Roundoff Error Analysis of Digital Filters*

UDO ZÖLZER, *AES Member*

Department of Telecommunications, Technical University of Hamburg, D-21071 Hamburg, Germany

Digital filtering of audio signals in professional mixing consoles requires low-roundoff-noise digital filter structures which are suitable for realizing very low pole frequencies. Without appropriate choice of the digital filter structure the effects of limited word length lead to high necessary internal word length of the signal processor. On the basis of a roundoff error analysis a new recursive digital filter structure is compared with already known filter structures, especially for very low cutoff frequencies. With the help of additional zeros in the noise transfer functions (error spectrum shaping) a further reduction of roundoff noise and the suppression of limit cycles can be achieved.

0 INTRODUCTION

For recursive digital filter structures the limited word length of a signal processor implementation leads to two different quantization errors. The quantization of the coefficients influences the linear distortion of the frequency response. The quantization of the signal is responsible for the maximum dynamic range and determines the roundoff noise behavior of the digital filter. Another effect of the signal quantization is the limit cycles. They can be separated into overflow limit cycles, small-scale limit cycles, and limit cycles correlated with the input signal. Limit cycles are very disturbing because of their small-band character. The roundoff noise and the coefficient sensitivity of a digital filter structure depend on the topology of the specific filter structure and on the pole frequency (cutoff frequency). This paper considers different filter topologies and their specific pole distributions. A statistical analysis of quantization in fixed-point realizations of recursive digital filters is demonstrated, and the application of error spectrum shaping is analyzed.

1 POLE DISTRIBUTIONS

The basis for the following considerations concerning recursive digital filter structures is the connection between coefficient sensitivity and roundoff noise. This was first stated by Fettweis [1] and can be used deriving filter structures with low roundoff noise. By increasing the pole density in a special region of the z plane the coefficient sensitivity and the roundoff noise in this region can be reduced. With this improvement the coefficient word length as well as the signal word length can be reduced to the requirements of the application.

Typical audio filters, such as high-pass, low-pass, peak, and shelving filters, can be approximated with the second-order transfer function

$$H(z) = \frac{a_0 + a_1 z^{-1} + a_2 z^{-2}}{1 + b_1 z^{-1} + a_2 z^{-2}} . \tag{1}$$

The recursive part of the difference equation which can be derived from Eq. (1) is considered first because of its great influence on the error behavior. For a quantization step size of 6 bit the quantization of the denominator coefficients b_1 and b_2 of Eq. (1) leads to the pole distribution shown in Fig. 1(a). The pole distribution in the second quadrant is the mirror image of the first quadrant. Fig. 1(b) shows the block diagram of the corresponding recursive part of Eq. (1).

A different description of the denominator polynomial is given by

$$H(z) = \frac{N(z)}{1 + 2r \cos \varphi z^{-1} + r^2 z^{-2}} \tag{2}$$

where $N(z)$ is the nominator polynomial, r the radius, and φ the phase of the complex-conjugate poles. The quantization of r and φ leads to a different pole distribution compared to the direct-form structure given by Eq. (1). The state-variable structure of [2], [3] is based on the approach by Gold and Rader [4], which

* Presented at the 91st Convention of the Audio Engineering Society, New York, 1991 October 4–8.

is given by

$$H(z) = \frac{N(z)}{1 + 2\mathrm{Re}\{z_\infty\}\, z^{-1} + [\mathrm{Re}\{z_\infty\}^2 + \mathrm{Im}\{z_\infty\}^2]\, z^{-2}} \tag{3}$$

where z_∞ denotes the pair of complex-conjugate poles, $\mathrm{Re}\{z_\infty\}$ the real, and $\mathrm{Im}\{z_\infty\}$ the imaginary part of the pole pair.

The possible pole locations are shown in Fig. 2(a). $\mathrm{Re}\{z_\infty\}$ and $\mathrm{Im}\{z_\infty\}$ are quantized to 6 bit. Again the pole distribution in the second quadrant is the mirror image of the first quadrant. The block diagram of the recursive part of Eq. (3) is shown in Fig. 2(b). Compared to the direct quantization of coefficients b_1 and b_2, the quantization of the real and imaginary parts of the complex-conjugate poles offers a uniform grid of pole locations.

The frequencies (20 Hz to 5 kHz) where the greatest influence is required are usually small compared to the sampling frequency of 48 kHz. This fact places the poles of the second-order transfer function close to the point $z = 1$ in the z plane. In order to increase the pole density for very low cutoff frequencies Kingsbury [5] proposed a digital filter structure with the transfer function

$$H(z) = \frac{N(z)}{1 - (2 - k_1k_2 - k_1^2)z^{-1} - (1 - k_1k_2)z^{-2}}. \tag{4}$$

Coefficients b_1 and b_2 of Eq. (1) are now realized by

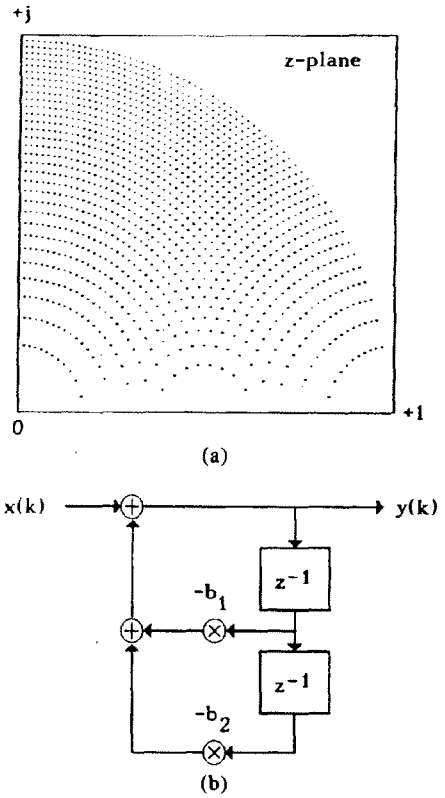


Fig. 1. Direct-form structure. (a) Pole distribution (6 bit). (b) Recursive part.

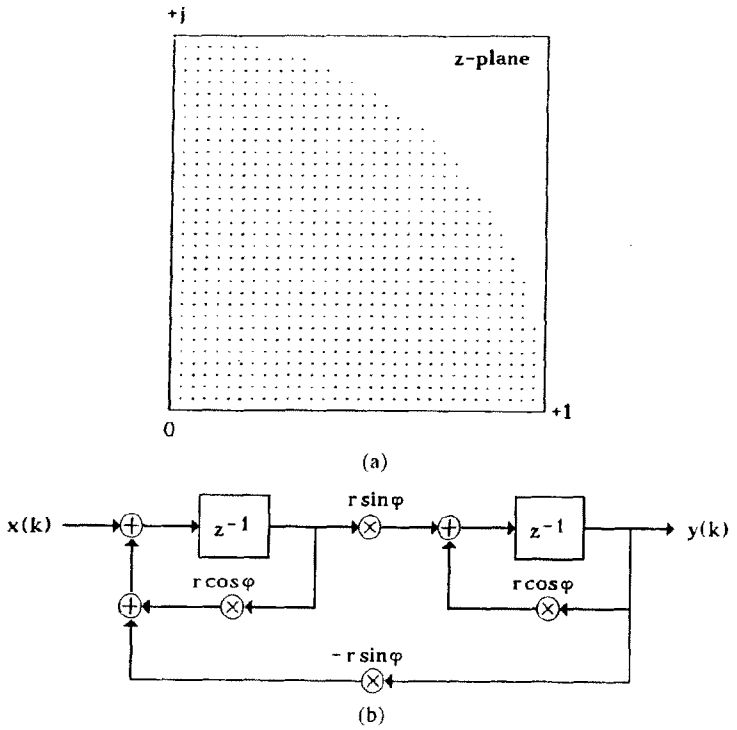


Fig. 2. Gold–Rader structure. (a) Pole distribution (6 bit). (b) Recursive part.

a linear combination of new coefficients k_1 and k_2, which are given by

$$k_1 = d = \sqrt{1 - 2r\cos\varphi + r^2} \tag{5}$$

$$k_2 = \frac{1 - r^2}{k_1} . \tag{6}$$

A geometric interpretation is shown in Fig. 3, where the distance $d = k_1$. The pole distribution is demonstrated in Fig. 4(a) and the block diagram of Eq. (4) in Fig. 4(b).

A further increase of the pole density near $z = 1$ can be achieved [6], [7]. This new second-order transfer function is given by

$$H(z) = \frac{N(z)}{1 - (2 - z_1z_2 - z_1^3)z^{-1} - (1 - z_1z_2)z^{-2}} . \tag{7}$$

It can be shown that

$$z_1 = \sqrt[3]{1 + b_1 + b_2} \tag{8}$$

$$z_2 = \frac{1 - b_2}{z_1} . \tag{9}$$

The distance d from the point $z = 1$ is related to the coefficient z_1 by

$$z_1 = \sqrt[3]{d^2} . \tag{10}$$

This nonlinear relationship between the distance d from the point $z = 1$ and the coefficient z_1 can be seen in Fig. 5(a), where the pole density of the new structure is shown. The improvement of the pole density at low frequencies is obvious. The corresponding block diagram is shown in Fig. 5(b).

The pole distributions of the Kingsbury structure and the new-filter structure show a reduction of the pole density for higher pole frequencies. For the pole density a symmetry to the imaginary axis as in the case of the direct-form structure and the Gold–Rader structure is not possible. But changing a sign in the recursive part of the difference equation results in a mirror image of the pole density. This mirror image can be achieved through the change of sign in the denominator polynomial

$$D(z) = 1 \overset{!}{\pm} (2 - z_1z_2 - z_1^3)z^{-1} - (1 - z_1z_2)z^{-2} . \tag{11}$$

The pole distributions for the Kingsbury structure and the new structure are shown in Fig. 6.

