Realization of Digital Filters Using Block-Floating-Point Arithmetic

ALAN V. OPPENHEIM, Member, IEEE
Massachusetts Institute of Technology
Lincoln Laboratory
Lexington, Mass. 02173
and
Department of Electrical Engineering, and
Research Laboratory of Electronics
Cambridge, Mass.

Abstract

Recently, statistical models for the effects of roundoff noise in fixed-point and floating-point realizations of digital filters have been proposed and verified, and a comparison between these realizations presented. In this paper a structure for implementing digital filters using block-floating-point arithmetic is proposed and a statistical analysis of the effects of roundoff noise is carried out. On the basis of this analysis, block-floating-point is compared to fixed-point and floating-point arithmetic with regard to roundoff noise effects.

Manuscript received August 8, 1969.

This work was sponsored by the Department of the Air Force.

Introduction

Recently, statistical models for the effects of roundoff noise in fixed-point and floating-point realizations of digital filters have been proposed and verified, and a comparison between these realizations has been suggested [1]–[3]. In general terms, the comparison revolves around the fact that while floating-point arithmetic has a larger dynamic range than fixed-point, the latter is more accurate when the full register length can be utilized. Because of the limited dynamic range of fixed-point arithmetic, for high-gain filters, the input signal must be attenuated to prevent overflow in the output. Thus, for sufficiently high gain, floating-point arithmetic leads to lower noise-to-signal ratio than fixed point. On the other hand, floating-point arithmetic implies a more complex hardware structure than fixed-point arithmetic.

An alternative realization, block-floating-point, has some of the advantages of both fixed point and floating point. In this paper a structure for implementing digital filters using block-floating-point arithmetic is proposed and a statistical analysis of the effects of roundoff noise presented. On the basis of this analysis, block-floating-point is compared to fixed-point and floating-point arithmetic with regard to roundoff noise effects.

A Structure for Block-Floating-Point Realization

In block-floating-point arithmetic the input and filter states (i.e., the outputs of the delay registers) are jointly normalized before the multiplications and adds are performed using fixed-point arithmetic. The scale factor obtained during the normalization is then applied to the final output to produce a fixed-point result. To illustrate, consider a first-order filter described by the difference equation

$$y_n = x_n + a_1 y_{n-1}. \tag{1}$$

For convenience we will treat all numbers as fixed-point fractions. To perform the computation in a block-floating-point manner, we define

$$A_n = \frac{1}{\mathrm{IP}[\max\{|x_n|, |y_{n-1}|\}]} \tag{2}$$

where IP$[M]$ is used to denote the integer power of two such that $M < \mathrm{IP}(M) \leq 2M$, i.e., with M written as $M = 2^m \cdot P$ with P between $\frac{1}{2}$ and 1, $\mathrm{IP}(M) = 2^m$. For M a fraction, 2^m is less than or equal to unity so that A_n is greater than or equal to unity. Thus A_n represents the power-of-two scaling which will jointly normalize x_n and y_{n-1}. Thus with block-floating-point we can compute y_n as

$$y_n = \frac{1}{A_n}[A_n x_n + a_1 A_n y_{n-1}] \tag{3}$$

where the multiplications and addition in (3) are carried out in a fixed-point manner.

Because of the recursive nature of the computation for a digital filter, it is advantageous to modify (3) as

$$\hat{y}_n = A_n x_n + a_1 \Delta_n w_{1n} \tag{4}$$

with

$$w_{1n} = A_{n-1} y_{n-1}$$

$$y_n = \frac{1}{A_n} \hat{y}_n$$

and

$$\Delta_n = A_n / A_{n-1}.$$

The difference between (3) and (4) is meant to imply that the number $A_n y_n$ rather than y_n is stored in the delay register of the filter. Because of (2), $A_n y_n$ is always more accurate (or as accurate) as y_n since multiplication by A_n corresponds to a left shift of the register.

