Equation (17) is an explicit formula for the elements of the function matrix $P=g(A)$. It does not (directly) require the knowledge of the similarity transformation matrices T and T^{-1} in (3). Thus, the elements of P can be computed directly from the knowledge of the companion matrix A, its eigenvalues, and the function g.

Now, it will be shown that (17) has also a remarkable recursive property (24), which allows one to compute the function matrix P using somewhat less than $6n^2$ multiplications and n^2 divisions. The preceding statement assumes, of course, that all eigenvalues of A are known, and all $g(\lambda_i)$ $(i=1, n)$ are also known. It is also interesting to note that the evaluation of P using the similarity transformation (3) would require about n^3 multiplications and n^3 additions more than is required using (24), which is derived below.

From (8) and (14) it is easy to see that

$$q_{i,j-1} = q_{i,j}\lambda_i + a_{n-j+1} \tag{18}$$

and

$$r_{i+1,j} = r_{i,j}\lambda_j. \tag{19}$$

Therefore,

$$\begin{aligned} p_{i+1,j-1} &= \sum_{l=1}^{n} r_{i+1,l}q_{l,j-1} = \sum_{l=1}^{n} r_{i,l}\lambda_l(q_{l,j}\lambda_l + a_{n-j+1}) \\ &= \sum_{l=1}^{n} r_{i,l}q_{l,j}\lambda_l^2 + \sum_{l=1}^{n} r_{i+1,l}a_{n-j+1}, \end{aligned} \tag{20}$$

but

$$\sum_{l=1}^{n} r_{i,l}q_{l,j}\lambda_l^2 = p_{i+2,j} \tag{21}$$

and

$$\sum_{l=1}^{n} r_{i+1,l}a_{n-j+1} = p_{i+1,n}a_{n-j+1}. \tag{22}$$

Therefore,

$$p_{i+1,j-1} = p_{i+2,j} + p_{i+1,n}a_{n-j+1} \tag{23}$$

or

$$p_{i,j} = p_{i+1,j+1} + p_{i,n}a_{n-j}. \tag{24}$$

Equation (24) is the desired recursive formula.

Example: In order to present an arithmetically simple example, a second-order companion matrix A with real eigenvalues is considered, and a square root function of this matrix is evaluated.

$$A = \begin{bmatrix} 0 & 1 \\ -4 & 5 \end{bmatrix}$$

$$g(A) = A^{\frac{1}{2}} = P.$$

The eigenvalues of A are

$$\lambda_1 = 1 \qquad \lambda_2 = 4.$$

In this case,

$$g(\lambda_1) = 1 \qquad g(\lambda_2) = 2$$

$$f'(\lambda_1) = -3 \qquad f'(\lambda_2) = 3$$

and

$$x_1 = -\tfrac{1}{3} \qquad x_2 = \tfrac{2}{3}.$$

From (17) it follows that

$$p_{1,2} = \tfrac{1}{3} \qquad p_{2,2} = \tfrac{7}{3} \qquad p_{3,2} = \tfrac{31}{3}.$$

These values are now used to obtain the remaining two elements of the matrix P, using the recursive formula (24). Note that the subscript i in (24) must vary faster than the subscript j. Also note that

$$j = n-1, n-2, \cdots, 1$$

$$i = 1, 2, \cdots, n+j-1.$$

Upon completion of these simple computations, the matrix P has the following form,

$$P = \begin{bmatrix} \tfrac{2}{3} & \tfrac{1}{3} \\ -\tfrac{4}{3} & \tfrac{7}{3} \end{bmatrix} = A^{\frac{1}{2}}.$$

And, of course, it can be easily checked that $P^2=A$.

The evaluation of $g=\exp(A)$ can be performed in a similar manner.

Acknowledgment

The author wishes to thank Dr. J. D. Lawson of the University of Waterloo for his valuable suggestions.

I. Kaufman
Dept. of Elec. Engrg. and
Computing Centre
University of Waterloo
Waterloo, Ontario, Canada

A Comparison of Roundoff Noise in Floating Point and Fixed Point Digital Filter Realizations

Abstract—A statistical model for roundoff noise in floating point digital filters, proposed by Kaneko and Liu, is tested experimentally for first- and second-order digital filters. Good agreement between theory and experiment is obtained. The model is used to specify a comparison between floating point and fixed point digital filter realizations on the basis of their output noise-to-signal ratio, and curves representing this comparison are presented. One can find values of the filter parameters at which the fixed and the floating point curves will cross, for equal total register lengths.