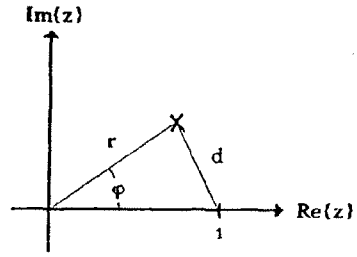


Fig. 3. Pole locations in z plane.

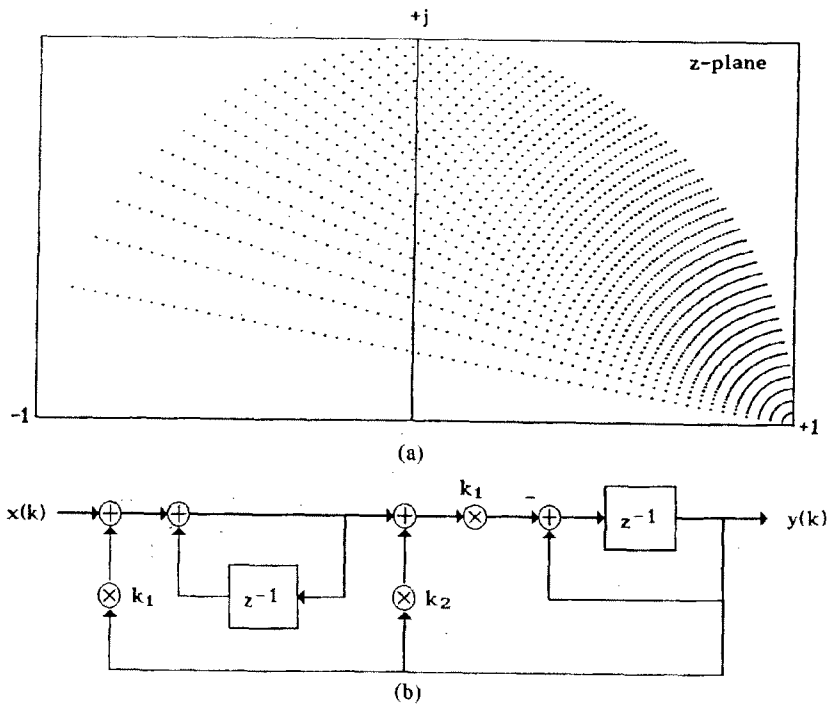


Fig. 4. Kingsbury structure. (a) Pole distribution (6 bit). (b) Recursive part.

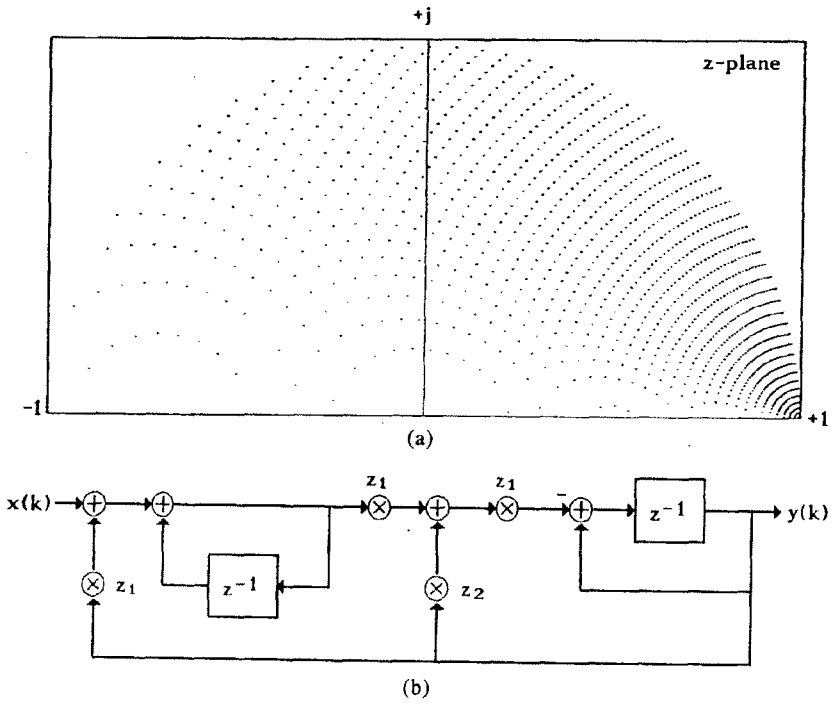


Fig. 5. New structure. (a) Pole distribution (6 bit). (b) Recursive part.

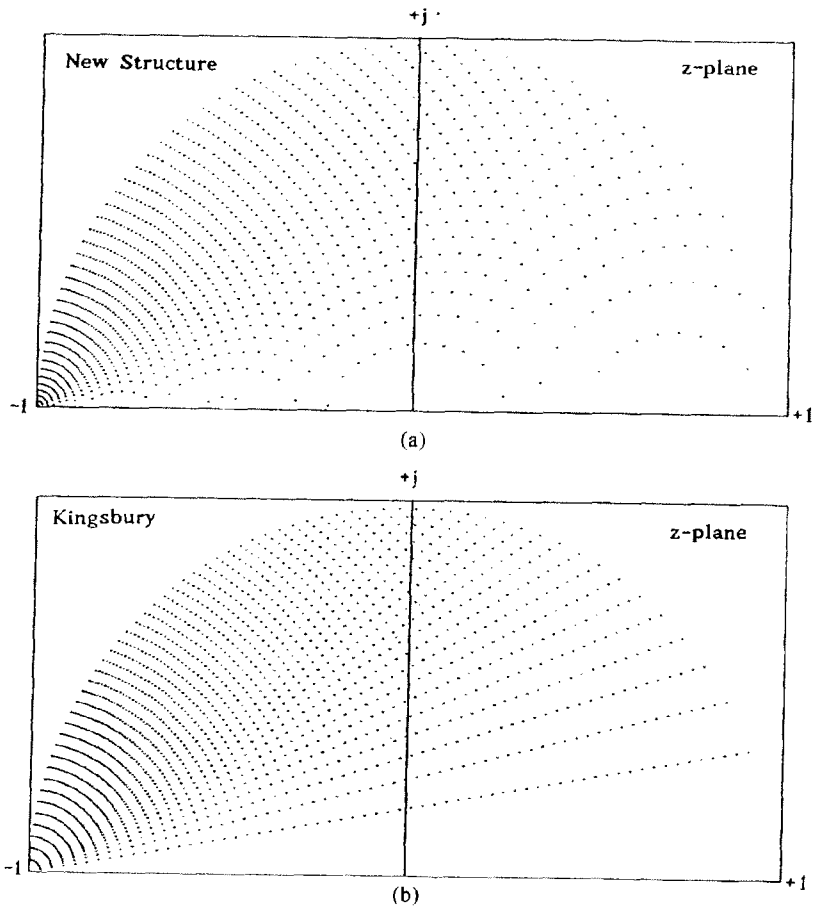


Fig. 6. Pole distributions (sign changed). (a) New structure. (b) Kingsbury structure.

2 ROUNDOFF ERROR ANALYSIS

In this section recursive filter structures are analyzed. The comparison is based on fixed-point arithmetic. The results can be adopted for floating-point arithmetic, because the mantissa of the floating-point arithmetic is mainly responsible for the roundoff noise behavior. The block diagrams of the different digital filter structures are the bases for the roundoff error analysis. First the general case is considered when quantization is performed after each multiplication. The following assumptions must be valid [8]–[11]:

1) The error sequence $e(k)$ is a white-noise stationary process.

2) The error sequence is uncorrelated with the exact signal $x(k)$.

3) The probability of the error process is uniform over the range of quantization error.

4) Different error sequences $e_i(k)$ are uncorrelated.

The variance of the quantization noise is easily shown to be

$$\sigma_e^2 = \frac{Q^2}{12} \tag{12}$$

where Q is the quantization step size. The quantization error is added at each quantization point in the filter structure and is filtered by a specific transfer function to the output of the filter. Therefore each transfer function $G_i(z)$ from the multiplier outputs to the output of the filter structure is determined. The variance of the output quantization noise is given by

$$\sigma_{ye}^2 = \sigma_e^2 \frac{1}{2\pi j} \oint_{z=e^{j\Omega}} G(z)G(z^{-1})z^{-1}\, dz \ . \tag{13}$$

Exact solutions of the ring integral of Eq. (13) can be found in [12] for transfer functions up to the order of 4. With the quadratic L_2 norm

$$\|G\|_2^2 = \frac{1}{\pi} \int_0^{\pi} G(e^{j\Omega})^2\, d\Omega \tag{14}$$

the superposition of the noise variances at the output is given by

$$\sigma_{ye}^2 = \sigma_e^2 \sum_i \|G_i\|_2^2 \ . \tag{15}$$

The signal-to-noise ratio (SNR) can be written as

$$\mathrm{SNR} = 10 \log \frac{0.5}{\sigma_{ye}^2} \ \ [\mathrm{dB}] \ . \tag{16}$$

