THE BELL SYSTEM
TECHNICAL JOURNAL
DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 49 February 1970 Number 2

On the Interaction of Roundoff Noise and Dynamic Range in Digital Filters*

By LELAND B. JACKSON

(Manuscript received October 22, 1969)

The interaction between the roundoff-noise output from a digital filter and the associated dynamic-range limitations is investigated for the case of uncorrelated rounding errors from sample to sample and from one error source to another. The required dynamic-range constraints are derived in terms of L_p norms of the input-signal spectrum and the transfer responses to selected nodes within the filter. The concept of "transpose configurations" is introduced and is found to be quite useful in digital-filter synthesis; for although such configurations have identical transfer functions, their roundoff-noise outputs and dynamic-range limitations can be quite different, in general. Two transpose configurations for the direct form of a digital filter are used to illustrate these results.

I. INTRODUCTION

With the rapid development of digital integrated circuits in the 1960's and the potential for large-scale integration (LSI) of these circuits in the 1970's, digital signal processing has become much more than a tool for the simulation of analog systems or a technique for the implementa-

* This paper is taken in part from a thesis submitted by Leland B. Jackson in partial fulfillment of the requirements for the degree of Doctor of Science in the Department of Electrical Engineering at Stevens Institute of Technology.[1]

tion of very complex and costly one-of-a-kind systems alone. The traditional advantages of digital systems, such as high accuracy, stable parameter values, and straight-forward realization, have been supplemented through the use of integrated circuits by the additional advantages of high reliability, small circuit size, and ever-decreasing cost. As a result, it now appears that many signal processing systems which have been in the exclusive domain of analog circuits may in the future be implemented using digital circuits; while other proposed systems which could not be implemented at all because of the practical limitations of analog circuits may now be realized with digital circuits.[2]

The key element in most of these new signal-processing systems is the digital filter. The term "digital filter" here denotes a time-invariant, discrete or sampled-data filter with finite accuracy in the representation of all data and parameter values.[3-5] That is, all data and parameters within the filter are "quantized" to a finite set of allowable values with, in general, some form of error being incurred as a result of the quantization process. Implicit in this quantization is a maximum value or set of maximum values for the magnitudes of these data and parameters which, in the case of the data, is usually referred to as the "dynamic range" of the filter.

Without the above quantization effects, linear discrete filters could be implemented exactly. Of course, one very significant feature of digital signal processing is that arbitrarily high accuracy can, in fact, be maintained once the initial analog-to-digital (A-D) conversion (if any) has taken place. However, there are still practical limitations to the accuracy of any physical system, and often it is desirable to minimize the accuracy of the implementation (while still satisfying the system specifications) in order to minimize the cost of the system. Hence, a thorough understanding of quantization errors in digital filters is quite important if the full potential of digital signal processing is ever to be realized.

II. QUANTIZATION ERRORS IN DIGITAL FILTERS

The specific sources of quantization error in the implementation and operation of a digital filter are as follows:

(*i*) The filter coefficients (multiplying constants) must be quantized to some finite number of digits (usually binary digits, or bits).

(*ii*) The input samples to the filter must also be quantized to a finite number of digits.

(*iii*) The products of the multiplications (of data by coefficients)

within the filter must usually be rounded or truncated to a smaller number of digits.

(*iv*) When floating-point arithmetic is used, rounding or truncation must usually be performed before or after additions as well.

The first source of error above is deterministic and straightforward to analyze in that the filter characteristics must simply be recomputed to reflect the (small) changes in the filter coefficients due to quantizing.[6,7] However, the inclusion of coefficient quantization in the initial filter synthesis procedure in order to minimize (in some sense) the resulting filter complexity produces a complex problem in nonlinear integer programming which has only begun to be investigated.

The second source of error is often referred to as "quantization noise". It is inherent in any A-D conversion process and has been studied in great depth.[8] Hence, input quantization has not been included in our investigation, except as it relates to other error sources of interest.

The third and fourth error sources are similar to the second since they also involve quantization of the data, but they differ in two respects: (i) The data to be quantized is already digital in form, and (ii) the rounding or truncation of the data takes place at various points *within* the filter, not just at its input. To distinguish these sources of error from the input quantization noise, the resulting error processes will be referred to as "roundoff noise" (to be used generically, whether rounding or truncation is actually employed). Because of (ii), the roundoff noise is potentially much larger than the input quantization noise, and it is one of the principal factors which determine the complexity of the digital filter implementation, especially when special-purpose hardware is used.

There are three variables in the filter implementation which determine the level and character of the roundoff noise for a given input signal:

(*i*) the number of digits (bits) used to represent the data within the filter,

(*ii*) the "mode" of arithmetic employed (that is, fixed-point or floating-point), and

(*iii*) the circuit configuration of the digital filter. The number of digits in the data may be thought of as determining either the quantization step size or the dynamic range of the filter. We choose here the latter interpretation in order to have the same step size for all filters. Therefore, with this interpretation, the number of data digits does not affect the level of the roundoff noise directly, but rather it limits the maximum allowable signal level and hence the realizable signal-to-noise ratio. Data within the filter must, of course, be properly "scaled" if the

maximum signal-to-noise ratio is to be maintained without exceeding the dynamic-range limitations. Among the principal results reported here are the determination of appropriate scaling for certain important classes of input signals and the calculation of the effect of this scaling on the output roundoff noise.

The output roundoff noise from a floating-point digital filter is usually (but not always) less than that from a fixed-point filter with the same total number of data digits because of the automatic scaling provided by floating-point arithmetic.[9,10] However, since floating-point arithmetic is significantly more complex and costly to implement, most special-purpose digital filters have been, and will probably continue to be, constructed with fixed-point hardware. Hence, we have considered only fixed-point digital filters in this work although much of the analysis could be adapted to floating-point filters. Oppenheim has recently proposed another interesting mode of arithmetic for digital filter implementation, called "block-floating-point", which provides a simplified form of automatic scaling of the filter data.[11] As would be expected, the performance of block-floating-point appears to lie somewhere between those of fixed-point and of floating-point.

The third variable in the implementation of a digital filter, that of circuit configuration, is the principal factor determining the character (spectrum) of the output roundoff noise and, along with mode of the arithmetic, ultimately determines the number of data digits required to satisfy the performance specifications. In fact, the key step in the synthesis of a digital filter is the selection of an appropriate configuration for the digital circuit. There are a multitude of equivalent circuit configurations for any given linear *discrete* filter (whose transfer function is expressible as a rational fraction in z); but in the implementation of the corresponding *digital* filter, these configurations are no longer equivalent, in general, because of the effects of coefficient quantization and roundoff noise. As noted previously, the effects of coefficient quantization are deterministic and can thus be accounted for exactly as a (typically small) change in the transfer function of the discrete filter. Therefore, assuming that the coefficients for the configurations under consideration have been (or can be) quantized satisfactorily, the choice between these configurations is then determined by the level and character of their output roundoff noise. As we will show, there can be very significant differences between the roundoff-noise outputs of otherwise equivalent digital filter configurations.

The content and complexity of any analysis of roundoff noise are determined to a large extent by the assumed correlation between round-

off errors. If these errors may be assumed to be uncorrelated from sample to sample and from multiplier (or other rounding point) to multiplier, then the roundoff-noise analysis is relatively straightforward, and the results are independent of the exact nature of the input signal to the filter. If, on the other hand, uncorrelated errors may not be assumed, then the analysis is much more complex, and the results are generally dependent on the particular input signal or class of input signals. This paper is concerned exclusively with the uncorrelated-error case because this assumption seems to be valid for most filters with input signals of reasonable amplitude and spectral content. Even in this case, the inclusion of the associated dynamic-range constraints makes the analysis reasonably involved and the corresponding synthesis problem quite complex.

