floating point arithmetic. We note that although floating point representation is used for the coefficients, the arithmetic hardware to be employed is actually fixed point. In the next section, we shall show that multiply operation in the NS does not involve addition of two exponents. The exponents are predetermined and built in the structure. In other words, we have effectively represented the filter coefficients in floating point with *t-bit* mantissa while actually using only t-bit fixed point arithmetic.

III. SIGNAL SCALING

Fig. 1 shows the flow graph of the nested structure where the output y_n is obtained through a set of sequential operations given by

$$\left.\begin{aligned} S_j &= x_{x,-pi} + \cdot\,'_{1*1} \\ P_j &= b_j S_j \end{aligned}\right\} \; j = (N-1) \text{ to } 0, \text{ with } P_N \wedge 0 \tag{25}$$

where S_i and P_i, respectively, are the outputs from the jth adder and the jth multiplier. A straightforward implementation of (25) may not always be feasible because of possible overflows at the summation nodes. The sum scaling method in which $|x_n|_{\max} \leq 1$ and $|c_n| < 1$, will no doubt ensure that $|y_n| = |P_o| < 1$, but there is always some possibility that overflows can occur at the intermediate summation nodes. Hence, scale factors must be provided at the summing inputs. Under the assumption that $|x_n|_{\max} < 1$, the bound on S_i is given by

$$\hat{S}_j \; |S_j|_{\max} = 1 + |b_{j+1}| \; \hat{S}_{j+1}, \qquad j = 0 \text{ to } N-1. \tag{26}$$

Then, by using (26) recursively, S_i can be expressed as

$$\hat{S}_j = \left(\sum_{k=j}^{N-1} |a_{p_k}|\right) \Big/ |a_{pj}| = \tilde{S}_j / |a_{pj}|, \qquad j = 0 \text{ to } N-1 \tag{27}$$

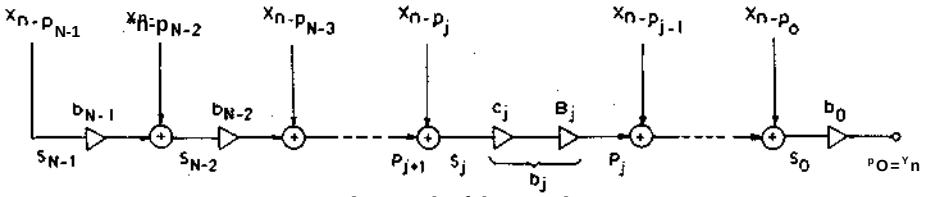


Fig. 1. Flow graph of the nested structure.

where

$$\tilde{b}_j = \sum_{k=j}^{N-1} |a_{p_k}|. \tag{28}$$

Let $\tilde{g}_j$ be scaled by a factor v_i ($= 2^t$, $U = 0,,1,2, * * e$), such that the scaled bound S_i' now lies between 3 and 1, i.e.,

$$\hat{S}_i A = v_i s_i, \quad \text{where } 3 < S_i < 1. \tag{29}$$

By doing so, we make sure that both multiplier and multiplicand can make full use of multiplier wordlength. From (26) and (29), we get

$$\omega_j' = \frac{1}{v_i} \cdot \frac{B_{i+1} I^c_{i+1} I^v_{j+1} S_{j+1}'}{v_j}$$
$$= (1/v_j) + F_{j+1} |c_{j+1}| S_{j+1}' \tag{30}$$

where

$$F_{j+1} \triangleq B_{j+1} v_{i+}, /v_i = \text{a power of 2}, \quad j = 0 \text{ to } N-2 \tag{31a}$$

and

$$F_o = B_o v_o. \tag{31b}$$

Fig. 2 shows how the scale factors are incorporated as per (30). It can be seen that $F_i > 2$; in most cases they turn out to be 1 or 3.

Although it appears that we use denormalization of the data before addition similar to that in floating point addition, we note that the addition scheme in the NS is not floating point because all scale factors are precomputed based on the upper bounds on S_i's, whereas in floating point addition, the scale factors are data-dependent and hence dynamic.