A disadvantage with (4) is that y_{n-1} must be available to compute A_n, and Δ_n must then be obtained from A_n and A_{n-1}. An alternative is represented by the set of equations

$$\hat{y}_n = \Delta_n \hat{x}_n + a_1 \Delta_n w_{1n} \tag{5a}$$

with

$$\hat{x}_n = A_{n-1} x_n \tag{5b}$$

and

$$\Delta_n = A_n / A_{n-1} = \frac{1}{\mathrm{IP}[\max\{|\hat{x}_n|, |w_{1n}|\}]} \cdot \tag{5c}$$

In this case, we first scale x_n by A_{n-1} to form $\hat{x}_n$ and then determine the incremental scaling using (5c). As in (4), the scaled value $\hat{y}_n$ is stored in the delay register and the output value y_n is determined from $\hat{y}_n$. If we consider the general case of an Nth order filter of the form

$$y_n = x_n + a_1 y_{n-1} + a_2 y_{n-2} + \cdots + a_N y_{n-N},$$

then the block-floating-point realization corresponding to (5) and represented in the direct form is depicted in Fig. 1. For the general case,

$$\Delta_n = \frac{1}{\mathrm{IP}\,[\max\{|\hat{x}_n|, |w_{1n}|, |w_{2n}|, \cdots, |w_{Nn}|\}]} \tag{6}$$

and

$$A_n = \frac{1}{\mathrm{IP}\,[\max\{|x_n|, |y_{n-1}|, |y_{n-2}|, \cdots, |y_{n-N}|\}]} \tag{7}$$
$$= A_{n-1}\Delta_n.$$

As an additional consideration, we note that because of the block normalization, there is the possibility of overflow in the addition, which cannot be avoided by an attenuation of the input. This possibility of overflow can be avoided by decreasing the normalization constant A_n by a fixed amount. Thus we modify (6) and (7) as

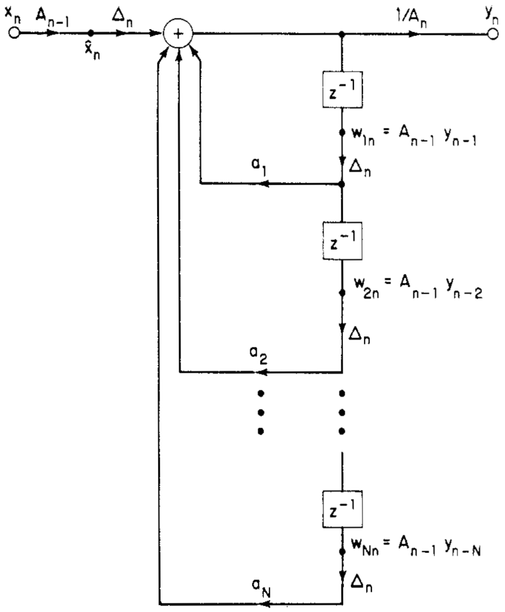

Fig. 1. Network for block-floating-point realization of an Nth-order filter.

$$\Delta_n = \frac{1}{\alpha\,\mathrm{IP}\,[\max\{|\hat{x}_n|, |w_{1n}|, |w_{2n}|, \cdots, |w_{Nn}|\}]} \tag{6'}$$

and

$$A_n = \frac{1}{\alpha\,\mathrm{IP}\,[\max\{|x_n|, |y_{n-1}|, |y_{n-2}|, \cdots, |y_{n-N}|\}]} \tag{7'}$$

where α is a constant that may be changed depending on the filter to be implemented. In a first-order filter, for example, α need never be greater than two.

The Effect of Roundoff Noise in Block-Floating-Point Filters

In evaluating the performance of the block-floating-point realization in the presence of roundoff noise, we will restrict attention to the implementation of (5) and Fig. 1 for the first- and second-order cases. We will assume that no roundoff occurs in the computation of $\hat{x}_n$ from x_n and the subsequent multiplication by Δ_n. Since A_{n-1} and $A_{n-1}\Delta_n$ are always nonnegative powers of two, that is, they always correspond to a positive scaling, the above assumption corresponds to allowing more bits in the representation of the intermediate variable $\hat{x}_n$. This is reasonable if we take the attitude that it is primarily in the variables used in the arithmetic computations that the register length is important.