Although the generic term "roundoff noise" has been used to include the case of truncation as well as rounding, we actually concentrate on the rounding case. As long as the assumption of uncorrelated errors can be made, our results are applicable to either case, with the error variance for truncation being four times that for rounding. However, as the input signals become less "random", the uncorrelated-error assumption tends to break down for truncation more readily than for rounding. Hence, additional care must be exercised in applying these results to the truncation case.

III. FILTER MODEL FOR UNCORRELATED-ROUNDOFF-NOISE ANALYSIS

The analyses appearing in the literature concerning roundoff noise in digital filters usually employ the simplifying and often reasonable assumption of uncorrelated roundoff errors from sample to sample and from one error source (multiplier or other rounding point) to another.[9,12,13] This assumption is based on the intuitively plausible and experimentally supported notion that for sufficiently large and dynamic signals within the filter, the small roundoff error made at one point in the network and/or in time should have little relationship to (that is, correlation with) the roundoff error made at any other point in the network and/or time. The advantage of assuming uncorrelated errors from one sample to another is that the noise injected into the filter by each rounding operation is then "white"; while the advantage of assuming uncorrelated error sources is that the output noise power spectrum may then be computed as simply the superposition of the (filtered) noise spectra due to the separate error sources.[12] Experimental results which support the validity of this assumption, even in the case

of a single sinusoidal input, are presented in Ref. 1. In this section, we introduce the notation and develop the analysis pertaining to uncorrelated roundoff noise for later use in investigating the synthesis of digital filters.

Digital filter networks are composed of three basic elements: adders, constant multipliers, and delays. The interconnection of these elements into a particular network configuration is the key step in digital filter synthesis. For our purposes here, we need only consider the network as a directed graph, with the multipliers and delays being represented by graph branches. The branch interconnection points, or nodes, will be divided into two types: "summation nodes", which correspond to the adders and have multiple inputs and a single output, and "branch nodes", which correspond to simple "wired" interconnections that have a single input and one or more outputs.

A digital filter network may thus be represented as shown in Fig. 1. The input to and output from the filter at time $t = nT$ are denoted by $u(n)$ and $y(n)$, respectively. The corresponding output from the i^{th} branch node is denoted by $v_i(n)$; while the roundoff error introduced into the filter at the j^{th} summation node is denoted by $e_j(n)$. Since with fixed-point arithmetic, rounding is performed only after multiplications, non-zero roundoff errors are "input" to the filter only at those summation nodes which follow constant (non-integer) multiplier branches, as depicted in Fig. 2.

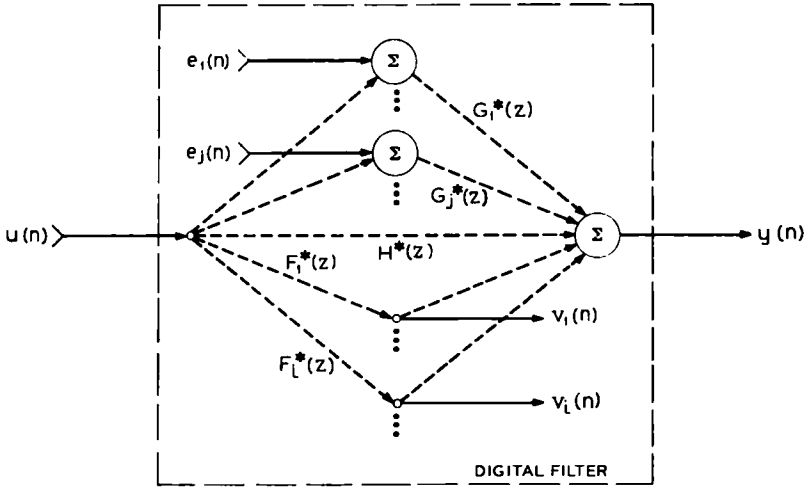

Fig. 1 — General digital filter model.

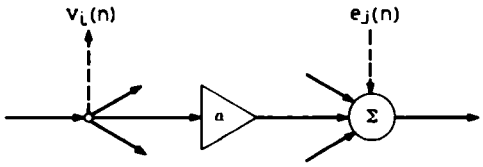

Fig. 2 — Constant multiplier with preceding branch node and succeeding summation node.

For a unit sample input to the filter at $t = 0$ and no rounding [that is, $u(0) = 1$, $u(n) = 0$ for $n \neq 0$, and $e_i(n) = 0$ for all j and n], the resulting output values $y(n)$ and $v_i(n)$ for all $n \geqq 0$ and all i are designated as $h(n)$ and $f_i(n)$, respectively. Alternatively, for a unit sample input to the j^{th} summation node and zero inputs otherwise [that is, $e_i(0) = 1$, $e_j(n) = 0$ for $n \neq 0$, and $e_k(n) = u(n) = 0$ for all n and for $k \neq j$], the resulting output values $y(n)$ for all $n \geqq 0$ are denoted by $g_j(n)$. We thus have the following transfer functions of interest, expressed in z-transform form:

From filter input to output:

$$H^*(z) = \sum_{n=0}^{\infty} h(n)z^{-n}. \tag{1}$$

From filter input to i^{th} branch-node output:

$$F_i^*(z) = \sum_{n=0}^{\infty} f_i(n)z^{-n}. \tag{2}$$

From j^{th} summation-node input to filter output:

$$G_j^*(z) = \sum_{n=0}^{\infty} g_j(n)z^{-n}. \tag{3}$$

These transfer functions are indicated in Fig. 1.

The frequency responses (Fourier transforms) corresponding to the above transfer functions are given by [3-5]

$$H(\omega) = H^*(e^{j\omega T}), \tag{4}$$

$$F_i(\omega) = F_i^*(e^{j\omega T}), \tag{5}$$

$$G_k(\omega) = G_k^*(e^{j\omega T}). \tag{6}$$

This notation will be used throughout this paper. That is, for any z-transform $A^*(z)$ which converges for $|z| = 1$, the corresponding Fourier transform is given by

$$A(\omega) = A^*(e^{j\omega T}).$$

If scaling has been included in the filter design in order to satisfy certain dynamic-range constraints, then prime marks (′) are added to denote this fact [for example, $F_i'(\omega)$, $F_j^{*\prime}(z)$].

Each error source (rounding operation) within the filter is assumed to inject white noise of uniform power-spectral density N_0. Assuming uniformly distributed rounding errors with zero mean, the variance of the roundoff noise from each error source is given by[12,13]

$$\sigma_0^2 = \Delta^2/12 \tag{7}$$

where Δ is the spacing of the quantization steps (after rounding). To eliminate the sampling period T from certain expressions of interest, we now define $N_0 = \sigma_0^2$. Hence, the variance, or total average power, corresponding to an arbitrary power-density spectrum $N(\omega)$ with no DC component (which implies a zero-mean process) is given by†

$$\sigma^2 = \frac{1}{\omega_s} \int_0^{\omega_s} N(\omega)\, d\omega \tag{8}$$

where ω_s is the radian sampling frequency given by

$$\omega_s = 2\pi/T. \tag{9}$$

Assume now that k_j error sources input to the j^{th} summation node. The spectral density of the roundoff error sequence $\{e_j(n)\}$ is then just $k_j N_0$ by our assumption of uncorrelated error sources. The total roundoff noise in the output of the filter thus has a power-density spectrum given by[12]

$$N_y(\omega) = \sigma_0^2 \sum_j k_j \mid G_j(\omega) \mid^2 \tag{10a}$$

where we have substituted σ_0^2 for N_0. If scaling has been included in the filter design, then the corresponding expression is just

$$N_y(\omega) = \sigma_0^2 \sum_j k_j' \mid G_j'(\omega) \mid^2 \tag{10b}$$

where $k_j' \geqq k_j$ to account for the additional scaling multipliers.