Figs. 7–10 contain the block diagrams, the corre-

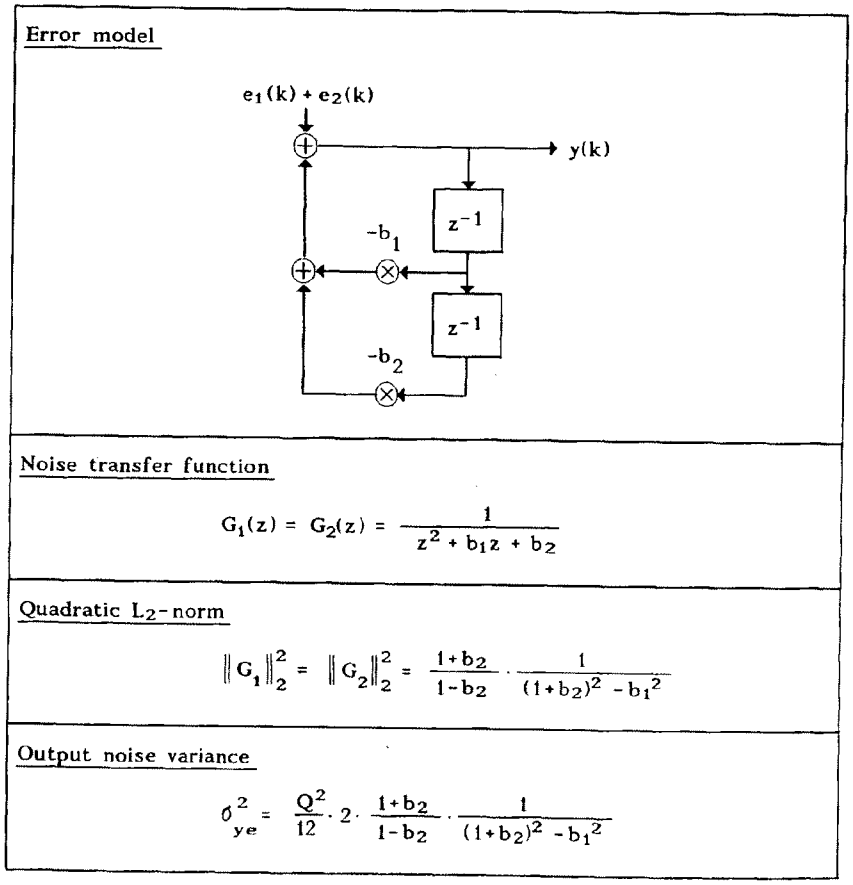


Fig. 7. Direct-form structure.

sponding noise transfer functions $G_i(z)$, the quadratic L_2 norms, and the noise variances of the different recursive filter structures. Especially the noise transfer functions are responsible for the spectrum shaping of the quantization errors.

The noise transfer function of the direct form contains only a denominator polynomial and no zeros. The realization of poles close to the unit circle leads to high amplification of roundoff noise. The influence of the pole radius on the noise variance can be seen in the equation for the output noise variance in Fig. 7, where the coefficient $b_2 = r^2$ approaches 1.

The Gold–Rader filter has an output noise variance (Fig. 8) that shows the dependence on the pole radius and no influence on the pole phase. This fact is based on the uniform pole grid of this recursive filter structure. One additional zero on the real axis ($z_0 = r \cos \varphi$) in one noise transfer function reduces the effect of the complex-conjugate poles.

The Kingsbury filter structure (Fig. 9) and the related new recursive filter structure (Fig. 10) exhibit a similar dependence on the pole radius. In the noise transfer functions additional zeros at $z = 1$ are present which are not influenced by the complex-conjugate poles.

Analytical results concerning the SNR of the new structure in comparison with the direct form, the Gold–Rader structure, and the Kingsbury structure are shown in Fig. 11. The state variables are quantized to 16 bit and the pole location is moved on a curve obtained by a bilinear transform of a low-pass filter with $Q = 0.7071$. In Fig. 11(a) the pole frequency is varied from 20 to 200 Hz. This is a typical region for very low-frequency audio filters. Compared with the other structures, the new filter structure has an improved SNR in this low-frequency region (sampling frequency $f_s = 48$ kHz). Up to 5 kHz the new structure yields better results. Approaching one quarter of the sampling frequency (12 kHz), reduction of the pole density can be noted [see Fig. 11(b)].

In view of a signal processor implementation the quantization does not have to be performed after each multiplication. As long as product terms can be added in an accumulator of double-precision word length, quantization can be avoided. Quantization must only be done when storing the results. This is demonstrated in Fig. 12 for the different filter structures. The resulting

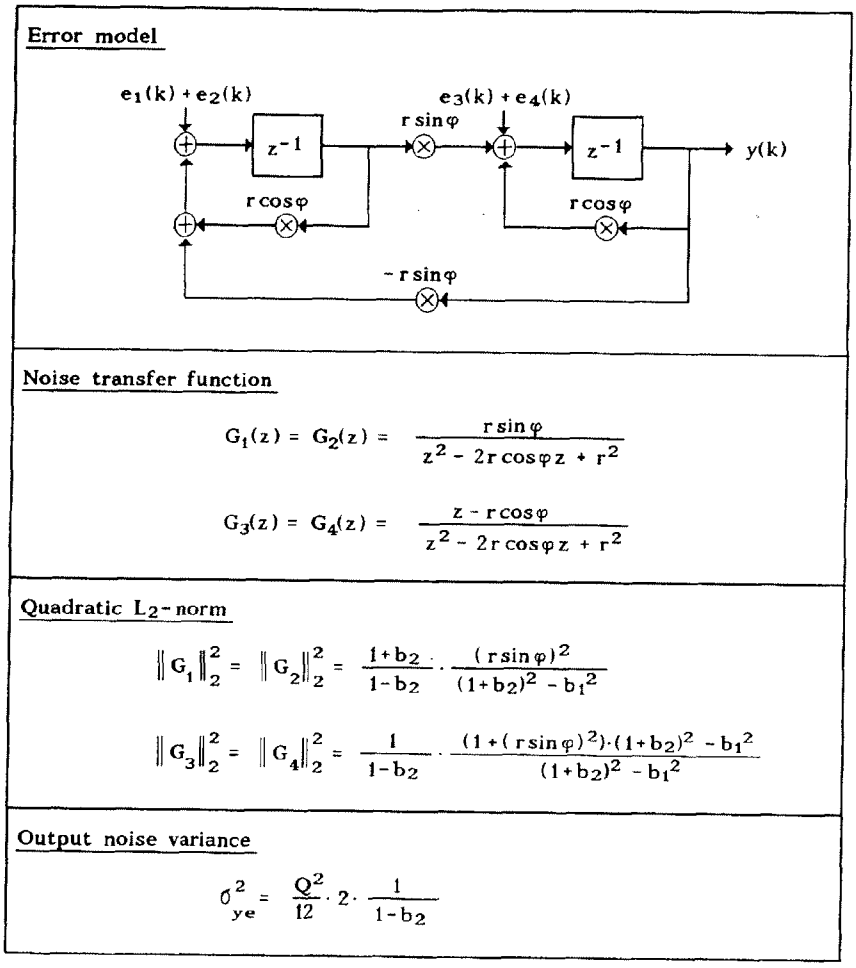


Fig. 8. Gold–Rader structure.

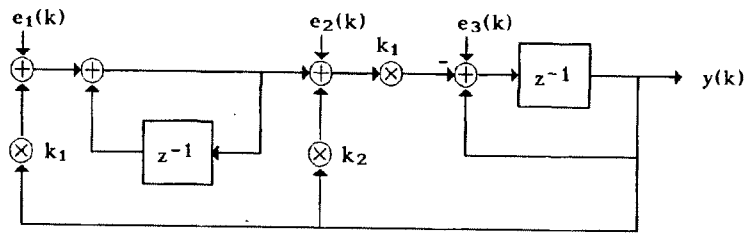
expressions for the noise variances are given above the block diagrams. The SNRs vesus the pole frequency are shown in Fig. 13. For the direct-form and the Gold–Rader structures an improvement of 3 dB is obtained. The Kingsbury and the Gold–Rader structures, show similar results, while the new structure is 3 dB better [see Fig. 13(a)]. In addition to the small coefficient sensitivity the new structure shows a better roundoff noise behavior for frequencies up to 2 kHz [Fig. 13(b)].

3 APPLICATION OF ERROR SPECTRUM SHAPING

The analysis of the noise transfer functions shows that filter structures with zeros in the noise transfer functions reduce the influence of the complex-conjugate poles. This becomes more significant when the poles approach the unit circle and especially the point $z = 1$. Roundoff noise and limit cycles [6] can be reduced further through artificial zeros in the noise transfer functions which cancel the effects of the poles. This technique is called error spectrum shaping (ESS) [13] because it performs a manipulation of the noise transfer function. In this section this technique is described and its application to recursive filter structures demonstrated.

Quantization can be modeled as the sum of the ideal output $x(k)$ and the error signal $e(k)$, as shown in Fig. 14. When there is access to the error signal $e(k)$, a modified model [see Fig. 15] can be considered. The access to the error signal is possible if the signal processing device has an accumulator of double-precision word length. The lower part of the accumulator represents the quantization error. As a simple example consider the feedback arrangement of Fig. 16. The

Error model



Noise transfer function

$$G_1(z) = \frac{-k_1 z}{z^2-(2-k_1k_2-k_1^2)z+(1-k_1k_2)}$$

$$G_2(z) = \frac{-k_1(z-1)}{z^2-(2-k_1k_2-k_1^2)z+(1-k_1k_2)}$$

$$G_3(z) = \frac{z-1}{z^2-(2-k_1k_2-k_1^2)z+(1-k_1k_2)}$$

Quadratic L_2-norm

$$\|G_1\|_2^2 = \frac{1}{k_1k_2}\cdot\frac{2-k_1k_2}{2\cdot(2-k_1k_2)-k_1^2}$$

$$\|G_2\|_2^2 = \frac{k_1}{k_2}\cdot\frac{2}{2\cdot(2-k_1k_2)-k_1^2}$$

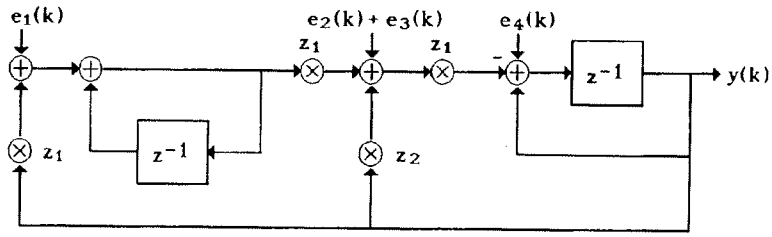
$$\|G_3\|_2^2 = \frac{1}{k_1k_2}\cdot\frac{2}{2\cdot(2-k_1k_2)-k_1^2}$$

Output noise variance

$$\sigma_{ye}^2 = \frac{Q^2}{12}\cdot\frac{5+2\cdot b_1+3\cdot b_2}{(1-b_2)\cdot(1+b_2-b_1)}$$

Fig. 9. Kingsbury structure.