Recently, Kaneko and Liu[1] used a statistical model to predict theoretically the effect of roundoff noise in digital filters realized with floating point arithmetic. This letter is concerned with providing an experimental verification of the model, and the use of the model in specifying a quantitative comparison between fixed point and floating point realizations. We restrict attention to first- and second-order filters, both in the interest of simplicity and because more complicated digital filters are often constructed as combinations of first- and second-order filters.

First-Order Case

For a first-order filter of the form

$$w_n = aw_{n-1} + x_n, \tag{1}$$

where x_n is the input and w_n is the output, the computed output y_n is

$$y_n = [ay_{n-1}(1+\varepsilon_n) + x_n](1+\xi_n). \tag{2}$$

The random variables ε_n and ξ_n account for the roundoff errors due to the floating point multiply and add, respectively, and are bounded by

$$|\varepsilon_n| \le 2^{-t}, \qquad |\xi_n| \le 2^{-t}. \tag{3}$$

Following Kaneko and Liu, we define the error $e_n=y_n-w_n$, subtract (1) from (2), neglect second-order terms in e, ε, and ξ, and obtain a difference equation for the error e_n, as

$$e_n - ae_{n-1} = aw_{n-1}(\varepsilon_n + \xi_n) + x_n\xi_n = u_n. \tag{4}$$

Manuscript received February 10, 1969. This work was sponsored by the U.S. Air Force.

[1] T. Kaneko and B. Liu, "Round-off error of floating-point digital filters," presented at the Sixth Annual Allerton Conf. on Circuit and System Theory, Allerton, Ill., October 1968. To be published in the *Proc. of the Conference.*

Assuming that ε_n and ξ_n are independent from sample to sample (white), and that ε_n, ξ_n, and the signal x_n are mutually independent, u_n in (4) is white noise with variance dictated by the statistics of x_n and the variances σ_ε^2 and σ_ξ^2 of ε_n and ξ_n. The variance σ_e^2 of the output noise e_n is obtained easily from the variance σ_u^2 of u_n as

$$\sigma_e^2 = \sigma_u^2 \sum_{n=0}^{\infty} h_n^2 = \frac{1}{1-a^2}\sigma_u^2 \tag{5}$$

where $h_n = a^n$ is the filter impulse response.

For example, if we assume that x_n is stationary white noise of variance σ_x^2, we obtain

$$\sigma_e^2 = \frac{\sigma_\xi^2 + a^2\sigma_\varepsilon^2}{(1-a^2)^2}\sigma_x^2. \tag{6}$$

For the case of a high gain filter, with $a = 1-\delta$, and δ small, (6) becomes

$$\sigma_e^2 = \frac{(\sigma_\varepsilon^2 + \sigma_\xi^2)\sigma_x^2}{4\delta^2}. \tag{6a}$$

If, instead, x_n is taken to be a sine wave of the form $A \sin(\omega_0 n + \phi)$ with ϕ uniformly distributed in $(0, 2\pi)$, then

$$\sigma_e^2 = \frac{A^2(\sigma_\xi^2 + a^2\sigma_\varepsilon^2)}{2(1-a^2)(a^2 - 2a\cos\omega_0 + 1)}. \tag{7}$$

To test the model, σ_e^2 was measured experimentally for white noise and sine wave inputs. Each input was applied to a filter using a 27-bit mantissa, and also a filter with the same coefficient a, but using a shorter (e.g., 12-bit) mantissa in the computation. The outputs of the two filters were then subtracted, squared, and averaged over a sufficiently long period to obtain a stable estimate of σ_e^2. Kaneko and Liu assumed that ξ_n and ε_n were both uniformly distributed in $(-2^{-t}, 2^{-t})$ with variances $\sigma_\varepsilon^2 = \sigma_\xi^2 = \frac{1}{3}2^{-2t}$. Actual measurements of the noise due to a multiply and an add verified that ε_n and ξ_n have zero mean, but indicated that the variances

$$\sigma_\varepsilon^2 = \sigma_\xi^2 = (0.23)(2^{-2t}) \tag{8}$$

would better represent these noise sources. Using (1), (6), (7), and (8), we can compute the output noise-to-signal ratio for both white noise and sinusoidal inputs for the first-order case as

$$\frac{\sigma_e^2}{\sigma_w^2} = (0.23)2^{-2t}\left(\frac{1+a^2}{1-a^2}\right) \tag{9}$$

In Fig. 1, experimental curves for noise-to-signal ratio are compared with the theoretical curve of (9).