IV. DYNAMIC-RANGE CONSTRAINTS

The ultimate objective of the synthesis procedures to be investigated will be the minimization of some norm of $N_y(\omega)$ for a given quantization step size Δ, subject to certain "constraints". One constraint is that the

† This normalization of $N(\omega)$ is further motivated by the derivation in Section V leading to equation (30b).

specified transfer function $H^*(z)$ must be maintained. Another fundamental, but often overlooked, constraint is the finite dynamic range of the filter. Specifically, the signals $v_i(n)$ at certain branch nodes within the filter cannot be allowed to "overflow" (that is, exceed the dynamic-range limitations), at least not more than some small percentage of the time, in order to prevent severe distortion in the filter output.

Overflow constraints are required only at certain branch nodes in the digital circuit because it is only the inputs to the constant multipliers which cannot be allowed to overflow when several standard numbering systems are used (for example, one's- or two's-complement binary).[14] Specifically, in the summation of more than two numbers, if the magnitude of the correct total sum is small enough to allow its representation by the K available digits, then in these numbering systems the correct total sum will be obtained regardless of the order in which the numbers are added, even if an overflow occurs in one of the partial sums. Hence, those node outputs which correspond to partial sums comprising a larger total sum may be allowed to overflow, as long as the total sum is constrained not to overflow. This property also applies when one of the inputs to a summation node has overflowed as a result of a multiplication by a coefficient of magnitude greater than one.

Turning to the formulation of the required overflow constraints, we may easily derive an upper bound on the magnitude of the signals $v_i(n)$ for all possible input sequences $\{u(n)\}$, neglecting the (small) error signals $e_i(n)$. Assuming zero initial conditions in the filter and $e_j(n) = 0$ for all j and n, the i^{th} branch-node output $v_i(n)$ is given by

$$v_i(n) = \sum_{k=0}^{\infty} f_i(k)u(n-k), \quad \text{all} \quad n. \tag{11}$$

Therefore, given that $u(n)$ is bounded in magnitude by some number M for all n, an upper bound on the magnitude of $v_i(n)$ is given by[15]

$$|v_i(n)| \leqq M \sum_{k=0}^{\infty} |f_i(k)|, \quad \text{all} \quad n. \tag{12}$$

Thus, if the node signal $v_i(n)$ is also to be bounded in magnitude by M for all possible input sequences, the associated scaling must ensure that

$$\sum_{k=0}^{\infty} |f_i'(k)| \leqq 1. \tag{13}$$

That (13) is not only a sufficient condition to rule out overflow for all possible input sequences $\{u(n)\}$, but also a necessary condition, is easily

shown by letting $u(n) = \pm M$ for all n, with $\text{sgn}\,[u(n_0 - k)] = \text{sgn}\,[f_i(k)]$ for some $n = n_0$ and all $k \geqq 0$. Then from equation (11) we see that (12) is satisfied with equality in this case, and thus (13) is a necessary condition, as well.

The norm of $f_i'(k)$ employed in (13) is not very useful in practice because of the difficulty of evaluating the indicated summation in all but the simplest cases. Also, for large classes of input signals, (12) and thus (13) are overly pessimistic. Therefore, we now derive alternate conditions on (the transform of) the scaled unit-sample response $\{f_i'(n)\}$ which ensure that for certain classes of input signals, the corresponding branch-node output $v_i(n)$ cannot overflow. The derivation of these conditions for discrete systems closely parallels the corresponding derivation for continuous systems, as given by Papoulis.[16]

An alternate expression for equation (11) in terms of z-transforms is derived as follows: Consider an (absolutely summable) deterministic input sequence $\{u(n)\}$ possessing the z-transform

$$U^*(z) = \sum_{n=-\infty}^{\infty} u(n)z^{-n}, \qquad a < |z| < b, \tag{14}$$

for some $a < 1$ and $b > 1$. Stability requires that $F_i^*(z)$, defined in equation (2), exist for all $|z| > c$ for some $c < 1$. Hence, the z-transform of $\{v_i(n)\}$ is given by[3]

$$V_i^*(z) = F_i^*(z)U^*(z), \qquad d < |z| < b, \tag{15}$$

where $d = \max(a, c)$. The inverse transform of equation (15) is given by[3]

$$v_i(n) = \frac{1}{2\pi j} \oint_\Gamma V_i^*(z)z^{n-1}\,dz \tag{16}$$

where the contour of integration Γ is contained in the region of convergence $d < |z| < b$. Since $d < 1$ and $b > 1$, let Γ be the unit circle in the z plane ($|z| = 1$), and perform the change of variables $z = e^{j\omega T}$ in equation (16). Using equation (15), the resulting equation becomes

$$v_i(n) = \frac{1}{\omega_s} \int_0^{\omega_s} F_i(\omega)U(\omega)e^{jn\omega T}\,d\omega. \tag{17}$$

The conditions to be derived from equation (17) are most easily expressed in terms of L_p norms, defined for an arbitrary periodic function $A(\cdot)$ with period ω_s by[17]

$$\|A\|_p = \left[\frac{1}{\omega_s} \int_0^{\omega_s} |A(\omega)|^p\,d\omega\right]^{1/p} \tag{18a}$$

for each real $p \geqq 1$ such that

$$\int_0^{\omega_s} | A(\omega) |^p \, d\omega < \infty .$$

It can be shown[17] that for $A(\cdot)$ continuous, the limit of equation (18a) as $p \to \infty$ exists and is given by

$$|| A ||_\infty = \max_{0 \leq \omega \leq \omega_s} | A(\omega) | . \tag{18b}$$

Assume now that $| U(\omega) |$ is bounded from above by some number M (that is, $|| U ||_\infty \leqq M$). Then, from equation (17),

$$| v_i(n) | \leqq M \frac{1}{\omega_s} \int_0^{\omega_s} | F_i(\omega) | \, d\omega$$

or

$$| v_i(n) | \leqq || F_i ||_1 \cdot || U ||_\infty . \tag{19}$$

In exactly the same manner, we may also show that

$$| v_i(n) | \leqq || F_i ||_\infty \cdot || U ||_1 . \tag{20}$$

Applying the Schwarz inequality to equation (17), on the other hand, yields that

$$| v_i(n) |^2 \leqq \frac{1}{\omega_s^2} \int_0^{\omega_s} | F_i(\omega) |^2 \, d\omega \int_0^{\omega_s} | U(\nu) |^2 \, d\nu$$

or

$$| v_i(n) | \leqq || F_i ||_2 \cdot || U ||_2 . \tag{21}$$

Note that (19), (20), and (21) are all of the form

$$| v_i(n) | \leqq || F_i ||_p \cdot || U ||_q , \qquad \left(\frac{1}{p} + \frac{1}{q} = 1 \right) \tag{22}$$

for $p, q = 1, 2$, and ∞. It can be shown[18] that (22) is true in general for all $p, q > 1$ satisfying $1/p + 1/q = 1$; and we have shown in (19) and (20) that if the L_∞ norms exist, then (22) holds for $p, q = 1$, as well. The general relation in (22) for all $p, q > 1$, is derived from Holder's inequality.

