which yield

$$\hat{K}' = \begin{bmatrix} 1.000 & 0.1189 & 0.7483 \\ 0.1189 & 1.000 & 0.1189 \\ 0.7483 & 0.1189 & 1.000 \end{bmatrix}$$

$$\hat{W}' = \begin{bmatrix} 0.2175 & 0.02586 & 0.1628 \\ 0.02586 & 0.2175 & 0.02586 \\ 0.1628 & 0.02586 & 0.2175 \end{bmatrix}.$$

The sensitivity measure and noise power gain for this optimal realization are $\hat{M}' = 5.610$ and $\hat{G}' = 0.6526$. $\hat{M}'$ is the minimum sensitivity measure and $\hat{G}'$ is the minimum noise power attainable for this transfer function. This is verified since for this realization $\hat{K}' = \rho^2 \hat{W}'$ with $\rho^2 = 4.597$ or by the aid of (10) and (11) using the eigenvalues of KW

$$\mu_1 = 0.01378 \quad \mu_2 = 0.2021 \quad \mu_3 = 0.6926.$$

VI. CONCLUSIONS

For scaled digital filters the relation between the sensitivity measure and the noise power gain is given which shows that the necessary and sufficient condition to minimize the noise power gain is to minimize the sensitivity measure of (11). It is shown that the minimization of this sensitivity measure can be carried out without taking into account the dynamic range constraint. A procedure is given that simultaneously minimizes the sensitivity measure and the noise power gain in scaled digital filters.

ACKNOWLEDGMENT

The authors wish to thank Prof. R. Saal for the opportunity to carry out this research at the Technical University of Munich and for his continuous encouragement. The first author also wishes to thank the Alexander von Humboldt Foundation for rendering a research fellowship.

REFERENCES

- [1] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 551-562, Sept. 1976.
- [2] L. B. Jackson, A. G. Lindgren, and Y. Kim, "Optimal synthesis of second-order state-space structures for digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 149-153, Mar. 1979.
- [3] S. Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. Acoustics Speech, Signal Processing*, vol. ASSP-25, pp. 273-281, Aug. 1977.
- [4] C. W. Barnes, "Roundoff noise and overflow in normal digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 154-159, Mar. 1979.
- [5] S. Y. Hwang, "Roundoff noise in state-space digital filtering: A general analysis," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-24, pp. 256-262, June 1976.
- [6] V. Tavsanoğlu, "Sensitivity analysis and optimum realizations using state-space techniques," Habilitation Thesis, Faculty of Electrical Engineering, Technical Univ. Istanbul, Istanbul, 1979.
- [7] V. Tavsanoğlu, "Minimization of pole sensitivities by similarity transformations," in *Proc. 4th Int. Symp. Network Theory*, Ljubljana, pp. 163-168, Sept. 1979.
- [8] F. R. Gantmacher, *The Theory of Matrices*. New York: Chelsea, 1959.
- [9] F. Neiss, *Determinanten und Matrizen*. Berlin, Germany: Springer-Verlag, 1967.

Combined Error at the Output of Two-Dimensional Recursive Digital Filters

B. G. MERTZIOS AND A. N. VENETSANOPOULOS

Abstract—The noise properties of two-dimensional (2-D) recursive digital filters, caused by finite register lengths, are considered. A combined error model is proposed, which takes into account three sources of error; roundoff accumulation, coefficient quantization, and input quantization. A

statistical approach is used, in the space domain, which allows the computation of the statistics of the errors at the filter output.

I. INTRODUCTION

The noise properties of two-dimensional (2-D) digital filters have been considered in the literature [1]–[5], but have not resulted in error models that can be easily utilized by the designer. The purpose of this paper is to present a combined error model which integrates, the three kinds of errors, roundoff accumulation, coefficient quantization, and input-quantization. While the combined error is in general a nonlinear combination of these three errors, neglecting the second-order terms appearing in this model, we arrive in a linear combination. These three individual errors are then considered separately. Finally the error statistics, for a fixed-point arithmetic direct realization are given in a table as functions of known filter parameters and the register length. Thus the designer can predict the filter quality and choose a register length adequate for the performance required in a given application.

The effect of coefficient quantization can be studied in terms of sensitivity analysis [6], but whenever the variation of the coefficients is unknown and can be described in a statistical manner, a statistical analysis is justified [1], [7, pp. 336–338].

In this paper fixed-point arithmetic is used and negative numbers are represented in any of the three usual forms. These forms, namely, the sign and magnitude, the one's complement and the two's complement representation, result in different error effects at the filter output. Distinction between rounding and truncation is made. The direct form of realization is presented here, but the analysis can be easily applied to other configurations (cascade, parallel, etc.).