Error model



Noise transfer function

$$G_1(z) = \frac{-z_1^2\, z}{z^2-(2-z_1z_2-z_1^3)z+(1-z_1z_2)}$$

$$G_2(z) = G_3(z) = \frac{-z_1(z-1)}{z^2-(2-z_1z_2-z_1^3)z+(1-z_1z_2)}$$

$$G_4(z) = \frac{z-1}{z^2-(2-z_1z_2-z_1^3)z+(1-z_1z_2)}$$

Quadratic L_2-norm

$$\|G_1\|_2^2 = \frac{z_1^4}{z_1z_2}\cdot\frac{2-z_1z_2}{2\cdot z_1^3\cdot(2-z_1z_2)-z_1^6}$$

$$\|G_2\|_2^2 = \|G_3\|_2^2 = \frac{z_1^6}{z_1z_2}\cdot\frac{2}{2\cdot z_1^3\cdot(2-z_1z_2)-z_1^6}$$

$$\|G_4\|_2^2 = \frac{z_1^3}{z_1z_2}\cdot\frac{2}{2\cdot z_1^3\cdot(2-z_1z_2)-z_1^6}$$

Output noise variance

$$\sigma_{ye}^2 = \frac{Q^2}{12}\cdot\frac{6+4\cdot(b_1+b_2)+(1+b_2)\cdot(1+b_1+b_2)^{1/3}}{(1-b_2)\cdot(1+b_2-b_1)}$$

Fig. 10. New structure.

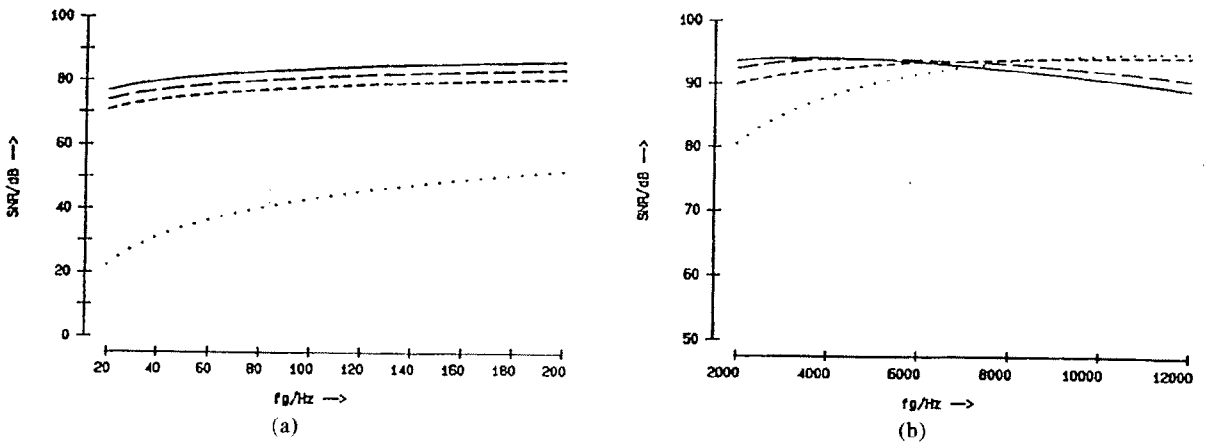
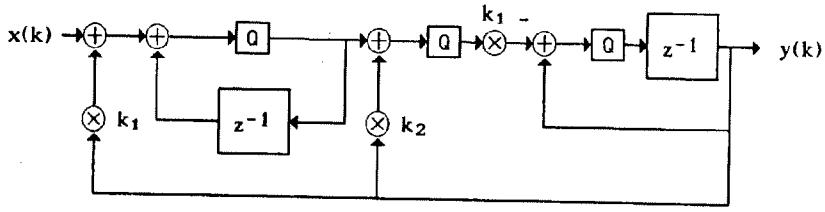
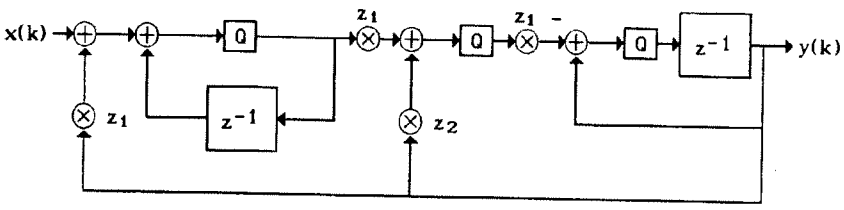


Fig. 11. SNR versus cutoff frequency (quantized produce terms). 16-bit word length. ——— New structure; — — — Kingsbury; - - - Gold–Rader; ····· Direct form.

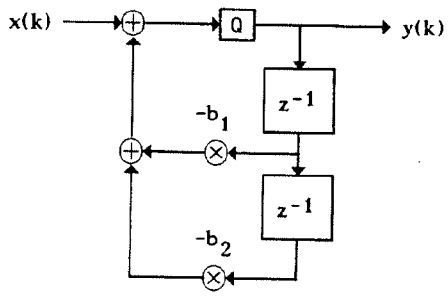
Kingsbury $\sigma_{ye}^2 = \frac{Q^2}{12} \cdot \frac{5 + 2 \cdot b_1 + 3 \cdot b_2}{(1-b_2) \cdot (1+b_2-b_1)}$



New structure $\sigma_{ye}^2 = \frac{Q^2}{12} \cdot \frac{2 \cdot (2 + b_1 + b_2) + (1+b_2) \cdot (1+b_1+b_2)^{1/3}}{(1-b_2) \cdot (1+b_2-b_1)}$



Direct form $\sigma_{ye}^2 = \frac{Q^2}{12} \cdot \frac{1+b_2}{1-b_2} \cdot \frac{1}{(1+b_2)^2 - b_1^2}$



Gold/Rader $\sigma_{ye}^2 = \frac{Q^2}{12} \cdot \frac{1}{1-b_2}$

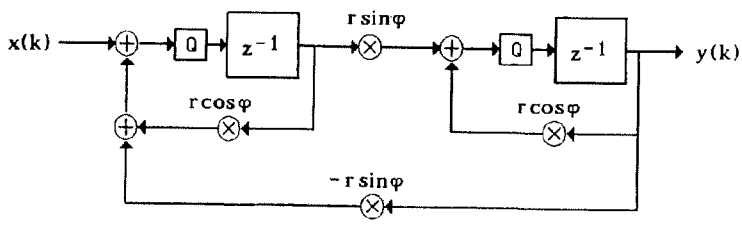


Fig. 12. Block diagrams (quantized sum terms).

quantized output signal can be written as

$$[x(k)]_Q = x(k) + e(k) - e(k-1) \tag{17}$$

and the output error signal is given by

$$e_o(k) = [x(k)]_Q - x(k) = e(k-1) \tag{18}$$

with the z transform

$$E_o(z) = (1 - z^{-1})E(z) \ . \tag{19}$$

Since there is an artificial zero at $z = 1$ for the noise source in Eq. (19), this ESS has no effect on the ideal input signal $x(k)$.

This simple error feedback will be first applied to the direct-form structure, giving insight into the use of this technique. One simple zero at $z = 1$ leads to the modified noise transfer function

$$G_1(z) = \frac{1 - z^{-1}}{1 + b_1 z^{-1} + a_2 z^{-2}} \ . \tag{20}$$

The resulting output noise variance and the block diagram are given in Fig. 17(a). The generation of a double zero at $z = 1$ is possible with the feedback structure shown in Fig. 17(b). The corresponding noise transfer function is given by

$$G_1(z) = \frac{1 - 2z^{-1} + z^{-2}}{1 + b_1 z^{-1} + a_2 z^{-2}} \ . \tag{21}$$

With decreasing pole frequency the coefficient b_1 in Eq. (21) approaches -2 and the coefficient b_2, 1. Thus the error source is filtered by a high-pass transfer function without any amplification. The shaping technique affects only the error source itself and not the input signal, which is only manipulated by the transfer function, Eq. (1). If one chooses the feedback coefficients for the error signal equal to the denominator coefficients b_1 and b_2, complex-conjugate zeros are generated in the nominator polynomial of Eq. (21), which exactly

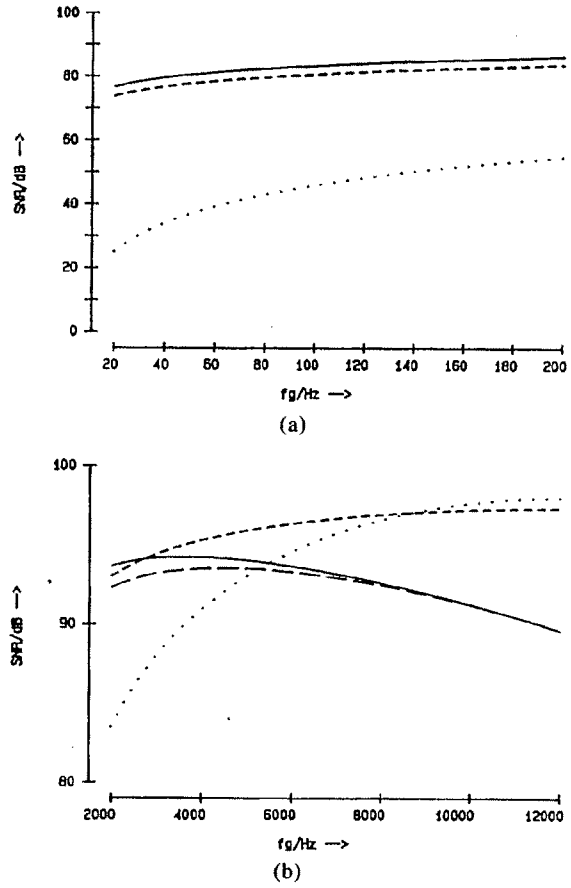


Fig. 13. SNR versus cutoff frequency (quantized sum terms). 16-bit wordlength. ——— New structure; — — — Kingsbury; - - - Gold–Rader; ····· Direct form.