A simple, but important special case of (22) results from letting $F_i^*(z) = F_i(\omega) = 1$. Since $|| 1 ||_p = 1$ for all $p \geqq 1$, we then have simply

$$| u(n) | \leqq || U ||_q , \quad \text{all} \quad q \geqq 1. \tag{23}$$

But since (23) holds for all sequences $\{u(n)\}$, it must also be true that

$$| v_i(n) | \leqq || V_i ||_r , \quad \text{all} \quad r \geqq 1.$$

This is, in fact, the basis of (22), for Holder's inequality actually states that

$$|| V_i' ||_1 \leqq || F_i ||_p || U ||_q , \qquad \left(\frac{1}{p} + \frac{1}{q} = 1\right).$$

Therefore, the real implication of (22) is that the mean absolute value of $V_i(\omega)$ is bounded by $|| F_i ||_p || U ||_q$, and this, in turn, provides a bound on $| v_i(n) |$.

Assume, therefore, that the input transform $U(\omega)$ satisfies $|| U ||_q \leqq M$ for some $q \geqq 1$. From (23) we immediately have that $| u(n) | \leqq M$ for all n. Then, if $| v_i(n) |$ is also to be bounded by M, (22) provides a sufficient condition on the scaling to ensure this, namely

$$|| F_i' ||_p \leqq 1, \qquad (|| U ||_q \leqq M) \tag{24}$$

for $p = q/(q-1)$. Inequality (24) is the desired condition to replace the more general, but often less useful condition given by (13).

From an engineering viewpoint, the most significant values for p and q would seem to be 1, 2, and ∞. The case $p = 1$, $q = \infty$ requires that the input transform $U(\omega)$ be everywhere bounded in magnitude by M (that is, $|| U ||_\infty \leqq M$), in which case only the L_1 norm of the scaled transfer function $F_i'(\omega)$ need satisfy (24). For an input of finite energy $E = \sum_n u^2(n)$, Parseval's identity implies that $|| U ||_2^2 = E$, and thus with $M \geqq (E)^{\frac{1}{2}}$, (24) can be satisfied for $p = q = 2$.

The case of $p = \infty$, $q = 1$ in (24) implies the most stringent condition on $F_i'(\omega)$ because from equation (18) it is evident that

$$|| F_i' ||_p \leqq || F_i' ||_\infty \tag{25}$$

for all $p \geqq 1$. It is clear, for example, that for a sinusoidal input of amplitude $A \leqq M$ and arbitrary frequency ω_0, we must have $| F_i'(\omega) | \leqq 1$ for all ω (that is, $|| F_i' ||_\infty \leqq 1$) to ensure that $| v_i(n) | \leqq M$ for all n. However, a sinusoidal input sequence $\{u(n)\}$ is not absolutely summable, and thus $U^*(z)$ as defined in equation (14) does not exist in this case. This difficulty may be circumvented, as is common in Fourier analysis, by assuming a finite sequence of length N and then passing to the limit as $N \to \infty$. The resulting (Fourier) transform of $\{u(n)\}$ is of the form

$$U_0(\omega) = \frac{A}{2} e^{i\theta}[\delta(\omega - \omega_0) + \delta(\omega - \omega_s + \omega_0)], \qquad (0 \leqq \omega \leqq \omega_s) \tag{26}$$

where $\delta(\omega)$ is the familiar Dirac delta function defined by

$$\begin{aligned} \delta(\omega) &= 0, \qquad \omega \neq 0, \\ \int_{-\infty}^{\infty} \delta(\omega)\, d\omega &= 1. \end{aligned} \tag{27}$$

$U_0(\omega)$ is, of course, periodic in ω with period ω_s. From equations (18a), (26), and (27), we immediately have that $\| U_0 \|_1 = A \leqq M$, and thus with $p = \infty$, (24) is applicable for sinusoidal input sequences, as expected.

V. RANDOM INPUT CASE

In the case of random input sequences, (24) is not directly applicable because the z-transform $U^*(z)$ is not defined. Similar conditions may be obtained, however, by considering the discrete autocorrelation function $\varphi(\cdot)$, defined for a (wide-sense) stationary sequence $\{w(n)\}$ by

$$\varphi_w(m) = E[w(n)w(n + m)] \tag{28}$$

where $E[\cdot]$ is the expected-value operator. A z-transform $\Phi_w^*(z)$ may be defined for the sequence $\{\varphi_w(m)\}$ as in equation (14) with an inverse transform as in (16). Assuming ergodicity and a zero mean ($E[w(n)] = 0$) for $\{w(n)\}$, we immediately have from equation (28) that the variance, or total average power, of $\{w(n)\}$ is given by

$$\varphi_w(0) = E[w^2(n)] = \sigma_w^2, \tag{29}$$

and from equation (16) we also have

$$\varphi_w(0) = \frac{1}{2\pi j} \oint_\Gamma \Phi_w^*(z) z^{-1}\, dz. \tag{30a}$$

Letting Γ be the unit circle ($z = e^{j\omega T}$), equations (29) and (30a) imply that

$$\sigma_w^2 = \frac{1}{\omega_s} \int_0^{\omega_s} \Phi_w(\omega)\, d\omega. \tag{30b}$$

Hence, from equation (8) we see that $\Phi_w(\omega)$ is just the power-density spectrum of the sequence $\{w(n)\}$.

For an input sequence $\{u(n)\}$ whose autocorrelation function has the z-transform $\Phi_u^*(z)$, it is well-known that the corresponding transform for the output $\{v_i(n)\}$ is given by

$$\Phi_{v_i}^*(z) = F_i^*(z) F_i^*(z^{-1}) \Phi_u^*(z) \tag{31a}$$

or

$$\Phi_{v_i}(\omega) = | F_i(\omega) |^2 \Phi_u(\omega). \tag{31b}$$

Equations (29) through (31) imply then that

$$\sigma_{v_i}^2 = \frac{1}{\omega_s} \int_0^{\omega_s} | F_i(\omega) |^2 \Phi_u(\omega) \, d\omega. \tag{32}$$

Since equation (32) is of the same basic form as (17), a derivation similar to that leading to (22) must yield the following relations for $p, q \geqq 1$:

$$\sigma_{v_i}^2 \leqq || F_i^2 ||_p \cdot || \Phi_u ||_q , \qquad \left(\frac{1}{p} + \frac{1}{q} = 1\right) \tag{33a}$$

or, from equation (17),

$$\sigma_{v_i}^2 \leqq || F_i ||_{2p}^2 \cdot || \Phi_u ||_q , \qquad \left(\frac{1}{p} + \frac{1}{q} = 1\right). \tag{33b}$$

Two cases of (33) are of particular interest, namely

$$\sigma_{v_i}^2 \leqq || F_i ||_2^2 \cdot || \Phi_u ||_\infty \tag{34}$$

and

$$\sigma_{v_i}^2 \leqq || F_i ||_\infty^2 \cdot || \Phi_u ||_1 . \tag{35}$$

In view of equation (25), we see that (34) implies the most stringent condition on the input spectrum $\Phi_u(\omega)$, whereas (35) yields the most stringent condition on the transfer function $F_i(\omega)$. From (34) and (30b), for example, we have that if the input power-density spectrum is "white" [that is, $\Phi_u(\omega) = \sigma_u^2$ for all ω], then $\sigma_{v_i}^2 \leqq || F_i ||_2^2 \sigma_u^2$. Hence, if the input sequence $\{u(n)\}$ is a Gaussian process,[19] the node output sequence $\{v_i(n)\}$ will overflow no more (in percentage of time) than does the input, provided only that

$$|| F_i' ||_2 \leqq 1. \tag{36}$$

The inequality in (35) requires, on the other hand, that for an input sinusoid of arbitrary amplitude and frequency, $F_i'(\omega)$ must satisfy

$$|| F_i' ||_\infty \leqq 1 \tag{37}$$

to ensure against overflow, as we have seen earlier from (24).