IV. ROUNDOFF NOISE

The roundoff noise model is shown in Fig. 3, where e_{vj} is the noise due to input scaling and e_{cj} is that due to multiplication roundoff. We assume that F_i is incorporated as part of the multiplier and rounding is performed after scaling. Unlike e_{ci}, which can be considered to have uniform probability density and a variance equal to $(Q^2/12)$, the noise e_{vj}, on the other hand, cannot be considered to be uniformly distributed since very few quantization levels exist because of scalings by 3, $ etc. For example, when $u_j = 4$, the possible values of e_{vj} are 0, ?Q/4 and $+Q/2$ with possibilities $, b, and $, respectively (we assume that random rounding is being used).

It can be shown that the variance of e_{vj} is equal to $d_j(Q^2/12)$, where

$$d_j = 1 + 2/v_f, \quad v_j \geq 2, \quad \text{i.e.,} \quad 1 < d_j < 1.5. \tag{32}$$

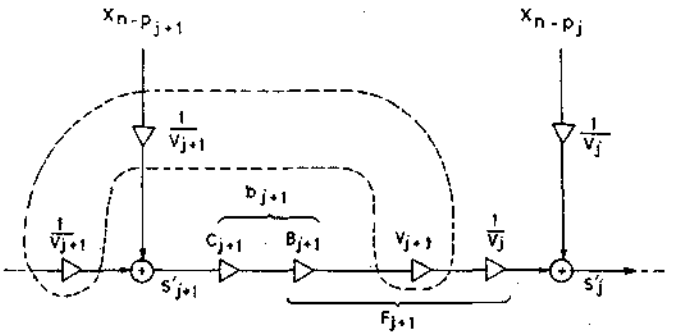


Fig. 2. Signal scaling.

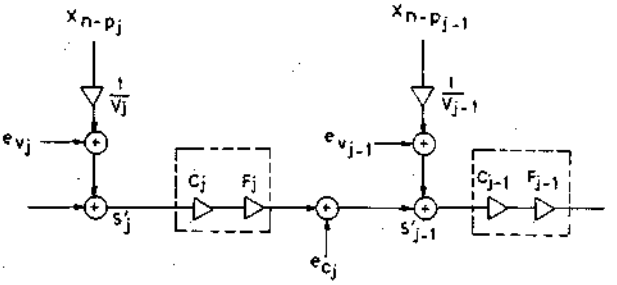


Fig. 3. Roundoff noise model.

It follows that for large u_j, the variance approaches $(Q^2/12)$, and the distribution is essentially uniform.

Since the transfer function from S_i to the output is $a_{pj} v_i$, the noise contribution from the jth node is

$$u_f = (1 \text{ t } d_i) (a_{p_j} v_i)' (Q^2/12), \quad j = 0 \text{ to } N-2. \tag{33}$$

Hence, the total output noise variance is

$$\sigma_T^2 = \frac{Q^2}{12} \left[1 + \sum_{j=0}^{N-2} (1+d_j) \tilde{a}_j^2 \right]$$
$$= \frac{Q^2}{12} [1+G] \tag{34}$$

where

$$\tilde{a}_j = a_{pj} v_i. \tag{35}$$

The first term within the brackets in (34) refers to the noise contribution from the front multiplier c_{ot} and G represents the noise due to the remaining multipliers and scalers. Since

$$|\tilde{a}_j| = |a_{p_j}| \hat{S}_j / S_j' = \left(\sum_{k=j}^{N-1} |a_{p_k}| \right) \Big/ S_j' = \tilde{S}_j / S_j' \tag{36}$$

and $\tilde{S}$ is monotonically decreasing from $\tilde{S}_0$ = 1 to $\tilde{S}_{N-1} = |a_{p_{N-1}}|$, it follows that the output noise is predominantly contributed from the first few front end sections. We shall now compare the NS noise results with that for the direct form realization. Depending on the way rounding is carried out, two cases arise in the DF [11. 1) The 2 t-bit products are first rounded to t-bits before being accumulated in a t-bit accumulator in which case the noise variance is

$$(\sigma^2)_{DF} = Ne^2/12. \tag{37a}$$

2) The products are first accumulated in a 2t-bit accumulator and then rounded to t-bits; this gives

$$(\sigma^2)_{IDF} = Q^2/12. \tag{37b}$$

It can be shown that a 2t-bit accumulator is not really necessary. By providing only log, N extra bits (guard-bits) in addition to the t-bits, we can achieve the same accuracy as that given in (37b).

Obviously, the case 2 direct form has the lowest noise; however, it requires an extra wide accumulator. Compared to the case 1 direct form, the nested structure will, however, exhibit reduced roundoff noise since in all practical cases we shall find $(1 + G) < N$. Also note that the noise does not grow linearly with N unlike in case 1 direct form because the tail end noise sources have negligible effect in the output.