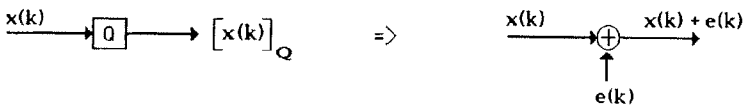


Fig. 14. Error model.

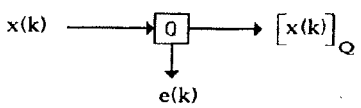


Fig. 15. Error model with access to error signal.

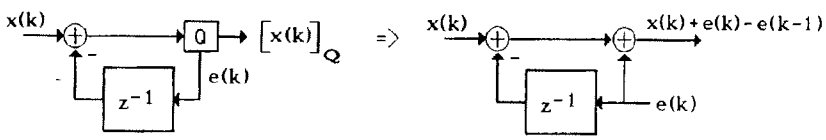


Fig. 16. Error model with feedback.

cancel the complex-conjugate poles. The noise transfer function reduces to the constant 1. This choice of the feedback coefficients can be viewed as a double-precise arithmetic and is a computation time consuming task. But simple nominator polynomials with easy to implement coefficients lead to very effective noise canceling with a very small computation load.

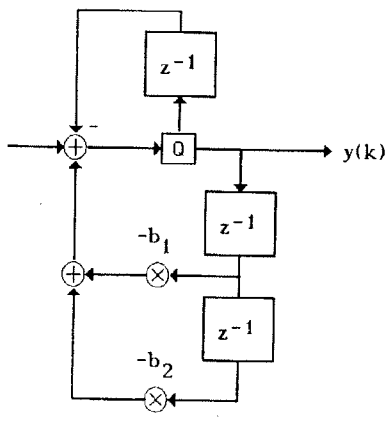
Since the Gold–Rader, Kingsbury, and new structures have already zeros in their noise transfer functions, a simple additional zero is sufficient. The application of error spectrum shaping to these structures is shown in Figs. 18–20 with the corresponding output noise variances.

The influence of the shaping technique on the SNR is analyzed with the methods of the previous section. The signal word length is chosen as 16 bit. The curves for the direct form are shown in Fig. 21 for a simple zero and a double zero at $z = 1$. Even the simple zero gives a great improvement. Fig. 22(a) demonstrates the ideal noise behavior for all structures (double zero in the direct form) in the low-frequency region. The curves for increasing pole frequencies are shown in Fig. 22(b), where the double zeros at $z = 1$ have reduced their influence. But the SNRs are still beyond the limit of 90 dB for a 16-bit arithmetic. Extension of the demonstrated results to larger signal word lengths can simply be done in adding 6 dB for every additional bit.

4 CONCLUSION

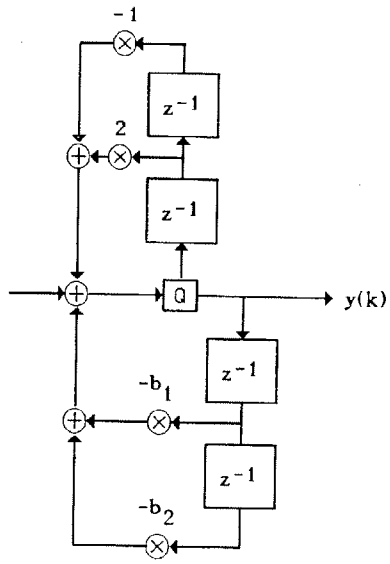
Considerations on different filter topologies and the connection between coefficient sensitivity and roundoff noise in digital filters led to the approach of a new digital filter structure with increased pole density at

$$\sigma^2_{ye} = \frac{Q^2}{12} \cdot \frac{2}{(1-b_2)\cdot(1+b_2-b_1)}$$



(a)

$$\sigma^2_{ye} = \frac{Q^2}{12} \cdot \frac{6 + 2\cdot b_1 - 2\cdot b_2}{(1-b_2)\cdot(1+b_2-b_1)}$$



(b)

Fig. 17. Direct-form structure with ESS. (a) One simple zero. (b) Double zero.

$$\sigma^2_{ye} = \frac{Q^2}{12} \cdot \frac{2+b_1}{(1-b_2)}$$

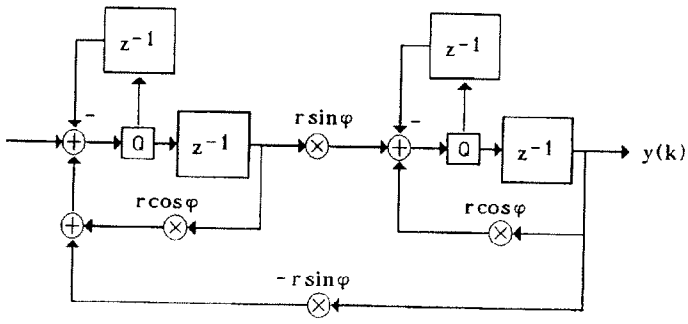


Fig. 18. Gold–Rader structure with ESS.

very low cutoff frequencies. A comparison with existing solutions on the basis of a roundoff error analysis shows a better performance of the new structure concerning the two main effects, namely, coefficient quantization and signal quantization. The application of the ESS technique is demonstrated for different filter topologies. It leads to a reduction of the signal quantization effects.

5 REFERENCES

[1] A. Fettweis, "On the Connection between Multiplier Wordlength Limitation and Roundoff Noise in Digital Filters," *IEEE Trans. Circuit Theory*, pp. 486–491 (1972 Sept.).

[2] C. T. Mullis and R. A. Roberts, "Synthesis of Minimum Roundoff Noise Fixed Point Digital Filters," *IEEE Trans. Circuits Sys.*, pp. 551–562 (1976 Sept.).

[3] B. W. Bomar, "New Second-Order State-Space Structures for Realizing Low Roundoff Noise Digital Filters," *IEEE Trans. Acoust., Speech, Signal Proc.*, pp. 106–110 (1985 Feb.).

[4] B. Gold and C. M. Rader, "Effects of Parameter Quantization on the Poles of a Digital Filter," *Proc. IEEE*, pp. 688–689 (1967 May).

[5] N. G. Kingsbury, "Second-Order Recursive Digital Filter Element for Poles near the Unit Circle and the Real Axis," *Electron. Lett.*, pp. 155–156 (1972 Mar.).

[6] U. Zoelzer, "Entwurf digitaler Filter für die Anwendung im Tonstudiobereich," Ph. D. thesis, Tech-

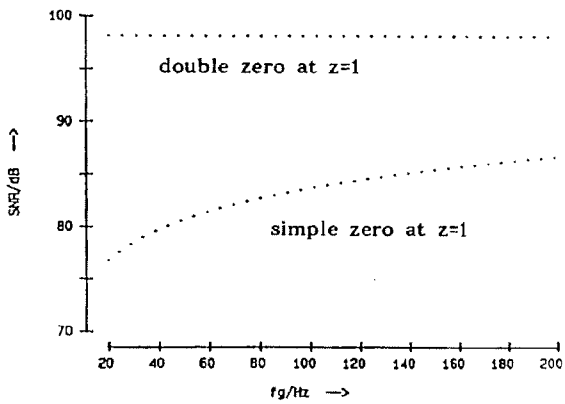


Fig. 21. Direct-form structure, SNR versus cutoff frequency (ESS).

$$\sigma_{ye}^2 = \frac{Q^2}{12} \cdot \frac{(1+k_1^2)\cdot((1+b_2)(6-2b_2)+2b_1^2+8b_1)+2k_1^2(1+b_1+b_2)}{(1-b_2)\,(1+b_2-b_1)\cdot(1+b_2+b_1)}$$

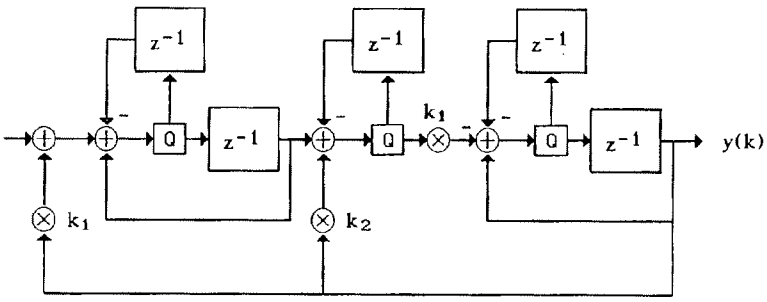


Fig. 19. Kingsbury structure with ESS.

$$\sigma_{ye}^2 = \frac{Q^2}{12} \cdot \frac{(1+z_1^2)\cdot((1+b_2)(6-2b_2)+2b_1^2+8b_1)+2z_1^4(1+b_1+b_2)}{(1-b_2)\,(1+b_2-b_1)\cdot(1+b_2+b_1)}$$

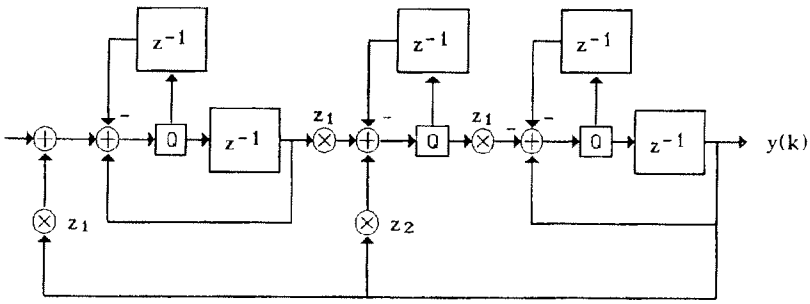


Fig. 20. New structure with ESS.

