THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 3, March, 1973
Printed in U.S.A.

Theory of Roundoff Noise in Cascade Realizations of Finite Impulse Response Digital Filters

By D. S. K. CHAN and L. R. RABINER

(Manuscript received September 14, 1972)

This paper presents a theoretical treatment of the roundoff noise problem for the special case of cascade realizations of Finite Impulse Response (FIR) digital filters.† *Explicit relations for evaluating roundoff noise with the usual assumption of uncorrelated samples are presented. Useful scaling methods are stated and classified as to conditions when these methods are optimum. Important differences between use of these scaling procedures for Infinite Impulse Response (IIR) filters and FIR filters are pointed out. Finally, useful properties of linear phase FIR filters are discussed.*

I. INTRODUCTION

In recent years, many techniques have been developed for the design of Finite Impulse Response (FIR) digital filters.[2–8] It is now possible to readily design filters with arbitrary frequency or time response characteristics using the windowing,[2,3] frequency sampling,[4] or optimal design[5–8] methods. While both the windowing and frequency sampling techniques yield suboptimal filters, they are useful because of their simplicity and ease of design. The optimal design technique is of special importance because the filters it generates can be proved to be optimum in a certain sense,[7] and because efficient algorithms exist for its implementation.[7,8]

As a result of these important developments, the FIR type of digital filter is becoming increasingly attractive as an alternative to the IIR (Infinite Impulse Response) type of filter for practical applications. A major advantage of FIR filters over IIR filters is that an FIR filter

† This paper is based on a thesis[1] submitted in partial fulfillment of the requirements for the degrees of Bachelor of Science and Master of Science in the Department of Electrical Engineering at the Massachusetts Institute of Technology in September 1972.

can have an exactly linear phase response while approximating an arbitrary magnitude frequency response. But even without considering this important advantage, current research[9] is revealing that in certain cases FIR filters are competitive with IIR filters in terms of speed and cost. Thus the implementation of FIR filters using finite-precision arithmetic is becoming an important area for research.

Up to the present, little is known as to how different types of FIR filter realizations behave with respect to quantization effects. Since hardware, specifically for the purpose of realizing FIR filters, has already been built for experimentation by various research groups,[10,11] it is important to obtain more knowledge to guide the implementation phase of FIR digital filter design. The purpose of this paper is to present a theoretical treatment of several problems associated with implementations of these filters.

II. PRELIMINARY REMARKS

The effects that quantization has on an IIR filter can be classified into three basic categories:

(*i*) Quantization of the values of samples derived from a continuous input waveform causes inaccuracies in the representation of the waveform (A-D noise).

(*ii*) Finite-precision representation of the infinite-precision filter coefficients alters the frequency response characteristics of the filter (coefficient accuracy problem).

(*iii*) Finite-precision arithmetic causes inaccuracies in the filter output (roundoff noise) which, together with the finite dynamic range of the filter, limit the signal-to-noise ratio attainable. Also, finite-precision arithmetic can lead to limit cycles where the output samples are generally highly correlated.

These same quantization effects also occur in finite wordlength FIR filters with the important exception that limit cycles cannot occur in nonrecursive realizations of FIR filters. In this paper only the third type of quantization effect, viz., roundoff noise, will be discussed. Furthermore, fixed-point arithmetic with rounding will be assumed.

Except for the first category above (A-D noise), all quantization effects depend in degree and character on the type of structure used to implement a filter. There are three well-known structures in which an FIR transfer function can be realized. They are the direct form, the cascade form, and the frequency-sampling structure.[12] Other less well-known structures based on polynomial interpolation formulas

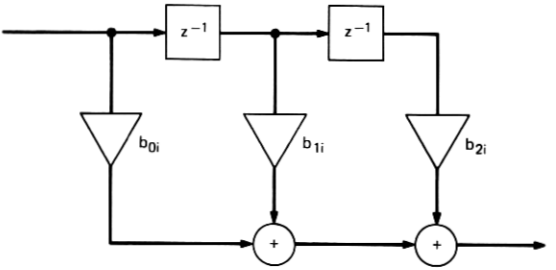

Fig. 1—Cascade-form filter section.

are also possible; these include the Lagrange, Newton, Hermite, and Taylor structures.[13] However, it is as yet unclear under what circumstances, if any, these other structures may be more advantageous than the well-known structures.

Only the cascade structure will be discussed in this paper. A second-order filter section, as shown in Fig. 1, will be used as the basic building block for the cascade structure. Although several minor variations to this configuration for the filter sections are possible,[1] the results presented here are sufficiently general so that they can be readily applied to other configurations as well.

Aside from section configuration, the prime issues that must be confronted in the realization of filters in cascade form are scaling and section ordering. Proper scaling must be performed on a filter in cascade form in order that full use of the dynamic range of each section can be made while avoiding error-producing overflows. By proper scaling, the signal-to-noise ratio of a filter can be maximized for a given quantization step size and section ordering.