Optimal Ordering of Coefficients

Recall that in the beginning of Section 11 we had mentioned about arbitrary permutation of the coefficients {an}. While such ordering does not affect the sensitivity derivations, it does affect the output roundoff noise. We shall now show that there exists an optimal ordering for which the roundoff noise is minimum.

From (34) and (36), we obtain

$$G = \sum_{j=0}^{N-2} (1 + d_j)(S_j')^2 \tilde{S}_j^2. \tag{38}$$

Since $\tilde{S}$ is monotonically decreasing for any ordering, it can be seen that $E\{\tilde{\$}\}$ and hence σ^2 will be minimum when a_i's are arranged in decreasing order of magnitude from j = 0 to $j = N - 1$, provided the ordering does not significantly alter d_i's and S_i^2's. However, this assumption may not be valid in general, and under certain conditions an exchange between adjacent coefficients (following the decreasing order) may indeed result in further reduction in noise (see Example 2). It may be noted that for all practical purposes, optimal ordering of the few largest a_i's is enough to ensure near optimum results.

V. GENERALIZED NESTED STRUCTURE

Equation (36) suggests that if $\tilde{S}_1$ can be reduced, $|\tilde{S}_1|$ will possibly be reduced. We can effectively do this by decomposing $H(z)$ into a parallel connection of p subfilters; for example, let

$$H(z) = \sum_{n=0}^{N-1} a_n z^{-n} = (a_0 + a_4 z^{-2} + a_4 z^{-4} + \cdots + a_{N-2} z^{-N+2}) + (a_1 z^{-1} + a_3 z^{-3} + \cdots + a_{N-1} z^{-N+1})$$
$$= H_1(z) + H_2(z) \quad \text{(say)}$$

so that

$$\tilde{S}_0 = \sum_{n=0}^{N-1} |a_n| = \sum_{k \in K_1} |a_k| + \sum_{k \in K_2} |a_k|,$$
$$K_1 = \{0, 2, \cdots, N-2\}, \qquad K_2 = \{1, 3, \cdots, N-1\}$$
$$= \sum_{i=1}^{2} \tilde{S}_{0i} \quad \text{(say)}$$

is distributed between two subfilters, and consequently 141's in a subfilter are likely to be reduced. Each $H_i(z)$ is then realized in a nested structure and their outputs are added to obtain the final output. Thus, from (34), it follows that for p subfilters, the total output noise variance is

$$\sigma_T^2 = \left(\mu + \sum_{i=1}^{\mu} G_i\right) Q^2/12. \tag{39}$$

Note that the generalized nested structure offers the possibility of using extra wide accumulator for the output additions; in this case, the noise reduces to

$$\sigma_T^2 = \left(1 + \sum_{i=1}^{\mu} G_i\right) Q^2/12. \tag{40}$$

Further note that the output accumulator needs to be only $(t + \log_2 N)$-bit long rather than $(t + \log_2 N)$-bit as in the case with direct form. Also, the generalized NS will require only $(p - 1)$ additions of $(t + \log_2 p)$-bit long words [the remaining $(N - p)$ additions being done with t-bit words] while the direct form requires $(N - 1)$ additions of $(t + \log_2 N)$-bit long words.

One simple way to decompose $H(z)$ is to consider the monotonic ordering (to be denoted by the set $\{a_i'\}$), and define $H_i(z)$ from the set $\{a_{i+mp}'\}$, m = 0, 1, . . . , thereby making $S_{OI} = S_{oi}$ for all i. Since F_{oi} and P are inversely related, from (39) it follows that there is an optimum p for which the noise is minimum. It may be noted that if $P = N$, the structure simply reduces to the direct form.

VI. EXAMPLES

Example 1: The preceding analysis is easily extended to the case of linear phase filters, by incorporating input adders s_i's as shown in Fig. 5, and with the assumption that $|x|_{max} < 4$ so that there is no overflow at the input adders. Consider a linear phase LPF with $N = 24$, F_p (passband cutoff frequency) = 0.08, F_s (stopband cutoff frequency) = 0.16, and δ_1/δ_2 (ripple ratio) = 1 [3, p. 189].

a) Coefficient Sensitivity: In Table I we have shown how the scaled multipliers $\{c_i\}$ are obtained for a word length r = 7 (excluding the sign bit), using (9), (19), (22), and (31) in that order. The last column of this table clearly shows that the relative accuracy is/of the oider of Q ($= 2^{-7} = 0.0078125$). Table I1 shows the bounds $\{\hat{S}_i\}$ and hence the scale factors $\{v_i\}$ and $\{F_i\}$ [(27), (29), and (31), respectively] . Table I11 shows the deviation as a function of word length t for the nested structure and the direct form. The NS is seen to exhibit about 3-4 bit superiority over the DF.