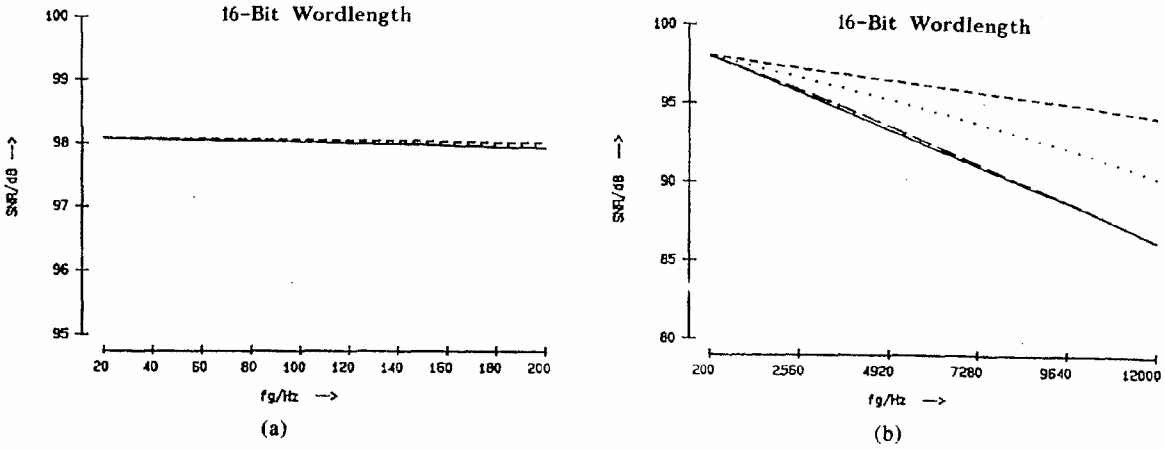


Fig. 22. SNR versus cutoff frequency (ESS). 16-bit word length. ——— New structure; — — — Kingsbury; - - - Gold–Rader; ····· Direct form.

nical University of Hamburg, Germany (1989).

[7] U. Zoelzer, "A Low Roundoff Noise Digital Audio Filter," in *Proc. EUSIPCO-90* (1990 Sept.), pp. 529–532.

[8] A. B. Sripad and D. L. Snyder, "A Necessary and Sufficient Condition for Quantization Errors to Be Uniform and White," *IEEE Trans. Acoust., Speech, Signal Proc.*, pp. 442–448 (1977 Oct.).

[9] T. A. C. M. Claasen and A. Jongepier, "Model for the Power Spectral Density of Quantization Noise," *IEEE Trans. Acoust., Speech, Signal Proc.*, pp. 914–917 (1981 Aug.).

[10] C. W. Barnes, B. N. Tran, and S. H. Leung, "On the Statistics of Fixed-Point Roundoff Error," *IEEE Trans. Acoust., Speech, Signal Proc.*, pp. 595–606 (1985 June).

[11] I. Tokaji and C. W. Barnes, "Roundoff Error Statistics for a Continous Range of Multiplier Coefficients," *IEEE Trans. Circuits Sys.*, pp. 52–59 (1987 Jan.).

[12] E. I. Jury, *Theory and Application of the z-Transform Method* (Wiley, New York, 1964).

[13] Tran-Thong and B. Liu, "Error Spectrum Shaping in Narrow Band Recursive Filters," *IEEE Trans. Acoust., Speech, Signal Proc.*, pp. 200–203 (1977 Apr.).

THE AUTHOR

Udo Zölzer was born in Arolsen, Germany, in 1958. He graduated from the University of Paderborn with the Diplom-Ingenieur degree in electrical engineering in 1985. After graduation he began working at the Department of Telecommunications, Technical University of Hamburg-Harburg, sponsored by Lawo Gerätebau GmbH, Germany, and since 1988 sponsored by Monitora Gerätebau GmbH, Germany. After completing his Ph.D. thesis on digital filters applied to audio processing in 1989, he became chief engineer at the Department of Telecommunications at the university. Since 1992 he has been a lecturer in digital audio processing. His research interests are multirate signal processing and the applications of digital signal processing to audio and acoustics.

Dr. Zölzer has had several papers published at AES conventions and at IEEE conferences in the field of audio signal processing. He is founder of ZIP Audio Technology and a member of the Audio Engineering Society and the IEEE.

Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT

ELSEVIER

Signal Processing 83 (2003) 1565–1582

SIGNAL PROCESSING

www.elsevier.com/locate/sigpro

Noise optimized IIR digital filter design—tutorial and some new aspects

Günter F. Dehner*

*Ingenieurbüro Dehner*** qDSP Design & Applications, Buckenhofer Weg 52, D-91058 Erlangen, Germany*

Received 30 October 2002; received in revised form 17 February 2003

Abstract

IIR digital filters can be implemented in a large number of different structures. In addition, there are various possibilities for the individual realization of these structures, depending on coefficient format, state variable format, quantization points and used overflow characteristics. An overview is given about the whole filter design procedure for cascades with second order sections (SOS) in the direct forms and in the state space form, including scaling and calculation of the noise figure for the selected structures and hardware realizations. For SOS in the state space form a noise optimized transformation is evaluated as a closed solution. The best pairing and ordering of poles and zeros within the cascade is calculated by "Dynamic Programming", an efficient optimization method for the allocation problem. Optimization is possible by the mean noise power as well as by the peak value of the noise power spectrum. Based on the exact solutions, rules of thumb are examined and further improved. The methods for optimized filter implementations are tested by an ensemble of filters with varying zero and pole positions. The results illustrate that, according to the transfer function, an appropriate choice of the SOS forms together with an optimization of pairing and ordering is useful for different hardware realizations.

1. Introduction

Digital filters, recursive with infinite impulse response (IIR) as well as non-recursive with finite-length impulse response (FIR), have been investigated in the early days of Digital Signal Processing by several research groups all over the world, in Europe especially, by Prof. Schüssler and his team at Erlangen-Nürnberg University. The results have been published in textbooks [34,35] and a large number of scientific contributions (e.g. [1–5,9,12–16,18]), covering the design of filters with low coefficient sensitivity as well as the quantization effects of state variables and the noise and limit-cycle behavior of systems with finite word length. The design of selective IIR filters with minimal coefficient word length as well as noise-optimized allocation of poles and zeros in cascade structures of second order sections in the two canonic forms were especially examined in [6–9]. The results contributed to a design and optimization program for IIR-filters in different hardware structures [10]. Results for finite word length FIR filter design were published in [19].

Early studies were mainly focused on the realization of the filters using discrete digital circuits or on the design of fast customer-designed chips. With the introduction of Digital Signal Processors in multi-step pipeline technique, some of the optimization strategies had to be thought over [28,36]. From the beginning, the focus was on digital filters, to be applied in digital measurement and communication techniques, and realized by special purpose hardware [16,17,25,29,33]. In the meantime, digital signal processing is more and more done by computers, and word-length effects no longer seem to play such an important role. However, with the introduction of portable and battery powered telecommunication devices such as digital mobile GSM-phones a new "optimization era" to low-cost and low-power systems started again.

Numerous different structures for the realization of digital filters were investigated in the past, including cascade and parallel structures of direct forms as well as wave digital filters as structurally loss-less implementations of allpass filters [30]. Some of them are integrated into commercially available filter design systems like the MATLAB® System [11]. Still, these procedures have to be improved and adapted to the needs of modern hardware design.

In this paper, cascade structures of digital IIR filters are reviewed, and some novel results with respect to scaling, to noise behavior, optimal allocation of the poles and zeros and therefore reduced hardware costs are presented. Some of the well-known analysis and optimization procedures are improved and modified for different filter structures, including the state space structure. Rules of thumbs for the filter design are given and are verified by comparison with results obtained by extensive computer based optimization procedures.

The paper is organized as follows:

- A tutorial on the fixed-point IIR filter design, basic literature and some design tools are given in the next section.
- In Section 3 we describe the components for fixed-point realizations in more detail, including entire, fractional and shift-multipliers, quantizers and scaling operation.
- In Section 4 we introduce cascades of second order sections in two direct forms and the state space form. Optimized parameters of the state space form are developed in a closed form.
- In Section 5 we explain the criteria—noise figure and relative peak noise power—as a measure for optimal allocation of poles and zeros and compare some improved rules of thumb with the optimal allocation obtained by 'Dynamic Programming'.
- Filter designs for different structures and hardware realizations are compared in Section 6 for low-passes of 8th order by the so-called "Allpass Test".

2. General IIR filter design procedure

The design process of recursive digital filters (IIR Filter) normally starts with the specification of the filter response in the frequency or time domain. For a frequency domain specification the desired tolerance scheme for the magnitude and/or group delay is given. The various methods for the design of the transfer function, independent of word length effects are described in text books e.g. [23,30,32,34,35]. Most of them include examples for the filter design, written in Fortran, C or a problem oriented language, like MATLAB® Language of Technical Computing [11]. In special, the design methods are supplemented and consolidated by MATLAB® procedures in [23,27,35]. Having designed the transfer functions, the choice of an appropriate structure and their realization forms are necessary. These realizations will not only differ in the general structure (signal flow graph), but also in the layout of the multipliers (coefficient and state variable format), the adders (single or double precision) as well as quantizers (rounding, truncation) and overflow management.

The whole design procedure consists of a number of separate steps, which are listed for completeness here again:

- Definition of filter transfer function by the coefficients of the numerator and denominator polynomial or by their zeros and poles.
- Building a cascade structure of Second Order Sections (SOS) and determination of pairing and ordering of the poles and zeros in these SOS by a rule of thumb or an optimization procedure minimizing the quantization noise at the filter output.
- Scaling the SOS structure for different fixed-point implementations depending on the chosen allocation of the poles and zeros or scaling and

* Tel.: +49-9131-38082; fax: +49-9131-710295.
E-mail address: guenter.dehner@t-online.de (G.F. Dehner).

0165-1684/03/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0165-1684(03)00075-6

calculation of the noise contribution as a part of the optimization process.

- Calculation of the scaled transfer function with finite word length and determination of the theoretical noise behavior of the final allocation.
- Simulation of the SOS structure in fixed-point arithmetic and measurement of the final frequency response and noise power spectrum.