To summarize, dynamic-range constraints of the form

$$|| F_i' ||_p \leqq 1, \qquad p \geqq 1 \tag{38}$$

have been derived for both deterministic and random inputs, where

$F_i'(\omega)$ is the (scaled) transfer response from the filter input to the i^{th} branch node and $|| \cdot ||_p$ denotes the L_p norm defined in equation (18). For a deterministic input with amplitude spectrum $U(\omega)$, (38) assumes that

$$|| U ||_q \leqq M, \qquad q = \frac{p}{p-1}, \tag{39}$$

where M is the maximum allowable signal amplitude. For a random input, on the other hand, the use of (38) requires appropriate conditions on $|| \Phi_u ||_r$, $r = p/(p-2)$ and $p \geqq 2$, where $\Phi_u(\omega)$ is the power-density spectrum of the input sequence.

The effect of (38) and (39) is to bound the mean absolute value of the amplitude spectrum at the i^{th} branch node (that is, $|| V_i ||_1$) which, in turn, bounds the peak signal amplitude at that node. The use of (38) in conjunction with (33), however, bounds only the average power at the i^{th} branch node, and thus the relationship between this average power and the peak signal amplitude at the node must also be determined in order to provide an effective dynamic-range constraint.

VI. TRANSPOSE SYSTEMS

In the evaluation of different circuit configurations for a given digital filter, a useful concept relating certain of these configurations is that of "transpose configurations". This relationship is a general property of linear graphs[20] and will be presented here in terms of a state-variable formulation.

The general state equations for a linear, time-invariant discrete system are given by[21]

$$\begin{aligned} \mathbf{x}(n+1) &= A\mathbf{x}(n) + B\mathbf{u}(n), \\ \mathbf{y}(n) &= C\mathbf{x}(n) + D\mathbf{u}(n) \end{aligned} \tag{40}$$

where $\mathbf{x}(n)$ is an N-dimensional vector describing the state of the system at time $t = nT$, $\mathbf{u}(n)$ is the corresponding J-dimensional input vector, $\mathbf{y}(n)$ is the corresponding I-dimensional output vector, and A, B, C, and D are fixed parameter matrices of the appropriate dimensions relating the input, state, and output vectors as given by equation (40). The $(N + I) \times (N + J)$ matrix S defined by

$$S = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{41}$$

provides a convenient single parameter matrix which describes the complete discrete system.

A transfer function matrix $\mathfrak{IC}_S^*(z)$ may be defined for the system (described by) S relating the input and output vector sequences $\{\mathbf{u}(n)\}$ and $\{\mathbf{y}(n)\}$ by

$$\mathbf{Y}^*(z) = \mathfrak{IC}_S^*(z)\mathbf{U}^*(z) \tag{42}$$

where $\mathbf{U}^*(z)$ and $\mathbf{Y}^*(z)$ are the vector z-transforms of $\{\mathbf{u}(n)\}$ and $\{\mathbf{y}(n)\}$, respectively. $\mathfrak{IC}_S^*(z)$ is readily shown to be given by[21]

$$\mathfrak{IC}_S^*(z) = C(zI - A)^{-1}B + D \tag{43}$$

where $(\cdot)^{-1}$ denotes the matrix inverse and I is the N-dimensional identity matrix.

Consider now a new system which is described by the parameter matrix S^t, that is,

$$S^t = \begin{bmatrix} A^t & C^t \\ B^t & D^t \end{bmatrix} \tag{44}$$

where $(\cdot)^t$ denotes the matrix transpose. From equations (41) and (43) it is easily seen that the transfer function matrix for the new system S^t is given by

$$\begin{aligned} \mathfrak{IC}_{S^t}^*(z) &= B^t(zI - A^t)^{-1}C^t + D^t \\ &= [\mathfrak{IC}_S^*(z)]^t. \end{aligned} \tag{45}$$

Thus, the transfer function matrix for the system S^t is simply the transpose of the transfer function matrix for the system S. That is, the element $H_{ij}^*(z)$ from $\mathfrak{IC}_S^*(z)$, which is the transfer function from the j^{th} input to the i^{th} output of system S, equals the element $H_{ji}^{*t}(z)$ from $\mathfrak{IC}_{S^t}^*(z)$, that is, the transfer function from the i^{th} input to the j^{th} output of S^t. Note also that while the system S has a total of J inputs and I outputs, the system S^t has I inputs and J outputs.

The concept of transpose systems will be particularly useful to us in conjunction with the digital-filter model introduced in Section III and depicted in Fig. 1. Defining the input and output vectors for the filter by

$$\mathbf{u}(n) = \begin{bmatrix} u(n) \\ e_1(n) \\ \vdots \\ e_J(n) \end{bmatrix} \quad \text{and} \quad \mathbf{y}(n) = \begin{bmatrix} y(n) \\ v_1(n) \\ \vdots \\ v_I(n) \end{bmatrix} \tag{46}$$

respectively, the transfer function matrix for the filter is given by

$$\mathfrak{IC}^*(z) = \begin{bmatrix} H^*(z) & G_1^*(z) & \cdots & G_J^*(z) \\ F_1^*(z) & \text{——} & \text{——} & \text{——} \\ \vdots & \text{——} & \text{——} & \text{——} \\ F_I^*(z) & \text{——} & \text{——} & \text{——} \end{bmatrix} \tag{47}$$

where the specific expressions for the elements in other than the first row and first column are unimportant for our purposes. By equation (45), the transfer function matrix for the corresponding transpose system is then simply

$$\mathfrak{IC}_t^*(z) = \begin{bmatrix} H^*(z) & F_1^*(z) & \cdots & F_I^*(z) \\ G_1^*(z) & \text{——} & \text{——} & \text{——} \\ \vdots & \text{——} & \text{——} & \text{——} \\ G_J^*(z) & \text{——} & \text{——} & \text{——} \end{bmatrix}. \tag{48}$$

Note, in particular, that the transfer function from input-1 to output-1 [that is, $H^*(z)$, the ideal transfer function from filter input to filter output] is the same for both systems.

As discussed more fully in Ref. 1, the circuit configuration realizing a given system S is not necessarily unique, and hence neither is the configuration for the transpose system S^t. However, given a particular configuration for the system S, a unique "transpose configuration", which realizes S^t, may be derived from the given configuration for S by simply reversing the direction of all branches in the given network! In particular, then, all delays and constant multipliers remain the same except for the change in direction. All summation nodes in the given configuration become branch nodes in the transpose configuration, and all branch nodes become summation nodes. Likewise, all inputs in the given configuration become outputs in the transpose configuration, and all outputs become inputs.†

That the transpose configuration defined above actually realizes the transpose system S^t is easily seen by considering the state equations in (40). The constant multiplier(s) corresponding to the element d_{ij} of the matrix D and relating the j^{th} input and the i^{th} output of the original configuration must relate the i^{th} input and the j^{th} output of the transpose

† Note that the transpose system S^t is fundamentally different from the "adjoint" system[22] because, although the signal flow is reversed in both, the transpose system does not run "backwards in time."

configuration, and thus $d_{ij} = d^t_{ji}$ for all i and j. The multiplier(s) corresponding to the element b_{ij} of B and relating the j^{th} input and the i^{th} state of the original configuration must, on the other hand, relate the i^{th} state and the j^{th} output of the transpose configuration, and thus $b_{ij} = c^t_{ji}$ for all i and j. Similarly, $c_{ij} = b^t_{ji}$ for all i and j. Finally, the multiplier(s) corresponding to a_{ij} and relating $x_j(n)$ and $x_i(n+1)$ in the original configuration must, in the transpose configuration, relate $x_i(n)$ and $x_j(n+1)$, and thus $a_{ij} = a^t_{ji}$ for all i and j. Therefore, the transpose configuration indeed realizes the system S^t.