Note that if a_n's are first multiplied by 4 and then rounded to t-bits, then the DF realization with these scaled coefficients will show improved performance over that shown in Table 111.

TABLE I
COEFFICIENT SCALING SCHEME FOR THE LPF OF EXAMPLE 1

n								
0	0. 233546135	0.233%06	1A	0.9*1»*2>*	0.1111000	0.23437500	0.23$37500	3.55.~-3
1	0.1885075e	0.8d+Z9875	1	0.8d+Z9875	0.1100111	0. SOW 8750	0.188591353	4.83 E-4
2	0.11540713	0.61191923	1	0.61191923	0.1001110	0.60337500	0.1192729	$4.16 $-3
3	0.1*0089 5M1 E-1	0.3'+882527	1/2	0.697650*9	0.1011001	0.3$765625	0.39955191 E-1	3.35 3-3
4	-0.14629169 E-1	-0.3661ma	1/2	-0.73227876	-0.1011110	-0.36718750	-0. 14671d'7 LI	2.s E-3
5	-0.38112177 6-1	2.59778171,	>*	O.6"t9iA-5i«-3	0.1010011	2.S375000	-0.38053028 8-1	1.55 .Z-3
6	-0.3$0853$3 S.1	0.e3573275	1	e$573275	0.1110011	0.e 38+3750	-0.341 88268 E-1	3.02 i 3
7	-0,15929992 B-1	0.465949C9	1/2	0.93189819	0.1110111	0.4648+375	-0.15812202 &-1	2.37 2-3
8	0.2915067 E-2	-0.15992162	iA	-O.639686M-7	-0.1010010	-0.16015625	0.2%52355 E-2	1.47 6-3
9	0.105693fDO E-1	V. 15260589	8	O.519O7573	0.1oooo10	b.12500000	0.1@+99C97-" E-l	4.65 E-3
10	0.1W38299 E-1	1.42281757	2	0.7111*0879	0.1011011	1.42187500	u.1$92@+03 E-1	6.62 6-4
11	0.337(e&17 E-2	0.22601e26	1 e	O.9OVO73O6	0.1110100	0.22656250	0.33822163 5-2	2.41 E-3

TABLE I1
SIGNAL SCALING SCHEME FOR THE LPF OF EXAMPLE 1

n	n		
0	3.04	4	1
1	2.S3	i	1
2	2.51	4	1
3	4.36	a	1
4	9 .I4	16	1
5	3.14	4	1
6	2.3B'	4	1
7	2.97	4	i
8	12:32	16	1
9	2.74	4	2
10	1.23	2	1
11	1.n	1	1/8

TABLE I11
MAGNITUDE RESPONSE ERROR FOR THE LPF OF EXAMPLE 1
DEVIATION ACTUAL RESPONSE-DE~IRED RESPONSE[

n			
0	1 .l250UOO	0:53750000	1.E677
1	1 .1250000	0.75439452	1.2094
2	1 .1250flOO	0.45970916	0.4491
3	1.0312500	0.31964153	0.2075
4	1.0078125	0.23473675	0.1106
5	1.12500000	0.15221211	0.0492
6	1.1250000	0.13675307	0.0397
7	1.1250000	O.OG356881	0.0086
8	1.DO78125	3.04072377	0.0033
9	1.1250000	0.04199639	0.0037
10	1 .5000000	0.029 85681	0.0022
11	0.o	0.0033a222	0.o

This realization will, however, require 2 extra bits in the accumulator to prevent overflow.

b) Roundoff Noise: We next calculate roundoff noise in a nested realization for different ordering of the coefficients *(a,}* and compare these with results for various forms of direct form implementation. We shall also compare these results with the noise in a floating point realization of direct form [4, pp. 441-4441.