To simplify the analysis some assumptions are made. These assumptions and their limitations are discussed below:

- 1) The sequence of error samples is a sample sequence of a stationary random process.
- 2) The quantization process is white. The random variables representing the error sources are uncorrelated and independent of the sampling rate.¹
- 3) The error sequence is uncorrelated with the sequence of exact samples. This is intuitively true if the amplitude of the exact samples varies in a random fashion, but is generally not considered to be a good assumption for one's complement and two's complement representations.
- 4) The quantization error has a uniform density function. This implies that the signal is equally likely to be anywhere within a quantization interval.
- 5) The statistical analysis of errors considered here is concerned with the steady-state case.

II. COMBINED QUANTIZATION ERROR DUE TO ROUND OFF ACCUMULATION, COEFFICIENT, AND INPUT ERRORS

Consider an infinite register length filter. Let x_{nm} , y_{nm} denote the filter input and output, respectively, and a_{ij} , b_{ij} the coefficients, with infinite precision arithmetic. Then the filtering process corresponding to a 2-D, first quadrant recursive filter is represented by the recursive difference equation

$$y_{nm} = \sum_{i=0}^{N_A} \sum_{j=0}^{M_A} a_{ij} x_{n-i, m-j} - \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} b_{ij} y_{n-i, m-j} \quad (1)$$

Manuscript received December 1, 1982; August 18, 1983.

The authors are with the Department of Electrical Engineering, University of Toronto, Toronto, Ont., Canada.

¹Thus the signals are considered fluctuating rapidly in a somehow unpredictable manner [8].

which corresponds to the 2-D first quadrant IIR filter

$$H(z_1, z_2) = \frac{A(z_1, z_2)}{D(z_1, z_2)} = \frac{\sum_{i=0}^{N_A} \sum_{j=0}^{M_A} a_{ij} z_1^i z_2^j}{\sum_{i=0}^{N_B} \sum_{j=0}^{M_B} b_{ij} z_1^i z_2^j}. \quad (2)$$

With finite length registers the implementation of the filtering process introduces errors due to

1) *Roundoff accumulation error*: Let q denote the quantity after multiplication and rounding or truncation.

2) *Coefficient quantization*: Let the result of quantizing the coefficients a_{ij} and b_{ij} be $\bar{a}_{ij} = a_{ij} - \mu_{ij}$, $\bar{b}_{ij} = b_{ij} - \nu_{ij}$.

3) *Input quantization error*: Quantization of the input occurs at the analog to digital converter. Further quantization may occur if the filter register length is shorter than that of the quantizer. Let the result of quantizing the input be $\bar{x}_{nm} = x_{nm} - e_{nm}$. Taking into account the three above-mentioned errors, the finite precision equation becomes

$$\bar{y}_{nm} = \sum_{i=0}^{N_A} \sum_{j=0}^{M_A} [\bar{a}_{ij} \bar{x}_{n-i, m-j}]_q - \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} [\bar{b}_{ij} \bar{y}_{n-i, m-j}]_q. \quad (3)$$

Let $f_{nm} \triangleq y_{nm} - \bar{y}_{nm}$ be the error at the output and define

$$\begin{aligned} \theta_{nm} &= \sum_{i=0}^{N_A} \sum_{j=0}^{M_A} \theta'_{ij, nm} \\ &= \sum_{i=0}^{N_A} \sum_{j=0}^{M_A} [\bar{a}_{ij} \bar{x}_{n-i, m-j} - (\bar{a}_{ij} \bar{x}_{n-i, m-j})_q] \\ \xi_{nm} &= \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} \xi'_{ij, nm} \\ &= \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} [\bar{b}_{ij} \bar{y}_{n-i, m-j} - (\bar{b}_{ij} \bar{y}_{n-i, m-j})_q]. \end{aligned} \quad (4)$$