Proper ordering must also be determined for a filter in cascade form if the filter is to be useful at all, since the noise output of a cascade filter can depend dramatically on the way it is ordered. For example, Schüssler[13] showed a 32nd-order FIR filter which, ordered two different ways, yields output noise variances that differ by a ratio on the order of 10^8. The problem of section ordering for cascade FIR filters has been investigated in depth,[1,14] and the results show that for higher-order filters, the variation of output noise variance across all orderings is much greater than 10^8.

Jackson[15] has formulated the roundoff noise problem for a general digital filter and has proposed an approach to the scaling of filters to satisfy dynamic range constraints. Most of his results can be specialized to the case of FIR filters by assuming a constant polynomial in the

denominator of the transfer function. However, the different perspective obtained by studying the FIR case separately affords many additional insights.

In this paper, formulas for the evaluation of roundoff noise variances in the FIR cascade structure are presented. Also, specific scaling methods for FIR filters are defined, two of which can be proved to be optimum for two classes of input signals. Finally, certain properties of the linear phase cascade structure useful in the study of section ordering are stated and can be proved. However, for reasons of space, no proofs are included in this paper. They can be found in Ref. 1.

III. DEFINITIONS

The general transfer function for an N-point FIR filter can be written in the form

$$H(z) = \sum_{k=0}^{N-1} h(k)z^{-k} \tag{1}$$

where the real-valued sequence $\{h(k), k = 0, \cdots, N-1\}$ is the impulse response of the filter. Alternatively, $H(z)$ can be expressed in the factored form

$$H(z) = \prod_{i=1}^{N_s} (b_{0i} + b_{1i}z^{-1} + b_{2i}z^{-2}) \tag{2}$$

where b_{ji}, $j = 0, 1, 2$, $i = 1, \cdots, N_s$ are real numbers and N_s, the number of factors, is defined as

$$N_s = \begin{cases} \dfrac{N-1}{2} & N \text{ odd} \\ \dfrac{N}{2} & N \text{ even} \end{cases}$$

and $b_{2N_s} = 0$ if N is even.

A linear phase filter is defined to be a filter, the transfer function $H(z)$ of which is expressible in the form

$$H(z)\,|_{z=e^{j\omega}} = H(e^{j\omega}) = \pm\,|H(e^{j\omega})|\,e^{-j\alpha\omega} \tag{3}$$

where α is a real positive constant with the physical significance of delay in number of samples. The factor $\pm$ is necessary since $H(e^{j\omega})$ actually is of the form

$$H(e^{j\omega}) = H^*(e^{j\omega})e^{-j\alpha\omega}$$

where $H^*(e^{j\omega})$ is a real function taking on both positive and negative values. It is useful to define a mirror-image polynomial (MIP) of degree N to be a polynomial of the form $\sum_{k=0}^{N} a_k z^k$, the coefficients of which satisfy the relation

$$a_k = a_{N-k} \qquad 0 \leqq k \leqq N.$$

Necessary and sufficient conditions on $H(z)$ such that a filter with transfer function $H(z)$ has an exactly linear phase response can then be stated as follows:

Theorem 1: $H(z)$ can be expressed in the form (3) *if and only if one of the following equivalent conditions hold:*

(*i*) $h(k) = h(N-1-k)$, $0 \leqq k \leqq N-1$.

(*ii*) *If z_i is a zero of $H(z)$, then z_i^{-1} is also a zero of $H(z)$. Also, if $z_i = +1$ is a zero of $H(z)$, then it occurs in even multiplicity.*

(*iii*) *Suppose z_i is a zero of the ith factor in* (2). *Let $S = \{i: z_i \text{ is real}\}$ and $Q = \{i: i \notin S\}$. Then $f(z) = \prod_{i \in S} (b_{0i} + b_{1i}z^{-1} + b_{2i}z^{-2})$ is a mirror-image polynomial in z^{-1}, and for all $i \in Q$, either $b_{0i} = b_{2i}$ or there exists $j \neq i$, $j \in Q$, such that*

$$\frac{b_{0i}}{b_{2j}} = \frac{b_{1i}}{b_{1j}} = \frac{b_{2i}}{b_{0j}}. \tag{4}$$

Furthermore, the following is a sufficient condition for $H(z)$ to be expressible in the form (3):

(*iv*) *In* (2), *for $1 \leqq i \leqq N_s$, either $b_{2i} = 0$ and $b_{0i} = b_{1i}$, or $b_{0i} = b_{2i}$, or there exists $j \neq i$, $1 \leqq j \leqq N_s$, such that*

$$\frac{b_{0i}}{b_{2j}} = \frac{b_{1i}}{b_{1j}} = \frac{b_{2i}}{b_{0j}}.$$

In all cases the value of α is $\alpha = (N-1)/2$.