1) Natural order: Table IV shows the various quantities involved in noise calculation. Observe that the noise is effectively contributed from the first few front-end sections. The result is

$$(4)\text{natord.} = 4.95(\pounds^2/\text{l}2).$$

It may be noted that the scale factors *Fi's* (see Table 11) can be moved back and forth, and can be combined '&th other scalers, with the idea that most of the scalers might turn out to be unity. In Table 11, F_7 can be combined with F_9 thereby making all $Fi = 1$. Such shifting has little effect on the output

TABLE IV
ROUNDOFT NOISE CALCULATIONS FOR THE LPF OF EXAMPLE :1

Filter Form	t bits	Deviation in Ou	
		Passband	Stopband
OF	5	-24.29	-21.71
NS	5	-39.02	-34.11
nF	7	-34.90	-30.74
N':	7	-37.89	-36.49
OF	9	-37.50	-34.51
NS	9	-38.06	-37.63
OF	11	-37.72	-37.33
NS	11	-38.08	-38.01
	00	-38.11	-38.11

noise provided this is carried out avoiding the first few front-end sections.

2) Decreasing order: $(\&)_{de..ord.} = 4.84\,(Q^2/12)$.

3) Optimal order: The optimal order is found to be

$$i^a n)\text{opt} = (^{f1}1.\ ^a2 > ^a0 \text{ » «6 , \#3 , 05 , } 0\text{-}1, \text{ «10 , \#4 , \#9 . } ^a l\ 1 > ^a 8\ \}\cdot$$

The noise is

$$(6)\text{optord.} = 4\_72(Q^2/12)\_$$

4) Generalized MDF: A 2 section NS formed as per the guidelines mentioned in Section V yields

$$(\sigma_T^2)_{\mu=2} = 3.13\,(Q^2/12).$$

By providing 1 extra bit for the output adder the noise reduces to $2.13(Q^2/12)$.

For comparison, in the direct form implementation, the noise variance is either 12 $(Q^2/12)$ or (Q2/12) depending on whether rounding of the products is performed before or after accumulation. Furthermore, in the floating point realization, the variance is 1.64 $(Q^2/12)$.

Example 2: Consider a linear phase bandpass filter of length N = 50 [5, p. 5161.

Band edges: Band 1 (stopband) = 0.0, 0.15, Band 2 (passband) = 0.2, 0.3, Band 3 (stopband) = 0.35, 0.5.

a) Coefficient Sensitivity:

		Deviation in dB		
Type	t	Band 1	Band 2	Band 3
NS	7	-44.93	-28.36	-48.83
DF	11	-45.20	-28.36	-54.01
NS	9	-47.20	-28.56	- 60.82
DF	13	-47.88	-28.55	-64.92
NS	11	-48.18	-28.62	-66.41
DF	15	-48.44	-28.59	-66.88
	oo	-48.62	-28.62	-68.62

As in Example 1, the deviation in the DF has been evaluated with unscaled a_n's. With scaled coefficients, performance will improve,

b) Roundoff Noise:

Type ofNS	(0;) 12/Q2
1) Natural order:	16.26
2) Decreasing order:	15.79
3) Optimal order:	8.97
4) Generalized NS:	5.11, $p = 3$
	5.18, $p = 4$
	5.88, $1.1 = 2$

The subfilters were obtained as per the guidelines in Section V. With $p = 3$, slight rearrangement of the coefficients yields a value 4.30 $(Q^2/12)$.

With $p = 4$ and a $(t\ t\ 2)$-bit long accumulator for the output additions, the noise variance is 2.18 $(Q^2/12)$.

In the direct form implementation the results are 25 $(Q^2/12)$, or $(Q^2/12)$. Finally, in a floating-point realization, the result is 2.16 $(Q^2/12)$.

VII. IMPLEMENTATION CONSIDERATIONS

In a serial implementation, the nested structure will require some software control for the multipliers and the scalers. To increase the throughout rate, one desires a parallel implementation. In a parallel implementation all scalers are hard-wired; hence the structure does not require additional complexity. Consider the transfer function

$$H(z) = a,\ \text{talz}^{-'}\ ta_2z\text{-}\sim\ \ ta_3z\text{-}\sim \tag{414}$$

which can be expressed in an NS as

$$H(z) = b_0(1 + b_1(z^{-1} + b_2(z^{-2} + b_3 z^{-3}))) \tag{41b}$$

$$= b_0(1 + b_1 z^{-1}(1 + b_2 z^{-1}(1 + b_3 z^{-1}))). \tag{41c}$$

Equation (41c) can be implemented as shown in Fig. 4(a), where we have omitted the scale factors. Note that the structure does not require any shift register memory. This is because each multiplier has a built-in delay of 1 clock cycle, thus the multiplier themselves provide the necessary unit delays required in the structure.