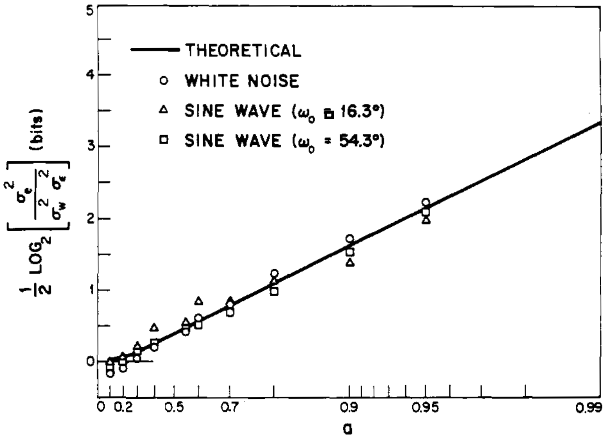

Fig. 1. Theoretical and experimental noise-to-signal ratio for a first-order filter, as a function of pole position. The noise-to-signal ratio is represented in bits.

Second-Order Case

An analysis similar to the above can be carried out for a second-order filter of the form

$$w_n = -r^2 w_{n-2} + 2r\cos\theta w_{n-1} + x_n, \tag{10}$$

with a complex conjugate pole pair at $z = re^{\pm j\theta}$. Based on experimental verification of (8) in the first-order case, we assume here that the ε's and ξ's representing the errors in the second-order case have the same variance, as given by (8).

When x_n is stationary white noise, we obtain for the variance of the noise e_n,

$$\sigma_e^2 = \sigma_\varepsilon^2\sigma_x^2\left[G + G^2\left(3r^4 + 12r^2\cos^2\theta - 16\frac{r^4\cos^2\theta}{1+r^2}\right)\right] \tag{11}$$

where

$$G = \frac{\sigma_e^2}{\sigma_u^2} = \sum_{n=0}^{\infty} h_n^2 = \left(\frac{1+r^2}{1-r^2}\right)\left(\frac{1}{r^4 + 1 - 4r^2\cos^2\theta + 2r^2}\right). \tag{12}$$

TABLE I

Theoretical and Experimental Noise-to-Signal Ratio for a Second-Order Filter, as a Function of Pole Position

r	θ	$\frac{1}{2}\log_2\left[\frac{\sigma_e^2}{\sigma_w^2\sigma_\varepsilon^2}\right]$ (bits)			
		White Noise		Sine Wave	
		Theoretical	Experimental	Theoretical	Experimental
0.55	22.5	1.48	1.66	1.54	1.64
0.7	22.5	2.16	2.33	2.23	2.38
0.9	22.5	3.32	3.33	3.35	3.45
0.55	45.0	0.93	1.08	0.97	0.94
0.7	45.0	1.36	1.44	1.37	1.51
0.9	45.0	2.28	2.51	2.22	2.14
0.55	67.5	0.42	0.46	0.39	0.33
0.7	67.5	0.75	0.88	0.65	0.62
0.9	67.5	1.63	1.97	1.45	0.99

For the case of a high gain filter, with $r = 1-\delta$, (11) becomes approximately

$$\sigma_e^2 = \sigma_\varepsilon^2\sigma_x^2\left(\frac{3+4\cos^2\theta}{16\delta^2\sin^2\theta}\right). \tag{13}$$

For the case of sinusoidal input, we obtain

$$\sigma_e^2 = A^2G\sigma_\varepsilon^2\left[\tfrac{3}{2}r^4|H|^2 + 6r^2\cos^2\theta|H|^2 + \tfrac{1}{2} - 4r^3|H|^2\cos\theta\cos\omega_0 - r^2|H|\cos(\phi - 2\omega_0) + 2r|H|\cos\theta\cos(\phi - \omega_0)\right] \tag{14}$$

where $|H|$ and ϕ represent the magnitude and phase of the filter system function at the input frequency ω_0. In Table I, a comparison of theoretical and experimental values for output noise-to-signal ratio are displayed for a second-order filter.