It should be pointed out that a section with $b_{0i} = b_{2i}$ is necessarily one which synthesizes either two complex conjugate zeros on the unit circle, or two reciprocal zeros on the real axis, or two identical zeros at $+1$ or -1. Furthermore, two sections which satisfy (4) are precisely those sections which synthesize reciprocal zeros (i.e., if z_i is a zero of one section, then z_i^{-1} is a zero of the other section). Thus, taking (2) as the basis for the FIR cascade form, condition (*iv*) of Theorem 1 provides a way to assign zeros to individual sections of the cascade structure so that linear phase is guaranteed independent

of scaling or ordering. Hence, in this paper the following convention of zeros assignment for linear phase filters will be adopted: complex zeros are grouped by conjugate pairs, real zeros that are reciprocals of each other are paired together, while doubled or higher multiplicity zeros are grouped by pairs of the same kind. In this way the only zero that can occur by itself in a section is $z = -1$ (since by Theorem 1, $z = +1$ is not allowed as a zero of odd multiplicity).

The definition (3) of a linear phase filter requires the filter to have both constant group delay and constant phase delay. However, if only constant group delay is desired, a second type of "linear phase" filter can be defined in which the phase of $H(e^{j\omega})$ is a piecewise linear function of ω, i.e.,

$$H(e^{j\omega}) = \pm |H(e^{j\omega})| e^{j(\beta - \alpha\omega)}. \tag{5}$$

It can be shown[1] that with the constraint (1) on the form of $H(z)$, the only possible solutions for $\beta \in [-\pi, \pi]$ is $\beta = \pm (k\pi/2)$, $k = 0, 1, 2$. If $\beta = 0, \pm\pi$, (5) reduces to (3). Thus the only new cases added are when $\beta = \pm \pi/2$. These cases arise exactly when $z_i = +1$ occurs as a zero of $H(z)$ in odd multiplicity, or, equivalently, when $\{h(k)\}$ satisfies

$$h(k) = -h(N - 1 - k) \qquad 0 \leqq k \leqq N - 1.$$

Filters of this special type are useful in the design of wideband differentiators.[16] However, this type of filter will not be considered in this paper and the term "linear phase filter" will be restricted to refer to those filters satisfying (3).

IV. THEORY OF FIR CASCADE STRUCTURES

4.1 *Roundoff Noise in the Cascade Structure*

The analysis of roundoff noise in this paper is based on the usual model used for such analyses in digital filters.[15,17,18] In particular, each multiplier in a filter is modeled as an infinite-precision multiplier followed by a summation node where roundoff noise is added to the product so that the overall result equals some quantized level. Each noise sample is modeled as a random variable with uniform probability density on the interval $(-Q/2, Q/2)$ and zero density elsewhere, where Q is the quantization step size. Thus each sample is a zero-mean random variable with a variance of $Q^2/12$.

Furthermore, the following assumptions are made:

(*i*) Any two different samples from the same noise source are uncorrelated.

(*ii*) Any two different noise sources (i.e., associated with different multipliers), regarded as random processes, are uncorrelated.

(*iii*) Each noise source is uncorrelated with the input signal.

Thus each noise source is modeled as a discrete stationary white random process with a uniform power density spectrum of magnitude $Q^2/12$.

Applying this model to the filter section shown in Fig. 1, the addition of a noise source to the output of any multiplier is seen to be equivalent to adding a noise source to the output of the section. Therefore, to model a section of a cascade filter, k_i noise sources are added to the output of the section, where k_i is the number of multipliers with non-integer coefficients in the section. Or, equivalently, by assumption (*ii*) above, one noise source of variance $k_i(Q^2/12)$ can be added instead.

For the configuration shown in Fig. 1, k_i is in general equal to 3. However, when $b_{0i} = b_{2i}$, the signals of the two branches feeding the multipliers with coefficients b_{0i} and b_{2i} can first be summed before being multiplied by the common coefficient, thus reducing k_i to 2. Furthermore, by a sacrifice in speed (assuming serial arithmetic), it is possible, as demonstrated by practical hardware,[10] to reduce k_i to 1 for all i by summing all products in each section before performing rounding. It is of interest to point out that the same can be done in the direct form, resulting in effectively only one noise source of variance $Q^2/12$ feeding into the output of the filter.

Before proceeding further, some notations need to be developed. Let $H_i(z)$ denote the transfer function of the ith section of a filter $H(z)$, i.e.,

$$H(z) = \prod_{i=1}^{N_s} H_i(z) \tag{6}$$

where

$$H_i(z) = b_{0i} + b_{1i}z^{-1} + b_{2i}z^{-2}. \tag{7}$$

As a convention, filter sections will be numbered in increasing numbers according to increasing distance from the filter input (i.e., the section at the input is called the 1st section).

Furthermore, define

$$G_i(z) = \begin{cases} \displaystyle\prod_{j=i+1}^{N_s} H_j(z) & 0 \leqq i \leqq N_s - 1 \\ 1 & i = N_s \end{cases} \tag{8}$$