In the above example, the nesting was done in the natural order. For any other ordering, the nested structure will require shift registers. For example, if we consider the other extreme case, i.e., if the ordering is reversed completely, so that

$$H(z) = b_o(z\sim^3 + (b_1 z^{-2} + (b_2\ 2^{-l}\ \ t\ b_3))) \tag{42}$$

then the implementation will require $2(N - 1)$ shift registers [Fig. 4(b)]. Thus the register length will vary between 0-2 (N - 1) depending on the ordering. Earlier we noted that a_n's should preferably be arranged in decreasing order of magnitude so as to minimize the roundoff noise. In most applications, the natural order is more likely to satisfy this requirement then the reversed order. In a parallel form realization, therefore, the nested structure will generally require less memory locations than a canonic form.

Finally, in Fig. 5, we have shown a fully parallel implementation of an odd-length $(N = 7)$, linear phase filter with the transfer function

$$H(z) = z^{-3}\left[a_0 + \sum_{n=1}^{3} (z^n + z^{-n})\, a_n\right]. \tag{43}$$

Here we have assumed that a_n's are arranged in their natural order. Note that the realized transfer function has a delay of 3 samples.

VIII. CONCLUSION

In this paper we have proposed a class of structures for FIR filters, which offers reduced coefficient sensitivity and superior roundoff noise properties as compared to some direct form implementations. A technique has been described to generate the multiplier coefficients such that using fixed-point arithmetic, it has been possible to realize floating-point accuracy. The output noise is predominantly contributed by the front-end noise sources, and hence, the noise does not grow linearly as in some direct form realizations using *t-bit* accumulator. In respect of roundoff noise also, these structures compare favorably with the floating-point implementation. The optimum structure to achieve the minimum roundoff noise can be found; however, the search in this direction is not as

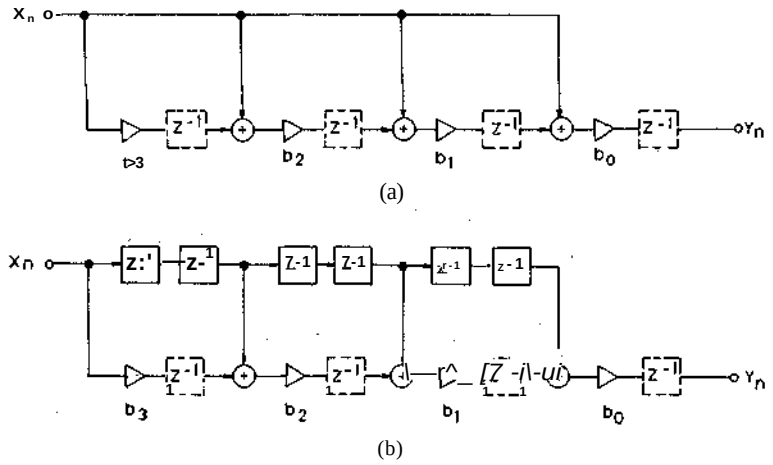


Fig. 4. Fully parallel form nested structure for (a) (41c); (b) (42).

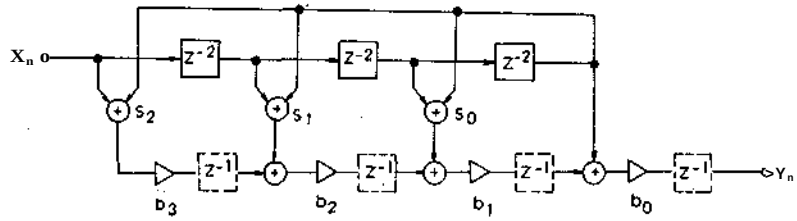


Fig. 5. Fully parallel form nested structure for the linear phase filter (43).

involved as in the cascade form, and in most cases, near optimum 'results can be easily achieved by simple permutations and combinations of the impulse response coefficients. While a serial form realization of these structures requires a certain amount of software complexity, a parallel form, on the other hand, does not require additional complexity. Further, in one form of nested structure, the realization does not require any shift register memory.

Since the output noise is predominantly contributed by the sources nearest to the output, it follows that computations at the tail-end sections can be carried out with a fewer number of bits, while prodding extra bits at the front end to achieve greater accuracy. H€me for a given cost, it should be Possible to arrive at an optimum word length configuration so as to minimize the output noise,, or vice versa. Such a cost effective design should be particularly attractive for a fully parallel realization.