VII. AN EXAMPLE: THE DIRECT FORM

To demonstrate the application of the results of the preceding sections, we now evaluate and compare the roundoff-noise outputs from two transpose configurations for a digital filter. The scaling required to satisfy the overflow constraints in (38) is derived, and the effect of this scaling on the output roundoff noise is determined.

The transfer function $H^*(z)$, defined in equation (1) and relating the input and output of the digital filter, may be expressed as a rational function in z of the form[3,4]

$$H^*(z) = \frac{\sum_{i=0}^{N} a_i z^{-i}}{1 + \sum_{i=1}^{N} b_i z^{-i}} = \frac{A^*(z)}{B^*(z)}. \tag{49}$$

Assuming that a_N and b_N are not both zero, N is referred to as the "order" of the filter. There are many different, but equivalent, forms in which equation (49) may be written, with a number of equivalent circuit configurations corresponding to each of these forms (at least two transpose configurations). Those forms such as equation (49) which require the minimum number of multiplications and additions in the general case (that is, $2N+1$ and $2N$, respectively) are referred to as "canonical" forms. In general, however, it is necessary to add additional scaling multipliers to these canonical forms in order to satisfy the overflow constraints in (38).

The form of $H^*(z)$ given in equation (49) is often called the "direct form" of a digital filter. It has been pointed out by Kaiser[6] that use of the direct form is usually to be avoided because of the sensitivity of the roots of higher-order polynomials to small variations (that is, quantization errors) in the polynomial coefficients. The roundoff-noise outputs from the direct form can also be much larger than from other canonical

forms.[15] Nevertheless, the direct form is of theoretical interest, and it provides a convenient illustration of our results. Similar investigations for the two canonical forms most commonly employed in practice—the cascade and parallel forms—are described in Ref. 1.

Two transpose configurations which implement the direct form with scaling are shown in Figs. 3 and 4. These configurations actually realize $H^*(z)$ in the form

$$H^*(z) = \frac{K_k' \sum_{i=0}^{N} {}_k a_i' z^{-i}}{1 + \sum_{i=1}^{N} b_i z^{-i}} \tag{50}$$

where ${}_k a_i' = a_i / K_k'$, and the additional scaling multipliers K_k', $k = 1, 2$, are required to satisfy (38) in the general case. The configuration in Fig. 3 will be designated as form 1 (that is, $k = 1$), and Fig. 4 as form 2 (that is, $k = 2$).

The branch nodes at which overflow constraints are required (because these signals input to multipliers) are indicated by (*). The dynamic-range limitations are obviously satisfied (by assumption) at the input to the filter, but for completeness, an overflow constraint is included there as indicated. The scaled transfer responses ${}_k F_i'(\omega)$ to these nodes are noted in Figs. 3 and 4, and the corresponding unscaled responses ${}_k F_i(\omega)$ apply, of course, when $K_k' = 1$.

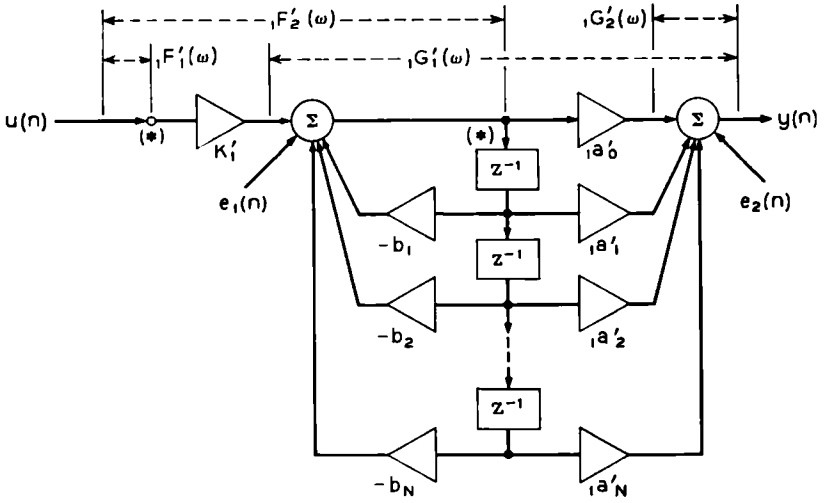

Fig. 3 — Direct form 1 with scaling.

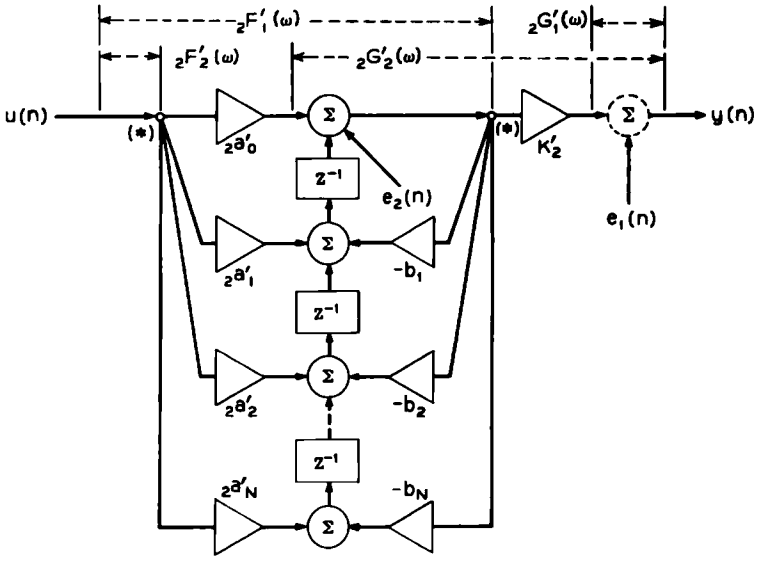

Fig. 4 — Direct form 2 with scaling.

It is intuitively clear that to preserve the greatest possible signal-to-noise ratio, the scaling should reduce the magnitude of ${}_kF_i'(\omega)$ no more than is necessary (or should increase it as much as possible, as the case may be). In other words, ${}_kF_i'(\omega)$ should satisfy

$$\| {}_kF_i' \|_p = 1. \tag{51}$$

This condition will be satisfied if the scaling factors ${}_ks_i$, defined by

$${}_kF_i'(\omega) = {}_ks_i \cdot {}_kF_i(\omega), \tag{52a}$$

are given by

$${}_ks_i = 1/\| {}_kF_i \|_p . \tag{52b}$$

It is readily seen from Figs. 3 and 4 that

$${}_1F_1'(\omega) = {}_2F_2'(\omega) = 1, \tag{53}$$

and hence equation (51) is automatically satisfied for these responses. Of more interest, however, are the responses

$${}_1F_2'(\omega) = \frac{K_1'}{B(\omega)} = K_1' \, {}_1F_2(\omega) \tag{54}$$

and

$$ {}_2F_1'(\omega) = \frac{H(\omega)}{K_2'} = \frac{{}_2F_1(\omega)}{K_2'}. \tag{55} $$