For the first-order case, roundoff noise is introduced in the multiplication of w_{1n} by Δ_n, the multiplication by a_1, and the final multiplication by $1/A_n$. The effects of multiplier roundoff will be modeled by representing the roundoff by additive white noise sources. We consider, for convenience, the fixed-point numbers in the registers to represent signed fractions, with the register length excluding sign denoted by t bits. Each of the roundoff noise generators is assumed to be white, mutually independent and independent of the input, and to have a variance $\sigma_\epsilon{}^2$ equal to $(1/12)\cdot 2^{-2t}$. The network for the first-order filter including the noise sources representing roundoff error is presented in Fig. 2(A). In Fig. 2(B) an equivalent representation is shown, where the noise sources are at the filter input. If we consider the input to be a stationary random signal, then the noise source ξ_n will be white stationary random noise with variance

$$\sigma_\xi{}^2 = \sigma_\epsilon{}^2(1 + a_1{}^2)k^2 \qquad (8)$$

where k^2 denotes the expected value of $(1/A_n)^2$. Letting η_n denote the noise in the filter output due to the noise ξ_n, the variance of the output noise η_n will be

$$\sigma_\eta{}^2 = \sigma_\xi{}^2 \sum_{n=0}^{\infty} h_n{}^2 + \overline{\epsilon_{3n}{}^2} = \sigma_\epsilon{}^2\left[1 + \frac{1 + a_1{}^2}{1 - a_1{}^2}k^2\right]. \qquad (9)$$

This result is derived by observing that in Fig. 2(B) the transmission from the noise source ξ_n to the output is that of a first-order filter with unit sample response h_n given by $h_n = a^n$. For the case of a second-order filter a similar procedure can be followed. Fig. 3(A) shows a second-order filter with the roundoff noise sources included. In Fig. 3(B) an equivalent representation is shown, where equivalent noise sources are introduced at the filter input. Again, considering the input to be a stationary random signal, then

$$\begin{aligned}\sigma_\xi{}^2 &= \overline{\left(\frac{1}{A_n}\right)^2}\sigma_\epsilon{}^2[4r^2\cos^2\theta + 2 + r^4] \\ &\quad + r^4\sigma_\epsilon{}^2\overline{\left(\frac{1}{A_{n-1}}\right)^2} \qquad (10) \\ &= k^2\sigma_\epsilon{}^2[4r^2\cos^2\theta + 2 + 2r^4]\end{aligned}$$

where we assume that the mean-square values of $(1/A_n)$ and $(1/A_{n-1})$ are equal. Hence the variance of the output noise η_n is

$$\sigma_\eta{}^2 = \sigma_\epsilon{}^2 + k^2\sigma_\epsilon{}^2G[4r^2\cos^2\theta + 2 + 2r^4] \qquad (11)$$

where

$$G = \left(\frac{1 + r^2}{1 - r^2}\right)\frac{1}{1 + r^4 - 4r^2\cos^2\theta + 2r^2}. \qquad (12)$$

Experimental Verification

To verify the validity of (9) and (11) the values of k^2 were measured and the values of $\sigma_\xi{}^2$ computed from (9) and (11) using these measured values. These results were taken as the theoretical results since they incorporate the

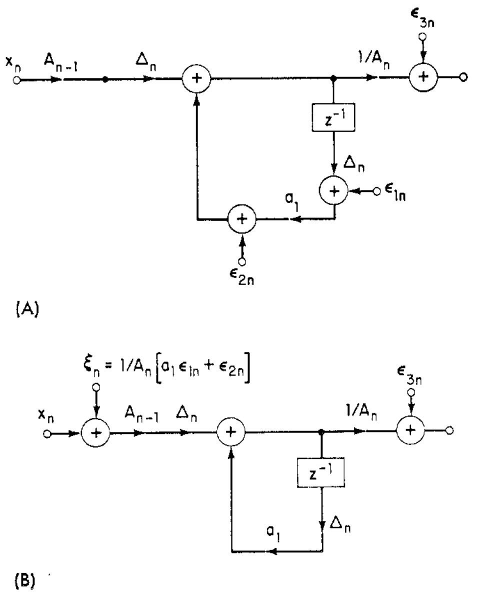

Fig. 2. (A) Noise model for first-order filter. (B) Equivalent noise model.

Fig. 3. (A) Noise model for second-order filter. (B) Equivalent noise model.