ACKNOWLEDGMENT

The authors thank the reviewers for their constructive criticism and helpful suggestions.

REFERENCES

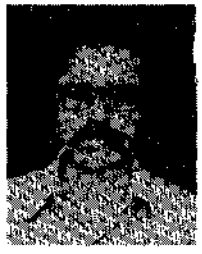
[1] D. S. K. Chan and L. R. Rabiner, "Analysis of quantization errors in the direct form for finite impulse response digital filters," *ZEEE Trans. Aui3io Electroacoust.*. vol. AU-21, pp. 354-366, Aug. 1973.

121 A. Gersho, B. Gopinath, and A. M. Odlyzko, "Coefficient inaccuracy in transversal filtering,"*BeZZSyst. Tech. J.*, vol. 58, pp. 2301-2316, Dec. 1979.

[3] L. R. Rabiner and B. *Gold,Theory and Application of Digital Signal Processing.* Englewood Cliffs, NJ: PrenticeHall, i975.

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

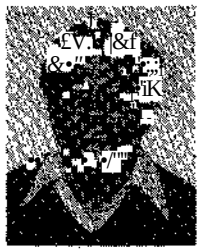
[5] J. H. McClellan, T. W. Parks, and L. R. Rabiner, "A computer program for designing optimum FIR linear phase digital filters," *ZEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 506-526, Dec.1973.



Anil Mahanta (S'76-S'79-M'794'80-M'80) was born in Gauhati, Assam, India on September 1, 1947. He received the B.E. degree in electrical communication engineering from the Indian Institute of Science, Bangalore in 1969, the M.Tech. degree in Control and Computation Engineering from the Indian Insitute of Techno'logy (IIT), Kharagpur, in 1971, and the Ph.D. degree in signal processing from the IIT Delhi in 1981.

He worked as a Scientific Officer at the Radio Astronomy Center of the Tata Institute of Fundamental Research from 1971 to 1974, and as a Lecturer in Assam Engineering College, Gauhati, India, during 1974-75. Since 1975, he has ,been with the IIT Delhi, first as a Research Scholar and then as a Research Associate in the Electrical Engineering Department.

Dr. Mahanta's research interests are in digital signal processing.



Ramesh C. Agarwal (S'74-M'74) was born in Gwalior, India, ,on May 8, 1947. He received the B.Tech. (Hons.) degree from the Indian Institute of Technology (IIT) Bombay, India, and the M.S. and Ph.D. degrees from Rice University, Houston, TX, all in i n gineering in 1968,1970, and 1974 respectively.

During 1971-72, he was an Associate Lecturer at the School of Radar Studies, IIT Delhi, India and from 1974 to 1977, he was with the IBM, T.J. Watson Research Center, Yorktown Heights, NY. He spent the period 1977 to 1981 as a Principal Scientific Officer at the Centre for Applied Research in Electronics at IIT Delhi, India, and returned to IBM in 1982. His research interests have included network synthesis, information theory and coding, number theoretic transforms, fast algorithms for computing convolution and DFT, application of digital signal processing to structure refinement of large biological molecules using X-ray diffraction data, and sonar signal processing.

Dr. Agarwal received the 1974 IEEE ASSP Senior Award, and is a member of Tau Beta Pi and Sigma Xi.

Suhash C. Dutta Roy was born in Mymensingh, now in Bangladesh. He received the B.Sc. (Hons.) degree in physics, the M.Sc. (Tech.) degree in radio physics and electronics, and the D.Phil, degree for research on network theory and solid state circuits, all from the University of Calcutta, Calcutta, India, in 1956, 1959, and 1965 respectively.

He worked with the Geological Survey of India; the River Research Institute, West Bengal, India; the University of Kalyani, West Bengal, India; and the University of Minnesota, Minneapolis, before joining the Indian Institute of Technology (IIT), New Delhi, in September 1968. He has been a Professor of Electrical Engineering at IIT since January 1970 and was the Chairman of the Department during 1970-73. During 1973-74, he was a Visiting Professor at the University of Leeds, England, and during 1978-1979, he was a Visiting Fellow at Iowa State University, Ames. He teaches circuits, systems, electronics, and signal processing courses, and conducts and supervises research in the same areas.