Subtracting (3) from (1) and using (4), (5) we obtain the following recursive relationship

$$\begin{aligned} f_{nm} &= \underbrace{\theta_{nm} - \xi_{nm}}_I + \underbrace{\sum_{i=0}^{N_A} \sum_{j=0}^{M_A} \bar{a}_{ij} e_{n-i, m-j}}_{II} \\ &+ \underbrace{\sum_{i=0}^{N_A} \sum_{j=0}^{M_A} \mu_{ij} x_{n-i, m-j} - \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} \nu_{ij} \bar{y}_{n-i, m-j}}_{III} \\ &- \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} b_{ij} f_{n-i, m-j} \end{aligned} \quad (6)$$

where the terms I, II, III describe the effect of roundoff accumulation, the input and the coefficient errors, respectively. The above model of the general error is nonlinear, since the input and the coefficient errors contain the following second-order terms, respectively,

$$\mu_{ij} e_{n-i, m-j}, \quad i=0, \dots, N_A; j=0, \dots, M_A$$

$$\nu_{ij} f_{n-i, m-j}, \quad i=0, \dots, N_B; j=0, \dots, M_B.$$

If we neglect the second-order terms, (6) yields in the linear equation

$$\begin{aligned} f_{nm} &= \theta_{nm} - \xi_{nm} + \sum_{i=0}^{N_A} \sum_{j=0}^{M_A} a_{ij} e_{n-i, m-j} + \sum_{i=0}^{N_A} \sum_{j=0}^{M_A} \mu_{ij} x_{n-i, m-j} \\ &- \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} \nu_{ij} y_{n-i, m-j} - \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} b_{ij} f_{n-i, m-j}. \end{aligned} \quad (7)$$

The late model of the general error results in a linear combination of the three individual errors.

Since in this case the general error becomes a linear combination of the three individual errors, we consider each error separately. If the errors are considered as uncorrelated sequences, the variance of the total is equal to the sum of the variances of the individual errors and it corresponds to the power of the total error at the output of the filter.

III. ROUND OFF ACCUMULATION ERROR

The output $\bar{y}_{nm}$ from the finite length register filter is given by

$$\bar{y}_{nm} = \sum_{i=0}^{N_A} \sum_{j=0}^{M_A} (a_{ij} x_{n-i, m-j})_q - \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} (b_{ij} \bar{y}_{n-i, m-j})_q. \quad (8)$$

Using (1) and (3) we obtain the error at the filter output

$$f_{nm} \triangleq y_{nm} - \bar{y}_{nm} = \lambda_{nm} - \epsilon_{nm} - \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} b_{ij} f_{n-i, m-j} \quad (9)$$

where

$$\begin{aligned} \lambda_{nm} &\triangleq \sum_{i=0}^{N_A} \sum_{j=0}^{M_A} \lambda'_{ij, nm} \\ &= \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} [a_{ij} x_{n-i, m-j} - (a_{ij} x_{n-i, m-j})_q] \end{aligned} \quad (10)$$

$$\begin{aligned} \epsilon_{nm} &\triangleq \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} \epsilon'_{ij, nm} \\ &= \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} [b_{ij} \bar{y}_{n-i, m-j} - (b_{ij} \bar{y}_{n-i, m-j})_q]. \end{aligned} \quad (11)$$

IV. COEFFICIENT QUANTIZATION ERROR

Here, the input x_{nm} is assumed to be a zero-mean and wide-sense stationary (W.S.S.) random process. The output is, therefore, also zero-mean and W.S.S. The recursive equation (1) implemented by finite precision coefficients then becomes

$$\bar{y}_{nm} = \sum_{i=0}^{N_A} \sum_{j=0}^{M_A} \bar{a}_{ij} x_{n-i, m-j} - \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} \bar{b}_{ij} \bar{y}_{n-i, m-j}. \quad (12)$$

Neglecting the second-order terms, the error at the filter output is found to be

$$f_{nm} \triangleq y_{nm} - \bar{y}_{nm} = u_{nm} - \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} b_{ij} f_{n-i, m-j} \quad (13)$$

Fig. 1. Block diagram for the direct form implementation and for the three kinds of errors. (a) Roundoff accumulation error. (b) Coefficient quantization error. (c) Input quantization error.

TABLE I
STATISTICS OF THE QUANTIZATION ERROR

REPRESENTATION		ROUNDING	TRUNCATION
Sign-and-Magnitude	$\bar{e}$	0	0
	σ_e^2	$q^2/12$	$q^2/3$
One's Complement	$\bar{e}$	0	0
	σ_e^2	$q^2/12$	$q^2/3$
Two's Complement	$\bar{e}$	0	$q/2$
	σ_e^2	$q^2/12$	$q^2/12$

Here q denotes the quantization interval and is equal to 2^{-b} where b is the register length.

where

$$u_{nm} \triangleq \sum_{i=0}^{N_A} \sum_{j=0}^{M_A} \mu_{ij} x_{n-i, m-j} - \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} \nu_{ij} y_{n-i, m-j} \quad (14)$$