From equations (52), (54), and (55), it follows that (51) is satisfied for these configurations if (and only if)

$$ K_1' = 1/||\ 1/B\ ||_p \tag{56} $$

and

$$ K_2' = ||\ H\ ||_p\ . \tag{57} $$

The rounding-error inputs $e_i(n)$ are also shown in Figs. 3 and 4 along with the transfer responses ${}_kG_i'(\omega)$ from these inputs to the output of the filter. Note that in form 2 (Fig. 4) the error input $e_2(n)$ incorporates the roundoff errors from all of the multipliers except K_2' even though these error sources are separated by delays (z^{-1}). This is done for convenience and is possible because of the assumption of uncorrelated errors from sample to sample and source to source. The noise weights k_i' [see equation (10)] for form 1 are thus

$$ {}_1k_1' = {}_1k_2' = N + 1; \tag{58a} $$

while for form 2,

$$ {}_2k_1' = 1 \quad \text{and} \quad {}_2k_2' = 2N + 1. \tag{58b} $$

The indices i and j of the ${}_kF_i(\omega)$ and ${}_kG_i(\omega)$ have been assigned in such a way that forms 1 and 2 are related as in equations (47) and (48). That is, these unscaled responses satisfy the following equations:

$$ {}_1F_i(\omega) = {}_2G_i(\omega), \qquad i = 1, 2, \tag{59a} $$

$$ {}_1G_j(\omega) = {}_2F_j(\omega), \qquad j = 1, 2. \tag{59b} $$

Note that the scaled responses ${}_kF_i'(\omega)$ and ${}_kG_j'(\omega)$ are not related as in equation (59) because, in general, $K_1' \neq K_2'$. In particular,

$$ {}_1G_1'(\omega) = \frac{H(\omega)}{K_1'} = \left(\frac{K_2'}{K_1'}\right){}_2F_1'(\omega); \tag{60} $$

while

$$ {}_2G_2'(\omega) = \frac{K_2'}{B(\omega)} = \left(\frac{K_2'}{K_1'}\right){}_1F_2'(\omega). \tag{61} $$

However, we do have, as in equation (53), that

$$ {}_1G_2'(\omega) = {}_2G_1'(\omega) = 1. \tag{62} $$

From equations (10) and (53) through (62), the power-spectral densities of the roundoff-noise outputs from these two configurations are thus computed to be

$$ {}_1N_y(\omega) = \sigma_0^2(N+1)\left\{1 + \left\| \frac{1}{B} \right\|_p^2 | H(\omega) |^2\right\} \tag{63a} $$

and

$$ {}_2N_y(\omega) = \sigma_0^2\left\{1 + (2N+1) \| H \|_p^2 \left| \frac{1}{B(\omega)} \right|^2\right\}. \tag{63b} $$

The variances, or total average powers of the output roundoff noise from these configurations are then, from equations (8) and (18), simply

$$ \| {}_1N_y \|_1 = \sigma_0^2(N+1)\left\{1 + \left\| \frac{1}{B} \right\|_p^2 \| H \|_2^2\right\} \tag{64a} $$

and

$$ \| {}_2N_y \|_1 = \sigma_0^2\left\{1 + (2N+1) \| H \|_p^2 \left\| \frac{1}{B} \right\|_2^2\right\}. \tag{64b} $$

The peak noise densities $\| {}_kN_y \|_\infty$ are, on the other hand, bounded by

$$ \| {}_1N_y \|_\infty \leqq \sigma_0^2(N+1)\left\{1 + \left\| \frac{1}{B} \right\|_p^2 \| H \|_\infty^2\right\} \tag{65a} $$

and

$$ \| {}_2N_y \|_\infty \leqq \sigma_0\left\{1 + (2N+1) \| H \|_p^2 \left\| \frac{1}{B} \right\|_\infty^2\right\}. \tag{65b} $$

We now compare direct forms 1 and 2 on the basis of (64) and (65). Although comparisons based on bounds for $\| {}_kN_y \|_\infty$ as in (65) do not, of course, necessarily hold for $\| {}_kN_y \|_\infty$ itself, experimental results have indicated that such comparisons are quite effective qualitatively, and often quantitatively as well.[1] Consider first the expressions in equation (64) for $p = 2$ and in (65) for $p = \infty$ (that is, $\| N_y \|_r$, $r = 1, \infty$, for $p = r + 1$]. In these two cases, the only difference between the (a) and (b) expressions for forms 1 and 2, respectively, are the k_i', as given in equation (58). In particular, for $\| 1/B \|_p^2 \| H \|_p^2 \gg 1$ as is often the case, the $\| N_y \|_r$ for form 1 are approximately half, or 3 db less than, those for form 2. This result simply reflects the fact that only half of the noise sources in form 1 input at other than the filter output; whereas in form 2, all but one input within the filter. Hence, if the gains from these inputs to the output are large, form 1 is preferable to form 2 by up to 3 db.

For $p \neq r + 1$, however, the differences in the k_i' are of secondary importance compared with the potential differences due to the mixture of L_2 and L_∞ norms in (64) and (65). In particular, letting

$$\theta_{pq} = \left\| \frac{1}{B} \right\|_p^2 \| H \|_q^2 , \tag{66a}$$

we immediately see that if $\theta_{\infty 2} \gg \theta_{2\infty}$, then form 2 is better for $p = \infty$ while form 1 is better for $p = 2$. If, on the other hand, $\theta_{\infty 2} \ll \theta_{2\infty}$, then the opposite applies.

To gain insight into the above conditions, we rewrite equation (66a) as

$$\theta_{pq} = \left\| \frac{1}{B} \right\|_p^2 \left\| \frac{A}{B} \right\|_q^2 . \tag{66b}$$

It is then clear that the difference between $\theta_{\infty 2}$ and $\theta_{2\infty}$ is due entirely to the effect of $A(\omega)$ on the L_q norms of $A(\omega)/B(\omega)$ for $q = 2, \infty$ versus the corresponding norms of $1/B(\omega)$. In particular, $A(\omega)$ affects the L_∞ norm in $\theta_{2\infty}$. But the L_∞ norm of a function "concentrates" exclusively on the maximum absolute value of that function; whereas the L_2 norm of a function reflects the r.m.s. absolute value of that function over all argument values. Therefore, the effect of $A(\omega)$ in $\theta_{2\infty}$ results from the alteration of the maxima of $| 1/B(\omega) |$ in $| A(\omega)/B(\omega) |$; while in $\theta_{\infty 2}$, the effect concerns the difference between $| 1/B(\omega) |$ and $| A(\omega)/B(\omega) |$ over all ω.

Intuitively, one expects that the former effect is potentially much greater; that is, in many cases $A(\omega)$ should affect the L_∞ norm in $\theta_{2\infty}$ much more than the L_2 norm in $\theta_{\infty 2}$. In particular, if $| A(\omega) |$ significantly attenuates the maxima of $| 1/B(\omega) |$ [as in a band-rejection filter, for example], then $\theta_{2\infty}$ should be much smaller than $\theta_{\infty 2}$. In this case, form 2 should be used for $p = \infty$, and form 1 for $p = 2$. If, however, $| A(\omega) |$ does not provide such attenuation, then $| A(\omega) |$ must be relatively constant within the band(s) where $| 1/B(\omega) |$ is largest [by the nature of $A(\omega)$], and hence

$$\left\| \frac{A}{B} \right\|_q \approx | A(\omega_0) | \cdot \left\| \frac{1}{B} \right\|_q \tag{67}$$

where ω_0 is a frequency at or near a maximum of $| 1/B(\omega) |$. But then,

$$\theta_{pq} \approx | A(\omega_0) | \left\| \frac{1}{B} \right\|_p \left\| \frac{1}{B} \right\|_q \approx \theta_{qp} , \tag{68}$$

and the difference between direct forms 1 and 2 should be less in this case.