and let $\{g_i(k)\}$ be the impulse response of $G_i(z)$, i.e.,

$$G_i(z) = \sum_k g_i(k)z^{-k}. \tag{9}$$

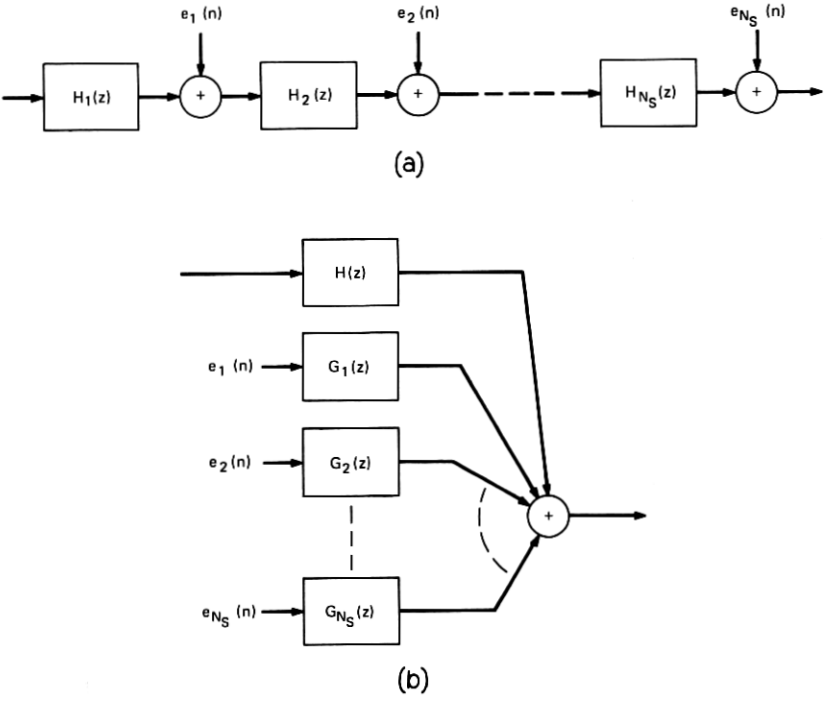

Fig. 2—Equivalent models for a filter in cascade form.

Then a cascade filter can be modeled as in Fig. 2a or equivalently as in Fig. 2b, where $\{e_i(n)\}$ is the noise source for the ith section. Letting $\{E_i(n)\}$ denote the noise sequence at the filter output due to the ith noise source alone gives

$$E_i(n) = \sum_k g_i(k)e_i(n - k). \tag{10}$$

By the stationarity of $\{e_i(n)\}$, the variance of $E_i(n)$ is independent of n; hence denoting this variance by σ_i^2, assumption (i) above leads to the relation

$$\begin{aligned}\sigma_i^2 &= \sum_k g_i^2(k)\overline{e_i(n - k)^2} \\ &= k_i \frac{Q^2}{12} \sum_k g_i^2(k).\end{aligned} \tag{11}$$

Now the total noise output is given by

$$E(n) = \sum_{i=1}^{N_s} E_i(n) = \sum_{i=1}^{N_s} \sum_k g_i(k)e_i(n - k). \tag{12}$$

Therefore by assumptions (*i*) and (*ii*)

$$\sigma^2 = \overline{E^2(n)} = \sum_{i=1}^{N_s} \sigma_i^2. \tag{13}$$

4.2 *Methods of Scaling to Meet Dynamic Range Constraints*

A practical digital filter, necessarily implemented as a physical device, must have a finite dynamic range. Especially when fixed-point arithmetic is employed, this dynamic range sets a practical limit to the maximum range of signal levels representable in a filter and acts to constrain the signal-to-noise ratio attainable.

In some filter structures such as the direct form, given the filter transfer function, the designer has no control over the relative signal levels at points within the filter. Only the gain of the overall filter can be varied. However, in a cascade realization with N_s sections there are $N_s - 1$ degrees of freedom available in addition to the overall filter gain and the ordering of sections.

To see this, a factorization for $H(z)$ is defined which is unique up to ordering of factors, in the form

$$H(z) = \beta \prod_{i=1}^{N_s} \hat{H}_i(z)$$

$$\hat{H}_i(z) = a_{0i} + a_{1i}z^{-1} + a_{2i}z^{-2} \tag{14}$$

where $\{a_{ij}\}$ satisfies

$$a_{0i} \geqq 0, \qquad \sum_{j=0}^{2} |a_{ji}| = 1 \qquad i = 1, \cdots, N_s. \tag{15}$$

Then the transfer function for the ith section in a cascade realization can be written as

$$H_i(z) = S_i \hat{H}_i(z) \tag{16}$$

where S_i is an arbitrary constant, subject only to the constraint that

$$\prod_{i=1}^{N_s} S_i = \beta. \tag{17}$$

Thus given β, $N_s - 1$ of the S_i's can be chosen at will.

Any rule for assigning values to $\{S_i\}$ will be referred to as a scaling method. Obviously, some scaling method must be employed in the design of a cascade filter whether or not one is concerned with dynamic range constraints, since numerical values must be assigned to the S_i's. When dynamic range is an issue, the constraints it imposes can be

met in some best manner by choosing the proper scaling method. In this paper, filters, designed so that no arithmetic overflow in them can cause distortion in the filter output, will be studied. Therefore, the investigation of scaling methods will be restricted to those methods which guarantee that for a given class of input signals no distortion-causing overflow occurs in the scaled filter.