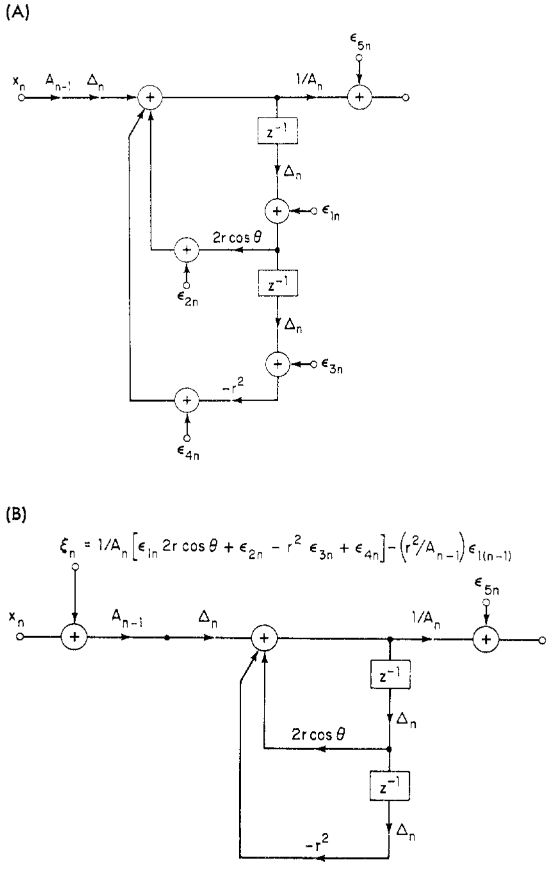

TABLE I

Measured Values of k^2 and Theoretical and Experimental Values of Output Noise Variance for a First-Order Filter with White Noise Input in the Range $|x_n| \leq 1/16$

a_1	k^2	(1/2) $\log_2[2^{2t}\sigma_\eta^2]$(bits) Theoretical	Experimental
0.1	0.0137	−1.7825	−1.7808
0.3	0.0161	−1.7787	−1.7737
0.4	0.0175	−1.7753	−1.7740
0.55	0.0200	−1.7760	−1.7640
0.7	0.0253	−1.7410	−1.7403
0.9	0.0612	−1.4613	−1.4457
0.95	0.1162	−0.9386	−0.9328

TABLE II

Measured Values of k^2 and Theoretical and Experimental Values of Output Noise Variance for a Second-Order Filter with White Noise Input in the Range $|x_n| \leq 1/128$

r	θ	k^2	(1/2) $\log_2[2^{2t}\sigma_\eta^2]$(bits) Theoretical	Experimental
0.55	22.5	0.00288	−1.774	−1.7539
0.55	45.0	0.00186	−1.7861	−1.7767
0.55	67.5	0.00151	−1.7893	−1.7822
0.7	22.5	0.00499	−1.7168	−1.6755
0.7	45.0	0.00248	−1.7780	−1.7661
0.7	67.5	0.00177	−1.7872	−1.7813
0.9	22.5	0.01667	−1.0268	−1.0078
0.9	45.0	0.00581	−1.6824	−1.6494
0.9	67.5	0.00385	−1.7576	−1.7242
0.95	22.5	0.03113	−0.2425	−0.2635
0.95	45.0	0.0110	−1.2369	−1.3760
0.95	67.5	0.00708	−1.6696	−1.5968
0.99	22.5	0.15527	2.0165	2.0199
0.99	45.0	0.04688	0.1605	0.3076
0.99	67.5	0.03125	−0.6065	−0.1395

assumptions of the model. The variance of the roundoff noise ξ_n was then measured experimentally. This was done by simulating the block-floating-point filter with a signed mantissa of 12 bits and comparing the output values with the output of an identical filter simulated with 36-bit fixed-point arithmetic. In all of these measurements the input was white noise with a uniform amplitude distribution. For the first-order filter, the value of α in (6′) and (7′) was taken as two. For the second-order filter, the value of α was taken as four.

In Table I, measured values of k^2 and the theoretical and experimental values of the variance of the roundoff noise for the first-order case are given. The input is white noise in the range $|x_n| \leq \frac{1}{16}$. In a similar manner, theoretical and experimental results for the second-order case are summarized in Table II. The input is white noise in the range $|x_n| \leq \frac{1}{128}$.