Fixed Versus Floating Point Comparison

The statistical model of floating point roundoff noise proposed by Kaneko and Liu and one of fixed point roundoff noise as presented for example by Gold and Rader[2] provide the framework for comparing these two structures on the basis of the resulting noise-to-signal ratio. We consider only the case of white noise input.

For the fixed point case, the register length must be chosen sufficiently long so that the output cannot overflow the fixed point word. If h_n denotes the impulse response of the filter, then output w_n is bounded according to

[2] 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. Spartan, 1966, pp. 213–219.

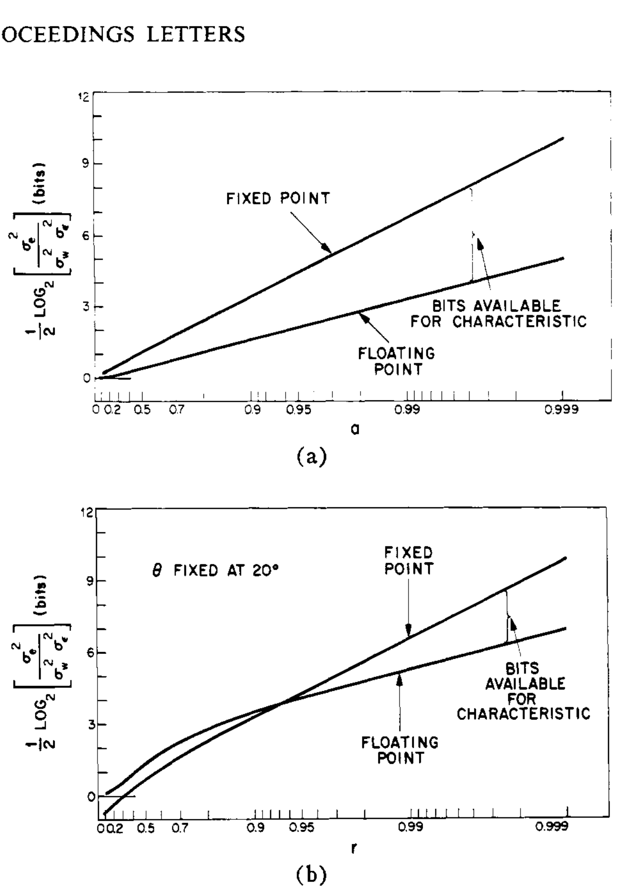

Fig. 2. Comparison of fixed point and floating point noise-to-signal ratios. (a) First-order filter. (b) Second-order filter, $\theta = 20°$.

$$\max(|w_n|) = \max(|x_n|) \sum_{n=0}^{\infty} |h_n|. \tag{15}$$

Interpreting the fixed point numbers as signed fractions, we require for no overflows that $|w_n| < 1$, restricting x_n to the range

$$-\frac{1}{\sum_{n=0}^{\infty} |h_n|} < x_n < +\frac{1}{\sum_{n=0}^{\infty} |h_n|}. \tag{16}$$

With x_n white and uniformly distributed between the limits in (16), the resulting output noise-to-signal ratio for a first-order filter is

$$\frac{\sigma_e^2}{\sigma_w^2} = \frac{1}{4} 2^{-2t} \left(\sum_{n=0}^{\infty} |h_n| \right)^2 = \frac{1}{4} \frac{2^{-2t}}{(1-a)^2}, \tag{17}$$

and for a second-order filter

$$\frac{\sigma_e^2}{\sigma_w^2} = \frac{1}{2} 2^{-2t} \left(\sum_{n=0}^{\infty} |h_n| \right)^2 = \frac{1}{2} 2^{-2t} \left(\frac{1}{\sin\theta} \sum_{n=0}^{\infty} r^n |\sin[(n+1)\theta]| \right)^2. \tag{18}$$

The variance of the roundoff noise due to a multiplication is taken as $\frac{1}{12} 2^{-2t}$ with t denoting the fixed point register length.