It can be shown[19] that, in an addition operation, if two's complement arithmetic is used, as is usually the case, then as long as the final result is within the representable numerical range, individual partial sums can be allowed to overflow without causing inaccuracies in the result. In this paper, it is assumed that all additions in a filter are done using two's complement arithmetic. Then, to guarantee that no distortion caused by overflow occurs at a cascade filter's output, only the input and output of each filter section need be constrained not to overflow.

To simplify the discussion of scaling methods, the following definitions are used. Let

$$F_i(z) = \sum_{k=0}^{2i} f_i(k)z^{-k} = \prod_{j=1}^{i} H_j(z) \tag{18}$$

and

$$1 \leqq i \leqq N_s$$

$$\hat{F}_i(z) = \sum_{k=0}^{2i} \hat{f}_i(k)z^{-k} = \prod_{j=1}^{i} \hat{H}_j(z). \tag{19}$$

Also, let $\{v_i(n)\}$ be the output sequence of $F_i(z)$ or $H_i(z)$. Furthermore, assume that the maximum magnitude of numerical data representable in a filter is 1.0. Then the necessary overflow constraints on a cascade filter can be stated as

$$|v_i(n)| \leqq 1 \qquad 1 \leqq i \leqq N_s, \qquad \text{all } n. \tag{20}$$

Necessary and sufficient conditions for (20) to hold for two classes of input signals are given below. Theorem 2 deals with the class of input sequences $\{x(n)\}$ which satisfies $|x(n)| \leqq 1$ for all n. For simplicity, this class is refererd to as "class 1." Theorem 3 deals with the class of inputs with transform $X(e^{j\omega})$ which satisfies

$$\frac{1}{2\pi}\int_0^{2\pi} |X(e^{j\omega})|\, d\omega \leqq 1.$$

This class will be called "class 2." By virtue of the fact that

$$x(n) = \frac{1}{2\pi}\int_0^{2\pi} X(e^{j\omega})e^{j\omega n}d\omega \tag{21}$$

and hence

$$|x(n)| \leqq \frac{1}{2\pi} \int_0^{2\pi} |X(e^{j\omega})| \, d\omega, \tag{22}$$

class 2 is a subset of class 1.

Theorem 2: Suppose $|x(n)| \leqq 1$. *Then* $|v_i(n)| \leqq 1$, $1 \leqq i \leqq N_s$ *and all* n, *if and only if*

$$\sum_{k=0}^{2i} |f_i(k)| \leqq 1 \qquad i = 1, \cdots, N_s. \tag{23}$$

Theorem 3: Suppose $1/(2\pi) \int_0^{2\pi} |X(e^{j\omega})| \, d\omega \leqq 1$. *Then* $|v_i(n)| \leqq 1$, $1 \leqq i \leqq N_s$ *and all* n, *if and only if*

$$|F_i(e^{j\omega})| \leqq 1 \qquad i = 1, \cdots, N_s \qquad 0 \leqq \omega \leqq 2\pi. \tag{24}$$

Conditions (23) and (24) of Theorems 2 and 3 can be restated to give conditions on $\{S_i\}$. Recall that the $\hat{H}_i(z)$'s are unique once $H(z)$ is given, hence the $\hat{F}_i(z)$'s and $\{\hat{f}_i(k)\}$'s are also unique. Equations (16), (18), and (19) give

$$f_i(k) = \left(\prod_{j=1}^{i} S_j \right) \hat{f}_i(k) \tag{25}$$

and

$$F_i(z) = \left(\prod_{j=1}^{i} S_j \right) \hat{F}_i(z). \tag{26}$$

Therefore, conditions (23) and (24) can be restated respectively as

$$\prod_{j=1}^{i} |S_j| \leqq \left[\sum_{k=0}^{2i} |\hat{f}_i(k)| \right]^{-1} \tag{27}$$

and

$$\prod_{j=1}^{i} |S_j| \leqq \left[\max_{0 \leqq \omega \leqq 2\pi} |\hat{F}_i(e^{j\omega})| \right]^{-1} \tag{28}$$

$i = 1, \cdots, N_s.$

These then are conditions which, for the class of inputs concerned, a scaling method must satisfy. It will be shown next that in some sense optimum scaling methods are obtained when (27) and (28) are satisfied with equality. For ease of reference, two scaling methods will first be defined.

Define *sum scaling* to be the rule:

$$\prod_{j=1}^{i} S_j = \left[\sum_{k=0}^{2i} |\hat{f}_i(k)| \right]^{-1} \qquad i = 1, \cdots, N_s \tag{29}$$

or stated recursively,

$$S_i = \begin{cases} \left[\sum_{k=0}^{2} |\hat{f}_1(k)| \right]^{-1} & i = 1 \\ \left[\left(\prod_{j=1}^{i-1} S_j \right) \sum_{k=0}^{2i} |\hat{f}_i(k)| \right]^{-1} & i = 2, \cdots, N_s. \end{cases} \tag{30}$$