VIII. SUMMARY

The interaction between the roundoff-noise output from a digital filter and the associated dynamic-range limitations has been investigated for the case of uncorrelated rounding errors from sample to sample and from one error source to another. The spectrum of the output roundoff noise from fixed-point implementations was readily shown to be of the form

$$N_y(\omega) = \sigma_0^2 \sum_i k_i' \mid G_i'(\omega) \mid^2 \tag{69}$$

where the $G_i'(\omega)$ are scaled transfer responses from certain "summation nodes" in the digital circuit to the filter output. σ_0^2 is the variance of the rounding errors from each multiplier (or other rounding point), and the k_i' are integers indicating the number of error inputs to the respective summation nodes.

Defining $F_i'(\omega)$ to be the scaled transfer response from the input to the i^{th} "branch node" at which a dynamic-range constraint is required, constraints of the form

$$\| F_i' \|_p \leqq 1 \tag{70}$$

for $p \geqq 1$ were then derived, where $\| F_i' \|_p$ is the L_p norm of the response $F_i'(\omega)$. The appropriate value of p is determined by assumed conditions on the spectra of the input signals to the filter. The effect of (70) is to bound the maximum signal amplitude (for deterministic inputs) or the maximum average power (for random inputs) at the i^{th} branch node.

A state-variable description was employed to formulate the general concept of "transpose configurations" for a digital network and to illustrate the usefulness of this concept in digital-filter synthesis. A particularly important result is that for a given unscaled configuration with transpose responses $F_i(\omega)$ and $G_i(\omega)$, as described above, the responses $F_i'(\omega)$ and $G_i'(\omega)$ for the corresponding transpose configuration are given by

$$F_i'(\omega) = G_i(\omega) \quad \text{and} \quad G_i'(\omega) = F_i(\omega). \tag{71}$$

Hence, although the overall transfer functions for these two configurations are the same, their roundoff-noise outputs can be quite different, in general. The transpose configuration is obtained by simply reversing the direction of all branches in the given network configuration, and the poles and zeros of the network are thus realized in reverse order in the transpose configuration.

To illustrate these results, the roundoff-noise spectra $N_y(\omega)$ for two

transpose configurations for the direct form of a digital filter were calculated and compared. The direct form should usually be avoided in practice,[6] but it is still of theoretical interest and provides a convenient example of our general approach. Using a very natural assignment of the indices i and j for the unscaled $F_i(\omega)$ and $G_j(\omega)$, equation (69) was shown to be of the form

$$N_y(\omega) = \sigma_0^2\left\{k'_{M+1} + \sum_{i=1}^{M} k'_i \,||\, F_i \,||_p^2 \mid G_i(\omega) \mid^2\right\} \tag{72}$$

for these (scaled) configurations for the direct form, where M is the number of error inputs at other than the output of the filter. Hence, the variance, or total average power, of the output roundoff noise is simply

$$\sigma_y^2 = \sigma_0^2\left\{k'_{M+1} + \sum_{i=1}^{M} k'_i \,||\, F_i \,||_p^2 \,||\, G_i \,||_2^2\right\}; \tag{73}$$

while the peak spectral density $||\, N_y \,||_\infty$ is bounded by

$$||\, N_y \,||_\infty \leqq \sigma_0^2\left\{k'_{M+1} + \sum_{i=1}^{M} k'_i \,||\, F_i \,||_p^2 \,||\, G_i \,||_\infty^2\right\}. \tag{74}$$

Identical expressions to (72) through (74) can also be derived for the parallel and cascade forms of a digital filter.[1] The relationship between the noise outputs of corresponding transpose configurations is immediately indicated by (71) through (74) [although, in general, $k'_i \neq k'^t_i$].

REFERENCES

1. Jackson, L. B., *An Analysis of Roundoff Noise in Digital Filters*, Sc.D. Thesis, Stevens Institute of Technology, Hoboken, New Jersey (1969).
2. McDonald, H. S., "Impact of Large-Scale Integrated Circuits on Communication Equipment," *Proc. of the National Electronics Conf., 24* (December 1968), pp. 569–72.
3. Rader, C. M., and Gold, B., *Digital Processing of Signals*, New York: McGraw-Hill, 1969, pp. 1–130.
4. Kaiser, J. F., "Digital Filters," *System Analysis by Digital Computer*, New York: Wiley, 1966, pp. 218–85.
5. Rader, C. M., and Gold, B., "Digital Filter Design Techniques in the Frequency Domain," *Proc. IEEE, 55*, No. 2 (February 1967), pp. 149–71.
6. Kaiser, J. F., "Some Practical Considerations in the Realization of Linear Digital Filters," *Proc. Third Annual Allerton Conf. on Circuit and System Theory*, Monticello, Illinois, October 1965, pp. 621–33.
7. Knowles, J. B., and Olcayto, E. M., "Coefficient Accuracy and Digital Filter Response," *IEEE Trans. on Circuit Theory, CT-15*, No. 1 (March 1968), pp. 31–41.
8. Bennett, W. R., Spectra of Quantized Signals," B.S.T.J., *27*, No. 3 (July 1948), pp. 446–72.
9. Kaneko, T., and Liu, B., "Round-off Error of Floating-Point Digital Filters," *Proc. Sixth Annual Allerton Conf. on Circuit and System Theory*, Monticello, Illinois, October 1968, pp. 219–27.

10. Weinstein, C., and Oppenheim, A. V., "A Comparison of Roundoff Noise in Floating-Point and Fixed-Point Digital-Filter Realizations," *Proc. IEEE, 57,* No. 6 (June 1969), pp. 1181–3.
11. Oppenheim, A. V., "Block-Floating-Point Realization of Digital Filters," MIT Lincoln Laboratory, Technical Note 1969–19 (March 20, 1969).
12. Knowles, J. B., and Edwards, R., "Effects of a Finite-Word-Length Computer in a Sampled-Data Feedback System," *Proc. IEE, 112,* No. 6 (June 1965), pp. 1197–1207.
13. Gold, B., and Rader, C. M., "Effects of Quantization Noise in Digital Filters," *Proc. AFIPS,* 1966 SJCC, pp. 213–19.
14. Jackson, L. B., Kaiser, J. F., and McDonald, H. S., "An Approach to the Implementation of Digital Filters," *IEEE Trans. on Audio and Electroacoustics, AU-16,* No. 3 (September 1968), pp. 413–21.
15. Edwards, R., Bradley, J., and Knowles, J. B., "Comparison of Noise Performance of Programming Methods in the Realization of Digital Filters," *Proc. of the Symposium on Computer Processing in Communications, XIX,* PIB–MRI Symposia Series (1969).
16. Papoulis, A., "Limits on Bandlimited Signals," *Proc. IEEE, 55,* No. 10 (October 1967), pp. 1677–85.
17. Rice, J. R., *The Approximation of Functions,* Reading, Mass.: Addison-Wesley, 1964, pp. 4–10.
18. Bachman, G., and Naria, L., *Functional Analysis,* New York: Academic Press, 1966, pp. 110–11.
19. Davenport, W. B., Jr., and Root, W. L., *Random Signals and Noise,* New York: McGraw-Hill, 1958, pp. 154–7.
20. Mason, S. J., and Zimmerman, H. J., *Electronic Circuits, Signals and Systems,* New York: Wiley, 1960, pp. 122–3.
21. Freeman, H., *Discrete-Time Systems,* New York: Wiley, 1965, pp. 19–27.
22. Laning, J. H., Jr., and Battin, R. H., *Random Processes in Automatic Control,* New York: McGraw-Hill, 1956, pp. 239–43.