Ordering and scaling of the SOS is absolutely necessary for cascade structures in fixed-point arithmetic. For the optimization of the allocation of poles and zeros in the cascade a heuristic procedure was proposed by Avenhaus [1] and by Jackson [22]. An exact procedure using the methods of "Dynamic Programming" was introduced by Hwang [20] and Lüder [26]. These procedures were improved and integrated into a complete design program system by the author [7,9,10]. These optimization programs, once written in FORTRAN, are still of interest and are now transferred into MATLAB® Language for Technical Computing [11]. In the meantime, a large number of structures, including features for minimum hardware realizations are investigated with respect to their noise behavior and integrated into new and improved analysis and optimization procedures.

Before further discussions of the scaling procedures the elements for fixed-point arithmetic are reviewed or as far as necessary introduced in the next section. These elements are building blocks for the modeling of the hardware realizations of the cascade structures given in Section 4.

3. Fixed-point implementations

Fixed-point arithmetic is often defined for fractional multiplications only. In this case the magnitude of multiplicand and multiplier are fractional values <1 (sfrac).[1] For general applications and especially for the digital filter realizations in direct form, coefficients are not restricted to absolute values <1. Therefore we use general fixed-point numbers (sfix)[1] for the definition of the coefficients, allowing absolute values >1.

[1] sfix, sfrac: Definitions of fixed-point numbers with scaling factors [11].

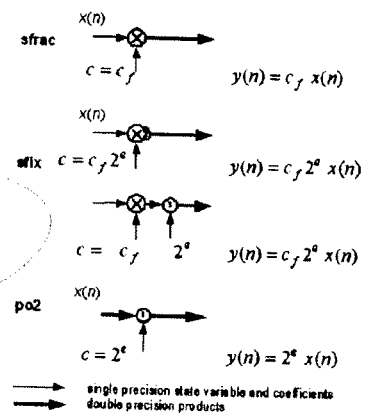


Fig. 1. Realization of fixed-point multipliers.

These entire multipliers can be implemented by a fractional multiplication followed by an additional shifter. Sometimes coefficients in sfix-format are also called "pseudo"- float-point coefficients, due to the fact that the mantissa is normally not normalized and the quantization of the coefficient is varying with the shift factor (or exponent).

Multipliers for fractional coefficients (sfrac) are marked in the block diagrams by a simple multiplier symbol, see Fig. 1. For non-fractional (sfix) or pseudo-float-point coefficients the symbol shows an additional shifter after the fractional multiplication. In some cases entire multipliers can be substituted by simple shifters using a power-of-two (po2-format) as multiplicand. To avoid additional quantization the shift-multiplication can be realized in double precision, see the last symbol in Fig. 1.

In a fixed-point arithmetic the multiplication of the state variable by the fixed-point coefficient, both in single precision, results in a double precision fixed-point product. Single or double precision numbers are marked in the block diagrams by normal and bold arrows. Summing up the double precision products, quantization is in any case necessary in front of the next multiplication. The quantizers are denoted in the block diagrams by the quantizer symbol, defined in Fig. 2. Quantization can be achieved by any of the

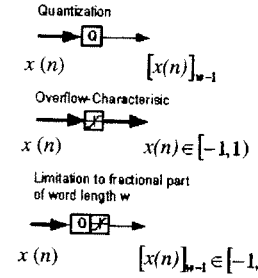


Fig. 2. Elements for quantization and overflow check.

known quantization characteristics described in the text books, e.g. [23] or [35].

After the addition of two or more fixed-point numbers an overflow may occur. To avoid overflows or to restrict their occurrence to a very low rate, scaling of the SOS blocks is necessary. For a limitation of the state variables by two's complement characteristic overflows have to be checked only in front of the next multiplication. Controlling overflows by the saturation characteristic, the overflows have to be corrected direct behind shift-multipliers and adders or the overflow information has to be stored intermediately, e.g. by "guard bits",[2] to control the characteristic at the "overflow point".

In Fig. 3 all products are added in double precision. Quantization and overflow check is done at the internal check point in front of the state variables and at the block output. To reduce the hardware cost, the numerator multiplier b_0^* is realized as a shifter with a coefficient in po2-format.

As already mentioned, to avoid overflows, the filter transfer function has to be scaled. To describe the principle of scaling, we start with the transfer function $H(z)$ of one SOS.

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}}. \tag{1}$$

[2] Guard bits, in DSP-Processors, extend the accumulation register to the left in order to accumulate the over-flow information.

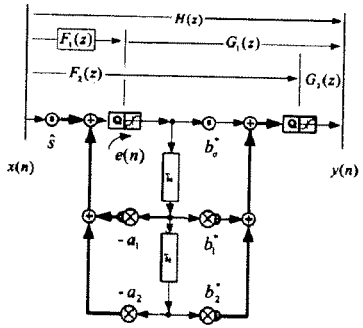


Fig. 3. Second order section (SOS) in df2-form.

Without considering the pre-multiplication $\hat{s}$, in Fig. 3 the transfer function $F(z)$ from the entrance to the scaling points is given by

$$F_1(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}}. \tag{2}$$

To bound the signal at the scaling point to one of the L_p-norms[3] [23], the maximal scaling factor s has to be chosen in accordance with the norm to

$$s = \|F_1(e^{j\Omega})\|_p^{-1}. \tag{3a}$$

The L_p-norms give an estimation for the maximum values of the state variables. Yet, scaling by these norms does not guarantee a prevention of overflows. To preclude any overflow at the state variables we have to scale by the absolute bound, the ℓ_1-norm [23].[4] Applications of the different norms will be compared in Section 6.

[3] L_p-norm of the frequency response:

$$\|X\|_p = \left(\frac{1}{2\pi}\int_0^{2\pi} |X(e^{j\Omega})|^p \, d\Omega\right)^{1/p};$$

$$\|X\|_\infty = \max_{0\le\Omega\le 2\pi}\{|X(e^{j\Omega})|\}.$$

[4] ℓ_1-norm of the impulse response:

$$\|X\|_{\ell_1} = \sum_{k=0}^{\infty} |x(k)|,$$

$$x(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\Omega}) e^{jk\Omega} \, d\Omega.$$

For a po2-scaling s has to be calculated to the next smaller po2-formatted number [5]

$$\hat{s} = \lfloor s \rfloor_{po2}. \tag{3b}$$

In order to scale the output of the SOS we have to consider the pre-scaling factor $\hat{s}$ and calculate the transfer function to the output, using the unscaled numerator coefficients,

$$F_2(z) = \hat{s}\,\frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{4}$$

and we get the maximal internal block gain factor for the numerator

$$g = \|F_2(e^{j\Omega})\|_p^{-1}. \tag{5a}$$

Taking care that the scaled numerator coefficients b_1^* and b_2^* do not exceed the coefficients number range c_{max} and the realization of the block gain by a power-of-two, we define the scaled internal block gain to

$$\hat{g} = \left\lfloor \min\left\{ g, \frac{c_{max}}{\max\{|b_1|, |b_2|\}} \right\} \right\rfloor_{po2}. \tag{5b}$$

With the scaled internal block gain we get the numerator coefficients

$$b_0^* = \hat{g} b_0;\ b_1^* = \hat{g} b_1;\ b_2^* = \hat{g} b_2 \tag{6}$$

for the scaled transfer function of the SOS in

$$\hat{H}(z) = s\,\frac{b_0^* + b_1^* z^{-1} + b_1^* z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}}. \tag{7}$$

This example shows a realization with minimal hardware efforts, suited to special custom designs. In general, scaling factor s and numerator coefficient b_0 are realized as entire or fractional multipliers.

Quantization is producing an additive error, $e(n)$ at the quantization points, shown in Fig. 3. For quantization by rounding all error sources are assumed to be uncorrelated. The mean power is determined to $\sigma^2 = Q^2/12$ for a fraction of w-bits and a quantization step $Q = 2^{-w}$.

[5] $\lfloor x \rfloor_{po2}$: next power-of-two number equal or smaller as x

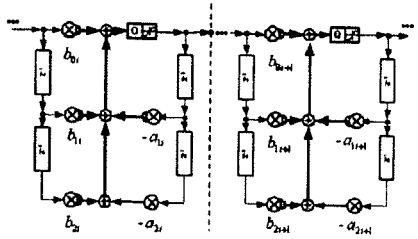


Fig. 4. Cascade with separate SOS in df1-form.

The quantization noise is transferred to the filter output by the scaled noise transfer function $\hat{G}_i(z)$. For a second order block, see Fig. 3, we get

$$\hat{G}_1(z) = \frac{b_0^* + b_1^* z^{-1} + b_2^* z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}},$$

$$\hat{G}_2(z) = 1. \tag{8}$$

The quantization noise power is proportional to the unfiltered noise of a single source. Therefore, we can compare different implementations with the relative noise power or the noise figure R_i, defined in [34]

$$R_i = \frac{\bar{N}_{si}}{\sigma_i^2} = \|\hat{G}_1\|_2^2 + 1. \tag{9}$$

The df1-form, see Fig. 4, can be scaled in the same way. For quantization and overflow check in front of the next multiplications, we have to consider only one scaling function $F_1(z)$ from the entrance to the block output

$$F_1(z) = H(z). \tag{10}$$

The noise transfer function is obviously

$$\hat{G}_1(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}}. \tag{11}$$

Different quantization characteristics and error models are evaluated in [9,12,28]. For other quantization characteristics (e.g. two's complement truncation) besides the uncorrelated noise also correlated errors can occur [9,28].

In the investigated examples in Section 6 we will restrict to rounding only.

Besides the two direct forms a large number of other structures are well-known. Depending on the applications and the available computing hardware, PC's DSP's, or special hardware with ASIC-, FPGA or full customer designs, these structures and their various realization forms may be more or less beneficial.

In the following sections we will concentrate on the so called "direct forms" and the state space structure for cascades of SOS, see Section 4. Direct forms are chosen due to the fact that they are very common and that they can be realized with a minimum number of noise sources. Yet, they may not be optimal according to their noise behavior. The state space structure shows excellent noise behavior for transfer functions with poles and zeros close to $z = \pm 1$ and, in general, show a better limit cycle behavior. The main disadvantage is the larger number of multiplications.