Also, define *peak scaling* to be the rule:

$$\prod_{j=1}^{i} S_j = \left[\max_{0 \leqq \omega \leqq 2\pi} |\hat{F}_i(e^{j\omega})| \right]^{-1} \qquad i = 1, \cdots, N_s \tag{31}$$

or

$$S_i = \begin{cases} \left[\max_{0 \leqq \omega \leqq 2\pi} |\hat{F}_1(e^{j\omega})| \right]^{-1} & i = 1 \\ \left[\left(\prod_{j=1}^{i-1} S_j \right) \max_{0 \leqq \omega \leqq 2\pi} |\hat{F}_i(e^{j\omega})| \right]^{-1} & i = 2, \cdots, N_s. \end{cases} \tag{32}$$

Theorem 4: Given an FIR transfer function to be realized in cascade form (as defined in Fig. 1) using fixed-point arithmetic of a given word-length, and given the ordering of filter sections, assume that:

(i) The number of noise sources in each section (i.e., k_i) is independent of the scaling method.
(ii) All filter coefficients can be represented to arbitrary precision.
(iii) No overflow is allowed to occur at the input and output of each section.
(iv) The overall gain of the filter is maximized subject to no overflow at the filter output.

Then each of the following scaling methods is optimum for the class of input signals stated, in the sense that it yields the minimum possible roundoff noise variance as defined in (13) among all scaling methods which satisfy conditions (iii) and (iv) above for the class of inputs considered.

(i) Sum scaling for class 1 signals.
(ii) Peak scaling for class 2 signals.

Thus optimal scaling methods are established for two classes of input signals. It is possible to define other classes of signals by considering the "*Lp* norm" of their transforms.[15,17] Specifically, the *Lp* norm of $X(e^{j\omega})$ is defined as

$$\|X(e^{j\omega})\|_p = \left[\frac{1}{2\pi} \int_0^{2\pi} |X(e^{j\omega})|^p d\omega \right]^{1/p} \qquad 1 \leqq p \leqq \infty \tag{33}$$

where for $p = \infty$ the limit as $p \to \infty$ of the right-hand side is meant. For each p, a class of signals can be defined consisting of those sequences with transforms which satisfy

$$\|X(e^{j\omega})\|_p \leqq 1. \tag{34}$$

Signals satisfying (34) will be referred to as Lp-norm constrained signals. Note that L_1-norm constrained signals are simply class 2 signals.

For proofs of the following useful theorem, refer to Refs. 1, 20, and 21.

Theorem 5: Let $X(e^{j\omega})$ and $Y(e^{j\omega})$ be transforms of sequences. Then

(*i*) $$\|X(e^{j\omega})\|_\infty = \max_{0 \leqq \omega \leqq 2\pi} |X(e^{j\omega})|$$

(*ii*) $$\|X(e^{j\omega})Y(e^{j\omega})\|_1 \leqq \|X(e^{j\omega})\|_p \|Y(e^{j\omega})\|_q$$

if $1/p + 1/q = 1, \quad 1 \leqq p,q \leqq \infty$

(*iii*) $$\|X(e^{j\omega})\|_r \leqq \|X(e^{j\omega})\|_s \quad \text{if} \quad 1 \leqq r \leqq s \leqq \infty .$$

Since with input $\{x(n)\}$,

$$v_i(n) = \frac{1}{2\pi} \int_0^{2\pi} F_i(e^{j\omega})X(e^{j\omega})e^{j\omega n}d\omega, \tag{35}$$

so that

$$|v_i(n)| \leqq \frac{1}{2\pi} \int_0^{2\pi} |F_i(e^{j\omega})X(e^{j\omega})|\,d\omega = \|F_i(e^{j\omega})X(e^{j\omega})\|_1, \tag{36}$$

by Theorem 5 (*ii*),

$$|v_i(n)| \leqq \|F_i(e^{j\omega})\|_p \|X(e^{j\omega})\|_q \qquad 1 \leqq i \leqq N_s. \tag{37}$$

Hence for L_q-norm constrained signals, i.e., if $\|X(e^{j\omega})\|_q \leqq 1$, the following scaling method (L_p-norm scaling) is obtained.

$$\|F_i(e^{j\omega})\|_p = 1 \qquad \begin{matrix} p = \dfrac{q}{q-1} \\ i = 1, \cdots, N_s, \end{matrix} \tag{38}$$

or stated in terms of $\{S_i\}$,

$$\prod_{j=1}^{i} S_j = [\|\hat{F}_i(e^{j\omega})\|_p]^{-1} \qquad i = 1, \cdots, N_s. \tag{39}$$

Notice that by virtue of part (*i*) of Theorem 5, L_∞-norm scaling is just peak scaling which has been shown to be optimum for class 2, or L_1-norm constrained, signals. Furthermore, by part (*iii*) of the theorem,

the following hierarchy of classes of signals is obtained:

class 1 $\supset$ class 2 $\supset$ L_p-norm constrained $\supset$ L_q-norm constrained
if $1 \leqq p \leqq q \leqq \infty$.

In general, class 1 and class 2 signals are the most useful to consider. L_2-norm constrained signals with L_2-norm scaling are useful when all inputs to a filter have finite energy bounded by a known value. For, by Parseval's Theorem, the energy of $\{x(n)\}$ is simply given by $(\|X(e^{j\omega})\|_2)^2$. Hence, if the input signals are first scaled so that their maximum energy is 1.0 (or the squared dynamic range of the filter), then L_2-norm scaling is sufficient to ensure no overflow.