Professor Dutta Roy is a Fellow of the Institution of Elecbonics and Telecommunication Engineers (IETE), India, on the Editorial Board of the *International Journal of Circuit Theory and Applications*, and the Honorary Editor for *Circuits and Systems, Journal of the IETE.* He was awarded the 1973 Professor Meghnad Saha Memorial Prize and the 1980 Ram Lal.Wadhwa Gold Medal by the IETE; the 1981 Shanti Swarup Bhatnagar Award by the Government of India; and the 1981 Vikram Sarabhai Research Award by the Physical Research Laboratory, Ahmedabad, India.

Recursive Lattice Forms for Spectral Estimation

BENJAMIN FRIEDLANDER, SENIOR MEMBER, IEEE

Abstmcf—A class of lattice prediction filters is proposed for high resolution spectral estimation. The square-root normalized lattice recursions are used to estimate a set of reflection coefficients from the data. The lattice variables determine the coefficients of a least-squares predictor, from which the spectrum can be evaluated. The prewindowed and sliding window (covariance) cases are considered for both AR and ARMA spectra. The behavior of the proposed spectral estimator is illustrated by simulation results.

I. INTRODUCTION

A LARGE number of spectral estimation techniques based on autoregressive (AR) modeling were developed in the last decade. The maximum entropy method is probably the best known technique of this kind [11. The idea of an AR model fitting of time series was treated extensively in the statistical literature, in speech processing, and in the general area of least-squares estimation. Many high resolution spectral estimation techniques that do not explicitly involve autoregressive modeling are very closely related to these techniques. Examples include: the maximum likelihood method, the extended Prony method, Pisarenko's Toeplitz and non-Toeplitz algorithms, and the Hildebrand-Prony method [2].

While most of the progress in modern spectral estimation involved AR modeling, considerable work was done on autoregressive moving-average (ARMA) modeling [3], [4]. The practical application of these techniques has been somewhat limited due, perhaps, to the relative complexity of the model fitting algorithms. Some of the more recent work seems to indicate that relatively efficient ARMA spectral estimation techniques are now available [5], [6].

The common basis of all the techniques discussed above is fitting a prediction model to the observed time series. The prediction model for an ARMA *(m, n)* processy, is given by

$$9t\backslash t\text{-}i = -\sum_{i=1}^{n} A_i y_{t-i} + \sum_{i=1}^{m} {}_B i^e t\text{-}i \tag{1}$$

where e_t is the prediction error sequence

$$e_t = y_t - \hat{y}_{t|t-1} \tag{2a}$$

$$e(z) = \frac{A(z)}{B(z)} Y(z). \tag{2b}$$

$$A(z) = 1 + A_1 z^{-1} + \cdots + A_n z^{-n}$$

$$B(z) = 1 + B_1 z^{-1} + \cdots + b_m z^{-m} \tag{3}$$

z^{-1} = unit delay operator, *i.e.*, $z\text{-}\sim x, = x, _{-l}$.

The spectrum $S(w)$ of the process y, is given by

$$S(\omega) = \frac{B(e^{j\omega})\, B(e^{-j\omega})}{A(e^{>''})A(e^{->''})^{J}}\, \sigma \tag{4}$$

where o^2 is the prediction error variance. The spectral estimation problem is thus reduced to estimating the predictor coefficients from an observed set of data $Cy,, 0 < t < T\}$. The spectral estimate is then computed by (4) with the true predictor coefficients $\{Ai,Bi\}$ replaced by their estimates. In the case of AR spectral estimation $B(z) = 1$.

The prediction filter can be realized in many different ways, leading to different parametrizations of its transfer function and of the related techniques. Usually, not much attention is paid to this issue. Most spectral estimation techniques (e.g., Pisarenko's method and the first step in Prony's method) pa&metrize the spectrum by the coefficients $\{A_i, B_i\}$ of the difference equation (1). This corresponds to a direct, or tapped delay line realization of the prediction filter. An alternative parametrization is to use the so-called partial correlation (PARCOR) or reflection coefficients related to lattice filters. This parametrization is widely used in speech processing applications [7] and is implicit in the maximum entropy method of spectral estimation [11.

Manuscript received May 18, 1981; revised June 18, 1982 and June 26, 1982. This work was supported by the Office of Naval Research under Contracts N00014-79-C-0743 and N00014-81-C-0300.

The author is with Systems Control 'Technology, Inc., Palo Alto, CA 94304.