A Comparison of Block-Floating-Point, Floating-Point, and Fixed-Point Realizations

Using the model presented in the previous section, the block-floating-point realization of digital filters can be compared with fixed-point and floating-point realizations. The comparison to be presented here will be on the basis of the output noise-to-signal ratio when the input is a random signal with a flat spectrum, using results presented by Gold and Rader [1], Kaneko and Liu [2], and Weinstein and Oppenheim [3]. With σ_η^2 denoting the variance of the roundoff noise as it appears in the output, we have for the first-order filter

$$\text{fixed point: } \sigma_\eta^2 = \frac{1}{12} \cdot 2^{-2t} \frac{1}{1 - a_1^2} \tag{13}$$

$$\text{floating point: } \sigma_\eta^2 = 0.23 \cdot 2^{-2t} \frac{1 + a_1^2}{1 - a_1^2} \sigma_y^2 \tag{14}$$

and for the second-order filter

$$\text{fixed point: } \sigma_\eta^2 = \frac{1}{6} \cdot 2^{-2t} G \tag{15}$$

$$\text{floating point: } \sigma_\eta^2 = 0.23 \cdot 2^{-2t} \cdot \left[1 + G\left(3r^4 + 12r^2 \cos^2\theta - \frac{16r^4 \cos^2\theta}{1 + r^2}\right)\right] \sigma_y^2 \tag{16}$$

where t is the number of bits in the mantissa, not including sign, σ_y^2 is the variance of the output signal, and G is given by (12). In the fixed-point case the output noise is independent of the output signal variance, and in the floating-point case the output noise is proportional to the output signal variance. The expression for block-floating-point noise has a term independent of the signal and a term which depends on the signal through the factor k^2. In both the fixed-point and block-floating-point cases, the dynamic range for the output is constrained by the register length. Consequently, as the filter gain increases, the input must be scaled down to prevent the output from overflowing the register length. Since the output is given by

$$y_n = \sum_{k=0}^{\infty} h_k x_{n-k},$$

then

$$|y_n| \leq \max(|x_n|) \sum_{k=0}^{\infty} |h_k|.$$

To insure that the output fits within a register length, we require that, with x_n and y_n interpreted as fractions,

$$|y_n| \leq 1$$

so that

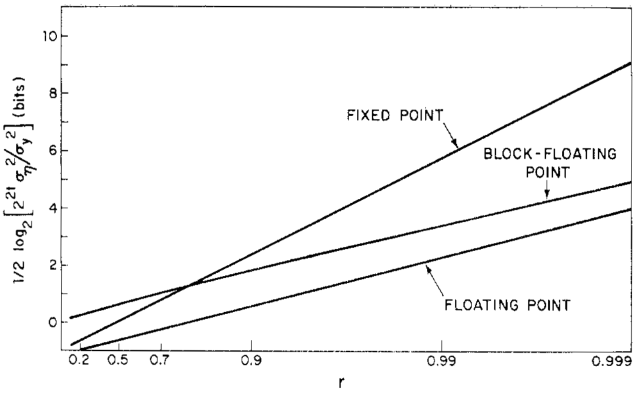

Fig. 4. Comparison of noise-to-signal ratios for first-order filter using fixed-point, floating-point, and block-floating-point arithmetic.

$$\frac{-1}{\sum_{k=0}^{\infty}|h_k|} \le x_n \le \frac{1}{\sum_{k=0}^{\infty}|h_k|}\,. \tag{17}$$

With this constraint on the input, we can then compute an output noise-to-signal ratio for fixed-point, floating-point, and block-floating realizations. Specifically, for the first-order case,

$$\left(\frac{\sigma_\eta^2}{\sigma_y^2}\right)_{\text{fixed point}} = \frac{1}{12}\,2^{-2t}\cdot\frac{3}{(1-a_1)^2} \tag{18}$$

$$\left(\frac{\sigma_\eta^2}{\sigma_y^2}\right)_{\text{floating point}} = 0.23\cdot 2^{-2t}\frac{1+a_1^2}{1-a_1^2} \tag{19}$$

$$\left(\frac{\sigma_\eta^2}{\sigma_y^2}\right)_{\text{block floating}} = \frac{1}{12}\cdot 2^{-2t}\left[\frac{3(1-a_1^2)}{(1-a_1)^2} + 3(1+a_1^2)\bar{k}^2\right] \tag{20}$$

where $\bar{k}^2$ is the value for k^2 when x_n is uniformly distributed between plus and minus unity.[1]