L_2-norm scaling finds greater application for FIR filters than for IIR filters because, in the former case, it is applicable for a larger class of input signals. In particular, an Nth-order FIR filter has only N samples of memory; thus if the input signal to an Nth-order FIR filter consists of bursts of energy spaced more than N samples apart with zero energy in between, then the filter will effectively "see" only one burst at a time. Hence, the maximum energy of a burst can be used as the bound on the energy of the input as far as scaling is concerned. Thus an infinite-energy signal can have the effect of a finite-energy signal on an FIR filter.

Clearly, sum scaling and peak scaling can also be applied to IIR filters.[15] In fact, Theorems 2 and 3 are also valid for IIR filters. However, the input sequence needed in Theorem 2 to prove necessity in the case of IIR filters is an infinite-duration sequence extending to $-\infty$ with full dynamic range magnitudes, and signs that match those of $\{f_i(k)\}$ for some i. Since $\{f_i(k)\}$ for IIR filters is infinite in duration for all i, clearly such an input sequence is highly improbable. Hence, class 1 signals have been deemed too restrictive a description for ordinary inputs to an IIR filter, resulting in too stringent a scaling method.[15]

However, for FIR filters it is not difficult to find an input sequence within dynamic range which will require sum scaling to ensure no overflow, since only a small, finite portion of the sequence need match up with the $\{f_i(k)\}$'s. For example, if $F_1(z)$ has a zero with angle ω_0, $\pi/2 \leqq \omega_0 < \pi$, then all three samples of $\{f_1(k)\}$ have the same sign; hence an input sequence need only have three consecutive samples of value 1 before $|v_1(n)| = \sum_k |f_1(k)|$ for some n.

4.3 *Properties of the Linear Phase Cascade Structure*

Two theorems regarding certain properties of the linear phase cascade form are now given. These results are useful in the investigation of ordering of cascade filter sections.[14]

Theorem 6: Let $H_i(z)$ be the transfer function for the ith section of a linear phase FIR filter in cascade form, where

$$H_i(z) = b_{0i} + b_{1i}z^{-1} + b_{2i}z^{-2},$$

and let ω_i be the angle of one of its zeros, $-\pi \leqq \omega_i \leqq \pi$. Then for all i:

$$(i) \qquad \max_{\omega} |H_i(e^{j\omega})| = \begin{cases} |H_i(e^{j\pi})| & 0 \leqq |\omega_i| < \dfrac{\pi}{2} \\ |H_i(e^{j0})| & \dfrac{\pi}{2} \leqq |\omega_i| \leqq \pi \end{cases}.$$

$$(ii) \qquad \sum_{l=0}^{2} |b_{li}| = \max_{\omega} |H_i(e^{j\omega})| = \max\left(|H_i(e^{j0})|, |H_i(e^{j\pi})|\right).$$

The next theorem is concerned with the equivalence of certain orderings with regard to output noise variance. In particular, it states that with peak scaling each pair of sections in a filter which synthesize reciprocal zeros is completely interchangeable without affecting the output noise variance of the filter. With sum scaling, however, this is not necessarily true. Nevertheless, a weaker condition can be stated which says that, with sum scaling, if every pair of sections which synthesize reciprocal zeros of a filter is interchanged in position, then output noise variance is not changed. Figure 3 illustrates two such equivalent orderings.

Theorem 7: Let $\{H_i(z)\}$ and $\{H_i'(z)\}$ be the section transfer functions of two orderings for a linear phase filter $H(z)$, both scaled by the same method, thus

$$H(z) = \prod_{i=1}^{N_s} H_i(z) = \prod_{i=1}^{N_s} H_i'(z).$$

Then filters with section transfer functions $\{H_i(z)\}$ and $\{H_i'(z)\}$ produce identical output noise variances if either of the following conditions is true:

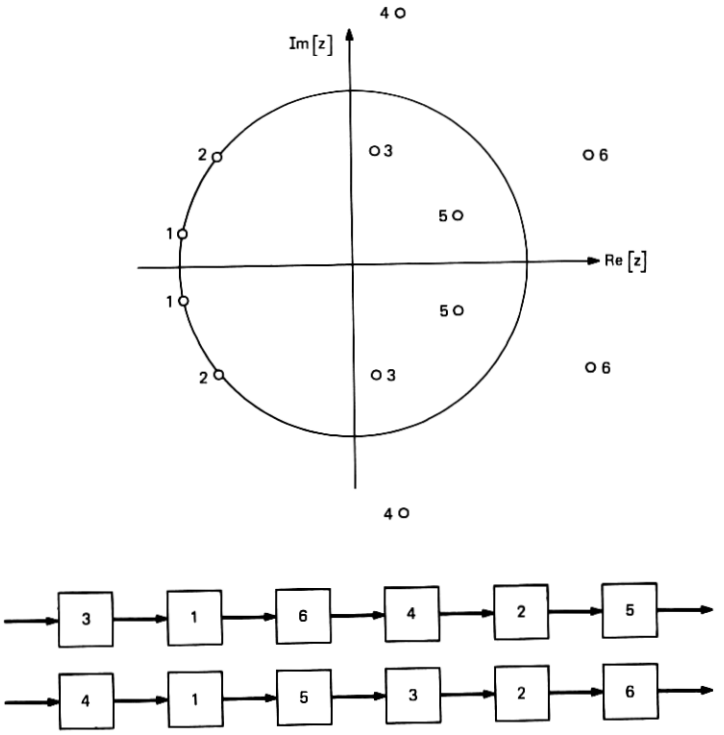

Fig. 3—Two orderings with equal output noise variances.