V. INPUT QUANTIZATION ERROR

The output of the filter when the input is quantized is given by

$$\bar{y}_{nm} = \sum_{i=0}^{N_A} \sum_{j=0}^{M_A} a_{ij} \bar{x}_{n-i, m-j} - \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} b_{ij} \bar{y}_{n-i, m-j} \quad (15)$$

Neglecting the second-order terms, the error at the filter output is then given by

$$f_{nm} \triangleq y_{nm} - \bar{y}_{nm} = \sum_{i=0}^{N_A} \sum_{j=0}^{M_A} a_{ij} e_{n-i, m-j} - \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} b_{ij} f_{n-i, m-j} \quad (16)$$

The statistics of the quantization error $\bar{e}$, σ_e^2 for rounding and truncation are given in Table I. The statistics of the three kinds of errors described individually in Sections III, IV, V, for the direct

TABLE II
STATISTICS OF THE ERROR FOR THE DIRECT FORM

Roundoff Accumulation Error	$\bar{f}$	$\bar{e}(\alpha - \beta)Q_1$
	σ_f^2	$\sigma_e^2(\alpha + \beta)Q_2$
Coefficient Quantization Error	$\bar{f}$	0
	σ_f^2	$\sigma_e^2(\alpha\sigma_x^2 + \beta\sigma_y^2)Q_2$
Input Quantization Error	$\bar{f}$	$\bar{e}P_1/Q_1$
	$-\sigma_f^2$	$\sigma_e^2P_2$

$$P_1 = \sum_{i=0}^{N_A} \sum_{j=0}^{M_A} a_{ij}^2 \quad Q_1 = \sum_{i=0}^{N_B} \sum_{j=0}^{M_B} b_{ij}^2$$

$$P_2 = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} h_{ij}^2 \quad Q_2 = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} d_{ij}^2$$

and h_{ij} , d_{ij} are the inverse Z transforms of $H(z_1, z_2)$ and $1/D(z_1, z_2)$, respectively.

form realization are given in Table II and are interpreted in Fig. 1. Note that α and β in Table II denote the number of the coefficients of the numerator and the denominator of the transfer function respectively, that require multiplications. Furthermore, σ_x^2 , σ_y^2 are the variances of the input and output signals respectively. Tables similar to 2 can be obtained for other filter configurations, such as canonic, cascade, array etc.

VI. CONCLUSIONS

The effects of finite precision in the implementation of two-dimensional recursive digital filters were analyzed. An expression for the total error at the filter output was derived and it was found that it is in general a nonlinear combination of the three individual errors. For sufficiently small individual errors it was found that it becomes a linear combination of the three errors. The statistical properties of the quantization errors, at the output of the filters were calculated. These statistics include the mean and the variance, from which the mean square error, the signal-to-noise ratio as well as error bounds can be computed.

The statistics obtained can be easily calculated by the designer in terms of the known coefficients and the register length. For the

calculation of the infinite sums, appearing in the calculation of the variances for the steady-state case, known results can be used [2], [3]. Therefore, it is possible to plot diagrams of the statistics against the register length used and choose on the basis of such diagrams the required register length for a predetermined output error-variance.

REFERENCES

- [1] M. D. Ni and J. K. Aggarwal, "Two-dimensional digital filtering and its error analysis," *IEEE Trans. Computers*, vol. C-23, September 1974.
- [2] P. Agathoklis, E. I. Jury, and M. Mansour, "Evaluation of quantization error in two-dimensional digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 273-279, June 1980.
- [3] L. M. Roytman and M. N. S. Swamy, "Determination of quantization error in two-dimensional digital filters," *Proc. IEEE*, vol. 69, pp. 832-834, 1981.
- [4] C. T. Mullis and R. A. Roberts, "Roundoff noise in digital filters: Frequency transformations and invariants," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 538-549, Dec. 1976.
- [5] M. D. Ni and J. K. Aggarwal, "Error analysis of two-dimensional recursive digital filters employing floating-point arithmetic," *IEEE Trans. Computers*, vol. C-25, pp. 755-759, 1976.
- [6] P. N. Paraskevopoulos and B. G. Mertzios, "Sensitivity analysis of 2-D systems," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 833-838, Aug. 1981.
- [7] L. B. Rabiner and B. Gold, *Theory and Applications of Digital Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.
- [8] A. N. Oppenheim and R. W. Schaffer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

Design of Linear CMOS Transconductance Elements

A. NEDUNGADI AND T. R. VISWANATHAN

Abstract—Some novel circuit techniques for realizing linear CMOS transconductance elements are proposed. The circuits discussed have superior linearity and input voltage range compared with the conventional source-coupled differential pair. Design tradeoffs are examined and computer simulation results are used to verify theoretical predictions. The results show close agreement between predicted behavior and simulated performance.