In a similar manner, for the second-order case,

$$\left(\frac{\sigma_\eta^2}{\sigma_y^2}\right)_{\text{fixed point}} = \frac{1}{2}\,2^{-2t}\left(\frac{1}{\sin\theta}\sum_{n=0}^{\infty} r^n\,|\sin[(n+1)\theta]|\right)^2 \tag{21}$$

$$\left(\frac{\sigma_\eta^2}{\sigma_y^2}\right)_{\text{floating point}} = (0.23)2^{-2t}\cdot\left[1 + G\left(3r^4 + 12r^2\cos^2\theta - 16\,\frac{r^4\cos^2\theta}{1+r^2}\right)\right] \tag{22}$$

$$\left(\frac{\sigma_\eta^2}{\sigma_y^2}\right)_{\text{block floating}} = \frac{1}{12}\,2^{-2t}\left\{\frac{3}{G}\left[\frac{1}{\sin\theta}\sum_{n=0}^{\infty} r^n\,|\sin(n+1)\theta|\right]^2 + 3\bar{k}^2(2 + 2r^4 + 4r^2\cos^2\theta)\right\}. \tag{23}$$

In Fig. 4, (18), (19), and (20) are compared. In Fig. 5, (21), (22), and (23) are compared. In these figures the noise-to-signal ratios are plotted in bits so that the difference between two of the curves reflects the number of bits that the mantissas should differ by to achieve the same noise-to-signal ratio. In each of the cases, the difference between floating-point and block-floating-point is approximately constant as the filter gain (or the proximity of the poles to the unit circle) increases. This difference is approximately one bit in the first-order case and two bits in the second-order case. In contrast, the fixed-point noise-to-signal ratio increases at a faster rate than floating-point or block-floating-point, and for low gain is better and for high gain is worse than block-floating-point.

In evaluating the comparison between fixed-point, floating-point, and block-floating-point filter realizations, it is important to note that Figs. 4 and 5 are based only on the mantissa length and do not reflect the additional bits needed to represent the characteristic in either floating-point or block-floating-point arithmetic.

An additional consideration which is not reflected in these curves is that in both fixed-point and block-floating-point the noise-to-signal ratio is computed on the assump-

[1] Equations (20) and (23) are expressed in terms of $\tilde{k}^2$ rather than k^2 to facilitate a later approximation. In deriving these equations, k^2 is given as $\tilde{k}^2$ times the square of the maximum value of the input as dictated by (17). Thus, although expressed in terms of a normalized value for k^2, (20) and (23) are consistent with the constraint that x_n be sufficiently small so that the output is less than unity.