(*i*) *Peak scaling is used and for each* i, $H_i(z)$ *and* $H_i'(z)$ *have either the same zeros or reciprocal zeros* [*i.e.*, $H_i(z_i) = 0$ *if* $H_i'(z_i^{-1}) = 0$].

(*ii*) *Sum scaling is used and for all* i, z_i^{-1} *is a zero of* $H_i'(z)$ *whenever* z_i *is a zero of* $H_i(z)$.

REFERENCES

1. Chan, D. S. K., "Roundoff Noise in Cascade Realization of Finite Impulse Response Digital Filters," S. B. and S. M. Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass., September 1972.
2. Rabiner, L. R., "Techniques for Designing Finite-Duration Impulse-Response Digital Filters," IEEE Trans. Commun. Tech., *COM-19*, No. 2 (April 1971), pp. 188–195.
3. Kaiser, J. F., "Digital Filters," ch. 7 in *Systems Analysis by Digital Computer*, F. F. Kuo and J. F. Kaiser, eds., New York: Wiley, 1966.
4. Rabiner, L. R., Gold, B., and McGonegal, C. A., "An Approach to the Approximation Problem for Nonrecursive Digital Filters," IEEE Trans. Audio Electroacoustics, *AU-18*, No. 2 (June 1970), pp. 83–106.
5. Herrmann, O., "Design of Nonrecursive Digital Filters with Linear Phase," Elec. Ltrs., *6*, No. 11, 1970, pp. 328–329.

6. Rabiner, L. R., "The Design of Finite Impulse Response Digital Filters Using Linear Programming Techniques," B.S.T.J., *51*, No. 6 (July-August 1972), pp. 1177–1198.
7. Parks, T. W., and McClellan, J. H., "Chebyshev Approximation for Nonrecursive Digital Filters with Linear Phase," IEEE Trans. Circuit Theory, *CT-19* (March 1972), pp. 189–194.
8. Hofstetter, E., Oppenheim, A. V., and Siegel, J., "A New Technique for the Design of Nonrecursive Digital Filters," Proc. Fifth Annual Princeton Conf. Inform. Sci. and Syst., 1971, pp. 64–72.
9. Current research by L. R. Rabiner, J. F. Kaiser, and O. Herrmann.
10. Schüssler, W., personal communication.
11. Stitt, J. R., and Sonderegger, R. E., "Development and Application of a Programmable Real-Time 200 Coefficient Nonrecursive Digital Filter," Hardware Technical Note of Electronic Communications, Inc., St. Petersburg, Florida, 1969.
12. Rabiner, L. R., and Schafer, R. W., "Recursive and Nonrecursive Realizations of Digital Filters Designed by Frequency Sampling Techniques," IEEE Trans. Audio Electroacoustics, *AU-19*, No. 3 (September 1971), pp. 200–207.
13. Schüssler, W., "On Structures for Nonrecursive Digital Filters," Archiv Für Elektronik Und Übertragungstechnik, Band 26, 1972.
14. Chan, D. S. K., and Rabiner, L. R., "An Algorithm for Minimizing Roundoff Noise in Cascade Realizations of Finite Impulse Response Digital Filters," B.S.T.J., this issue, pp. 347–385.
15. Jackson, L. B., "On the Interaction of Roundoff Noise and Dynamic Range in Digital Filters," B.S.T.J., *49*, No. 2 (February 1970), pp. 159–184.
16. Rabiner, L. R., and Steiglitz, K., "The Design of Wide-Band Recursive and Nonrecursive Digital Differentiators," IEEE Trans. Audio Electroacoustics, *AU-18*, No. 2 (June 1970), pp. 204–290.
17. Jackson, L. B., "Roundoff-Noise Analysis for Fixed-Point Digital Filters Realized in Cascade or Parallel Form," IEEE Trans. Audio Electroacoustics, *AU-18*, No. 2 (June 1970), pp. 107–122.
18. Weinstein, C. J. "Quantization Effects in Digital Filters," Technical Report 468, Lincoln Laboratory, Lexington, Mass., November 1969.
19. Chu, Y., *Digital Computer Design Fundamentals*, New York: McGraw-Hill, 1962.
20. Fleming, W. H., *Functions of Several Variables*, Reading, Massachusetts: Addison-Wesley, 1965, pp. 200–204.
21. Rice, J. R., *The Approximation of Functions*, Reading, Massachusetts: Addison-Wesley, 1964, pp. 4–10.