I. INTRODUCTION

Linear transconductance elements [1]–[7] are useful building blocks in analog signal-processing systems. The objective of this correspondence is to present some novel CMOS circuit realizations of linear transconductance blocks. The linearity and input voltage range of the proposed designs are superior to that of the source-coupled differential pair. Design tradeoffs are examined and computer simulation results are presented.

All FET's are assumed to be enhancement-mode types biased in saturation with substrates connected to their respective sources. Thus body effect is ignored and the transistor behavior is approximated by the relation [6]

$$I_D = k(V_{GS} - V_{th})^2 \quad (1)$$

where V_{th} is the threshold voltage, $k = k_0 W/L$ is the transconductance parameter, W and L are width and length of the channel, and $k_0 = \mu C_{ox}/2$, where μ is the effective surface carrier

mobility and C_{ox} is the gate oxide capacitance per unit area. The effect of channel length modulation is not included in (1) but is separately taken into account when considering output conductance in the proposed circuits.

II. THE BASIC SOURCE COUPLED PAIR

Consider a source-coupled n -channel differential pair biased by a current sink $2I$. For a differential input voltage v , let the drain currents of the transistors be represented by $I + i_1$, and $I - i_1$, where i_1 is the incremental current due to v . If the drain currents are differenced using a p-channel current mirror of unity-gain to form a singled ended current output i , assuming matched devices, the large-signal $i-v$ transfer characteristic will be given by

$$i = 2i_1 = 2\sqrt{kI} v \sqrt{1 - (kv^2/4I)}, \quad |v| \leq \sqrt{2I/k} \quad (2a)$$

$$i = 2I \operatorname{sgn}(v), \quad |v| > \sqrt{2I/k}. \quad (2b)$$

Defining $V_b = \sqrt{2I/k}$, (2) can be rewritten using normalized variables $x = v/V_b$ and $y = i/I$, as

$$y = 2x\sqrt{1 - x^2/4}, \quad |x| \leq \sqrt{2} \quad (3a)$$

$$y = 2 \operatorname{sgn}(x), \quad |x| > \sqrt{2} \quad (3b)$$

For $|x| \ll 1$ (i.e., $|v| \ll V_b$), we get a linear characteristic with transconductance $g_m = i/v = 2\sqrt{kI}$. For larger x the output becomes increasingly nonlinear. In order to obtain better than 1 percent linearity, $|v| < 0.28V_b$. To increase the linear input voltage range, the value of $V_b = \sqrt{2I/k}$ must be made large by choosing a large bias current I and/or by making the input devices long [8]. This would necessitate the use of a larger negative supply voltage to maintain the current sink device in saturation, especially during negative common-mode input swings. That is, for a given negative supply voltage the common-mode input range voltage would be severely restricted. Another important consequence is that the input offset voltage, due to k mismatches in the differential pair, increases directly in proportion to V_b [9]. The circuits proposed provide substantial improvement in linearity without requiring excessive bias voltage. The attempt is to linearize the $i-v$ characteristics over the entire range defined by $|v| \leq V_b$, or equivalently $|x| \leq 1$.

III. THE CROSS-COUPLED QUAD CELL

Consider a matched pair (M_1, M_2) coupled by two voltage sources of equal value ($V_{th} + V_x$) as shown in Fig. 1(a). Representing $V_1 - V_2$ by v , using (1) gives

$$I_1 = k(V_x + v)^2 \quad (5a)$$

$$I_2 = k(V_x - v)^2 \quad (5b)$$

The output current $i = I_1 - I_2$, obtained by a simple p-channel current mirror, will be

$$i = I_1 - I_2 = 4kV_x v. \quad (6)$$

Thus this configuration exhibits a perfectly linear transconductance of value $g_m = 4kV_x$.

Consider implementing the sources ($V_{th} + V_x$) by two additional n -channel transistors M_3 and M_4 biased by constant currents. Note that M_1 and M_4 share a p-well connected to their common source node, whereas M_2 and M_3 are in a separate well to avoid back-gate bias effects. For obtaining reasonable voltage sources with low source impedance, these devices must be wide

Manuscript received December 5, 1983.

A. Nedungadi is with the Department of Electrical Engineering, Texas A&M University, College Station, TX 77843.

T. R. Viswanathan is with the Department of Electrical Engineering, University of Waterloo, Waterloo, Ont., Canada N2L 3G1.