For the case of floating point computation, the noise-to-signal ratio for the first-order filter is

$$\frac{\sigma_e^2}{\sigma_w^2} = (0.23) 2^{-2t} \frac{1+a^2}{1-a^2} \tag{19}$$

where t is the number of bits in the mantissa. For the second-order filter, we have

$$\frac{\sigma_e^2}{\sigma_w^2} = (0.23) 2^{-2t} \left[1 + G\left(3r^4 + 12r^2 \cos^2\theta - 16 \frac{r^4 \cos^2\theta}{1+r^2} \right) \right]. \tag{20}$$

For a comparison of floating and fixed point arithmetic in the case of a first-order filter, Fig. 2(a) presents curves of $\frac{1}{2} \log_2 (\sigma_e^2 / \sigma_\varepsilon^2 \sigma_w^2)$ as determined from (8), (17), and (19). These curves represent a comparison of the rms noise-to-signal ratio for the two cases, in units of bits. In Fig. 2(b), a similar comparison is illustrated for the second-order case. For the purpose of the illustration, θ was kept fixed and only r varied.

Fig. 2(a) and (b) indicates that floating point arithmetic leads to a lower noise-to-signal ratio than fixed point if the floating point mantissa is equal in length to the fixed point word. We notice that for high gain filters, as a increases toward unity in the first-order case, and as r increases toward unity for θ fixed in the second-order case, the noise-to-signal ratio for fixed point increases faster than for floating point.

However, this comparison does not account for the number of bits needed for the characteristic in floating point. If c denotes the number of bits in the characteristic, this would be accounted for in Fig. 2 by numerically adding the constant c to the floating point data. This shift will cause the floating and fixed point curves to cross at a point where the noise-to-signal ratios are equal for equal total register lengths.

For the sake of the comparison, we provide just enough bits in the characteristic to allow the same dynamic range for both the floating and the fixed point filters. If t_{fx} denotes the fixed point word length, then the requirement of identical dynamic range requires that

$$c = \log_2 t_{fx}. \tag{21}$$

Assuming for example that $t_{fx} = 16$ so that $c = 4$, crossover points in the noise-to-signal ratio will occur at $a = 0.996$ in the first-order case, and at $r = 0.99975$, $\theta = 20°$, in the second-order case depicted by Fig. 2(b).

Clifford Weinstein
Alan V. Oppenheim
M.I.T. Lincoln Lab.
Lexington, Mass. 02173

A Physical Proof of Tellegen's Theorem

Abstract—A physically oriented proof is given for the generalized Tellegen's theorem. The proof consists of applying the law of energy conservation to a hypothetical network in which all branch voltages and currents are determined by independent sources.

Of all the recently discovered (or rediscovered) network theorems, few have been as stimulating or useful as Tellegen's theorem.[1] It has been used in sensitivity studies, energy and power relations, etc., for both linear and nonlinear networks, active as well as passive, time-variable as well as time-invariant.[2] It is even applicable in other branches of physics.[2] More recently, it was also utilized to justify and define the adjoint network concept[3] which promises to be of great importance in computer-aided circuit analysis and optimization.

In its most general formulation,[2] Tellegen's theorem relates the electrical quantities of two networks, $\boldsymbol{N_a}$ and $\boldsymbol{N_b}$. These networks must be of identical topology, but may be arbitrarily different in the way their branches are constructed from impedances and generators. Then the relation

$$[\Lambda_a(\bar{i}_a)]^T [\Lambda_b(\bar{v}_b)] = 0 \tag{1}$$

holds. In (1), $\bar{i}_a$ is the branch current vector of $\boldsymbol{N_a}$, $\bar{v}_b$ is the branch voltage vector of $\boldsymbol{N_b}$, and Λ_a (Λ_b) is a scalar operator such that when applied to a current (voltage) vector satisfying the Kirchhoff laws, the resulting vector still obeys these same laws.

Tellegen's theorem is usually proved from the Kirchhoff laws, using the incidence or loop matrices of the networks.[4] An alternative proof, perhaps

Manuscript received February 19, 1969.

[1] B. D. H. Tellegen, "A general network theorem, with application," *Philips Res. Repts.*, vol. 7, pp. 259–269, August 1952.

[2] P. Penfield, R. Spence, and S. Duinker, "Tellegen's theorem," M.I.T. Dept. of Elec. Engrg., Cambridge, Mass., Internal Memo 153, December 2, 1968.

[3] S. W. Director and R. A. Rohrer, "Automated network design: the frequency domain case," to be published in *IEEE Trans. Circuit Theory*, vol. CT-16, August 1969.

[4] See, for example, Penfield, Spence, and Duinker.[2]