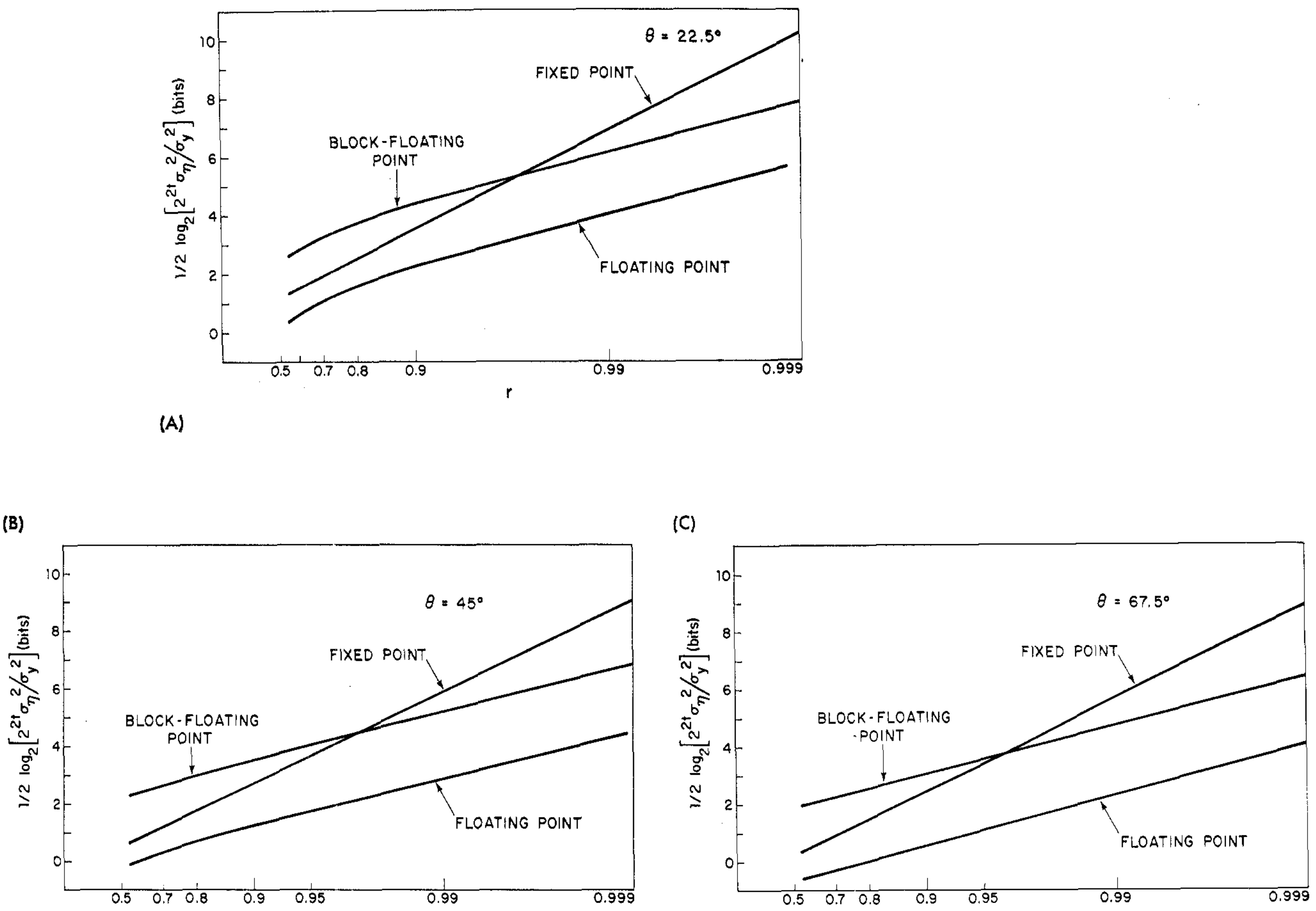

Fig. 5. Comparison of noise-to-signal ratios for second-order filter using fixed-point, floating-point, and block-floating-point arithmetic.

tion that the input signal is as large as possible consistent with the requirement that the output fit within the register length. If the input signal is in fact smaller than permitted, then the noise-to-signal ratio for the fixed-point case will be proportionately higher. For block-floating-point, as the input signal decreases, k^2 decreases, thus reducing the output noise. From (9) and (12) we observe that as the input signal decreases the output noise variance asymptotically approaches σ_ϵ^2.

For the case of high-gain filters, (18) through (23) can be approximated by asymptotic expressions which place in evidence the relationship between them. For the high-gain case, that is, for a_1 close to unity in the first-order filter and r close to unity and θ small in the second-order filter, we will assume that $|x_n|$ is always smaller than $|y_n|$ so that $(1/A_n) \cong 2|y_n|$ for the first-order filter and $(1/A_n) \cong 4|y_n|$ for the second-order filter. Then, if we consider y_n as a random variable with a symmetric probability density,

$$\bar{k}^2 \cong \frac{4}{3}\,\frac{1}{1+a_1^2}$$

in (20) and

$$\bar{k}^2 \cong \frac{16}{3} G$$

in (23).

Representing a as $1-\delta$ for the first-order case and r as $1-\delta$ for the second-order case, with δ small, we can approximate (18) through (20) as

$$2^{2t}\left(\frac{\sigma_\eta^2}{\sigma_y^2}\right)_{\text{fixed point}} \cong \frac{0.25}{\delta^2} \tag{24}$$

$$2^{2t}\left(\frac{\sigma_\eta^2}{\sigma_y^2}\right)_{\text{floating point}} \cong 0.23\,\frac{1}{\delta} \tag{25}$$

$$2^{2t}\left(\frac{\sigma_\eta^2}{\sigma_y^2}\right)_{\text{block floating}} \cong 0.83\,\frac{1}{\delta}\,\cdot \tag{26}$$

For the second-order case we will want to bracket the expression

$$\frac{1}{\sin\theta}\sum_{n=0}^{\infty} r^n \left| \sin(n+1)\theta \right| .$$

This sum is the sum of the absolute values of the impulse response and as such is an upper bound on the filter output with a maximum input of unity. Consequently, it must be greater than or equal to the response of the second-order filter to a sinusoid of unity amplitude at the resonant frequency. This resonance response is given by $1/(1-r)(1+r^2-2r\cos 2\theta)^{1/2}$ or, with the high-gain approximation, $1/(2\delta\sin\theta)$. An upper bound is easily obtained on the sum as

$$\left(\frac{1}{\sin\theta}\sum_{n=0}^{\infty} r^n\right) \quad \text{or} \quad \frac{1}{\delta\sin\theta} .$$

Furthermore, for the high-gain case we approximate G as $G \cong 1/4\delta\sin^2\theta$. We can then write that

$$\frac{1}{8}\,\frac{1}{\delta^2\sin^2\theta} \leq 2^{2t}\overset{2}{\left(\frac{\sigma_\eta}{\sigma_y{}^2}\right)}_{\text{fixed point}} \leq \frac{1}{2}\,\frac{1}{\delta^2\sin^2\theta} \qquad (27)$$

$$2^{2t}\overset{2}{\left(\frac{\sigma_\eta}{\sigma_y{}^2}\right)}_{\text{floating point}} = 0.23\left[1+\frac{3+4\cos^2\theta}{4\delta\sin^2\theta}\right] \qquad (28)$$

$$\frac{1}{\delta}\left(\frac{1}{4}+\frac{4(1+\cos^2\theta)}{3\sin^2\theta}\right) \leq 2^{2t}\overset{2}{\left(\frac{\sigma_\eta}{\sigma_y{}^2}\right)}_{\text{block floating}} \leq \frac{7+\cos^2\theta}{3\delta\sin^2\theta} . \qquad (29)$$

We note that the behavior of these expressions as a function of δ is consistent with the results plotted in Figs. 4 and 5.

Acknowledgment

The author would like to thank Dr. C. Weinstein for his contributions to this work through many discussions and Mrs. A. Fried for her help in carrying out many of the computations involved.

References

[1] B. Gold and C. M. Rader, "Effect of quantization noise in digital filters," *1966 Spring Joint Computer Conf., AFIPS Proc.*, vol. 28. Washington, D. C.: Thompson, 1966, pp. 213–219.

[2] T. Kaneko and B. Liu, "Round-off error of floating-point digital filters," *Proc. 6th Ann. Allerton Conf. on Circuit and System Theory*, October 2–4, 1968.

[3] C. Weinstein and A. V. Oppenheim, "A comparison of roundoff noise in floating point and fixed point digital filter realizations," *Proc. IEEE* (Letters), vol. 57, pp. 1181–1183, June 1969.

Alan V. Oppenheim (S'57–M'65) was born in New York, N. Y., on November 11, 1937. He received the S.B. and S.M. degrees in 1961 and the Sc.D. degree in 1964 in electrical engineering from the Massachusetts Institute of Technology, Cambridge.

From 1961 to 1964 he was a member of the M.I.T. Research Laboratory of Electronics and an Instructor in the Department of Electrical Engineering, where he received a departmental teaching award in 1963. During this period his primary activities centered on system and communication theory. His doctoral research involved the application of modern algebra to the characterization of nonlinear systems. Since 1964 he has been an Assistant Professor in the M.I.T. Department of Electrical Engineering and a Staff Member of the Research Laboratory of Electronics. His present research activities are in the areas of speech communication and digital waveform processing. He is presently on a leave of absence at the M.I.T. Lincoln Laboratory, Lexington, Mass.

Dr. Oppenheim is a member of Tau Beta Pi, Eta Kappa Nu, Sigma Xi, and the Acoustical Society of America.