Further on, all investigations will be done for addition of the double precision products, rounding and two's complement overflow characteristic.

In Section 5, the mean noise power is used as an optimization criterion for the allocation of poles and zeros in the cascade.

4. Cascade of second order sections

The transfer function of a cascade of SOS is given by

$$H(z) = \prod_{i=1}^{L} \frac{b_{0i} + b_{1i} z^{-1} + b_{2i} z^{-2}}{1 + a_{1i} z^{-1} + a_{2i} z^{-2}}. \tag{12}$$

The cascades are realized without additional block scaling factors and the SOS are implemented by multiplications in the sfix-format, summation of double precision products and overflow points only in front of the next multiplication or state variable storage.

4.1. Direct forms

The transfer function of an SOS, (1), can be separated into the numerator and the denominator part and both parts can be calculated separately. This direct approach is called "Direct Form". Evaluating the numerator first, separate delay registers are necessary for the numerator and the denominator. Exchanging the two parts and calculating denominator before numerator leads to a minimum of delay elements, too.

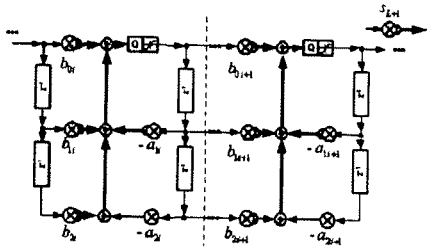


Fig. 5. Cascade with combined SOS in df1-form.

Two structures with a minimum number of state registers and multiplications were defined by Schüssler in [34] as the first and second canonical form. The second canonical form is the transposed one to the first. Oppenheim defined in [32] four direct forms, which all use a minimum number of multipliers, but only two of them are canonical with respect to the delay registers. In accordance with [11] we further on use the terms "df1-form", "df2-form" [6] and the hereto transposed forms, "df1t-form" and "df2t-form". [7]

For further investigations we will restrict to the direct structure of SOS in the df1-form (Fig. 4) and the df2-form (Fig. 6).

As already mentioned, the df1-form is not canonic with respect to the state registers. Yet, within the cascade the denominator delay elements of one block and the numerator delay elements of the next block keep always the same values. Therefore, these elements can be merged, as done in Fig. 5 and a total number of $n + 2$ registers is necessary for a system of order n.

It is obvious that all double precision products can be accumulated in one register and therefore the DSP's MAC-operation, [8] can be used very efficiently. Furthermore, the first scaling point is controlled by the first numerator, b_{11}^*, and thus no additional entrance multiplier is necessary. For scaling the filter output by a norm other than the L_∞-norm or due to hardware restrictions to the numerator coefficients an additional

[6] df2-form is identical to the second canonical form in [34].

[7] df2t-form is identical to the first canonical form in [34].

[8] MAC: multiplication and accumulation within one DSP processing loop.

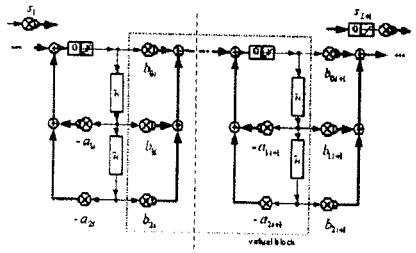


Fig. 6. Cascade with combined SOS in df2-form.

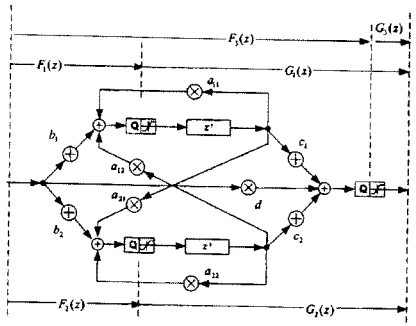


Fig. 7. Second order section in state space form.

scaling factor s_{l+1} is necessary at the output to verify the total gain of the transfer function. This additional multiplication also allows a comparison of different structures at the same output level. It is pointed out that no quantizer is meant behind this last scaling multiplier. The necessary word length reduction of the output variable can be assigned to a following D/A converter or to the input quantization of the next system component.

The cascade of SOS in the df2-form is shown in the block diagram in Fig. 6. The location of the adders in successive blocks allows an accumulation of the five double precision products within the cascade.

In contrast to the df1-form, the df2-form needs in any case an extra entrance-multiplier to scale the denominator of the first block. Again, the post-multiplier s_{l+1} is necessary to scale the overall transfer function.

A benefit of the df1-form and df2-form is that both structures need only one quantizer per block and therefore they have a minimum number of noise sources.

Moreover, the direct forms own the pleasant property that the numerator coefficients are independent of the denominator coefficients which simplifies a lot the coefficient rounding.

Filters realized in the direct forms, with poles located in the z-plane close to $z = \pm 1$ show frequently scaling problems due to the large resonance effect of the denominator transfer function. The necessary consequence is a very large quantization noise power. That means an extension of the state variable word length has to be considered in this case. Furthermore, the coefficient sensitivity for such realizations is critical.

4.2. State space forms

Robert and Mullis in [31] as well as Hwang in [21] proposed an optimized synthesis of digital filters, using the complete state matrix as free parameter space for the optimization of the scaling and noise behavior of the structure. The proposed structure, implemented as one full state matrix, size n by n, shows an excellent noise behavior, but, unfortunately the number of multiplications is growing extremely. To combine the low noise behavior with a moderate increase of multiplications Jackson et al. proposed in [24] a cascade of such Second Order State Space Sections (SSS). The block diagram of one SSS including quantizers and overflow checks is given in Fig. 7. The SSS are arranged in a cascade for filters of higher order. One SSS needs nine multipliers and, summing the double precision products, three quantizers are allocated. Consequently, we get three noise sources per block, which is an increase compared with the direct forms. However, this increase will be balanced in the critical cases by the optimal scaling and noise behavior of the state space structure.

The parameter of a single SSS can be calculated with the equations given in [23, p. 412].

We get for the transfer function of the SSS

$$H(z) = d + \underline{c}^t (z\mathbf{I} - \mathbf{A})^{-1} \underline{b} \tag{13}$$

using

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & b_{22} \end{pmatrix},$$

$$\underline{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \quad \underline{c}^t = (c_1\ c_2).$$

Corresponding to the direct forms, we introduce the scaling of a single SSS block, first.

The internal scaling functions, marked in Fig. 7, are given by

$$\underline{F}(z) = \begin{bmatrix} F_1(z) \\ F_2(z) \end{bmatrix} = (z\mathbf{I} - \mathbf{A})^{-1} \underline{b}. \tag{14}$$

To avoid overflows and to minimize the quantization noise the functions $F_1(z)$ and $F_2(z)$ have to be scaled by a transformation matrix $\mathbf{T}$ in a way that the chosen norms are observed at both scaling points.

According to [23] and with the conditions

$$\hat{\underline{F}}(z) = \mathbf{T}\underline{F}(z) \quad \text{and} \quad \|\hat{\underline{F}}\|_p = 1,$$

we get for the $\mathbf{T}$-Matrix

$$\mathbf{T} = \begin{bmatrix} \|F_1\|_p^{-1} & 0 \\ 0 & \|F_2\|_p^{-1} \end{bmatrix}. \tag{15}$$

Further the block output is scaled by the scaling function

$$F_3(z) = H(z) \tag{16}$$

and according to the condition

$$\|F_3\|_p = \|H\|_p = 1.$$

We get for the block gain factor

$$g = \frac{1}{\|H\|_p}. \tag{17}$$

The scaled transfer function of the state space structure is given by

$$\hat{H}(z) = \hat{d} + \hat{\underline{c}}^t (z\mathbf{I} - \hat{\mathbf{A}})^{-1} \hat{\underline{b}} \tag{18}$$

with the state matrices

$$\hat{\mathbf{A}} = \mathbf{T}\mathbf{A}\mathbf{T}^{-1}$$

$$\hat{\underline{b}} = \mathbf{T}\underline{b}, \quad \hat{\underline{c}}^t = g\underline{c}^t\mathbf{T}^{-1}, \quad \hat{d} = gd.$$

The scaled noise transfer functions can be calculated to

$$\hat{\underline{G}}(z) = \begin{bmatrix} \hat{G}_1(z) \\ \hat{G}_2(z) \end{bmatrix} = (z\mathbf{I} - \hat{\mathbf{A}}^t)^{-1} \hat{\underline{c}},$$

$$\hat{G}_3(z) = 1. \tag{19}$$

The parameters a_{11} and a_{22} of the state-matrix only depend on the real part of the pole positions. All other SSS parameters depend on the numerator and denominator coefficients of the transfer function. These parameters are not necessarily bounded to absolute values less than one. Specially for the case that poles and zeros of the SSS are both located in the pole-zero-diagram on a vertical line that means $\Re(z_p) = \Re(z_z) = const.$, the coefficients a_{12} or a_{21} possess points of discontinuity by variation of the pole and zero locations along these vertical lines. However, these problems can be solved by scaling the SSS according to the optimality criterion, given in [23]. The scaled SSS parameters can be calculated with (14), (15) for L_2-norm in a *closed form*. The parameters of the scaled transfer function are listed in the appendix.

The direct calculation of the scaled parameters may not only overcome numerical problems within the transformation of direct form coefficients into the SSS parameters, but also solve numerical problems in the implementation of the filter. For scaling by the L_∞- and ℓ_1-norm an additional post-scaling is necessary. It can be shown that for a single SSS the values of the scaled matrix $\mathbf{A}$ and the vector $\underline{b}$ are <1. For transfer functions, bounded to $|H(e^{j\Omega})| < 1$, the absolute values of the vector $\underline{c}^t$ are also <1 and in this case all coefficients can be realized by fractional multipliers.

To scale one block within a cascade of SSS, we calculate the norms of the transfer functions from the entrance of the filter to the internal scaling points and to the output of this block, build the transformation matrix (15) and the block gain factor (17) and recalculate the matrices of this block by (18). Scaling the complete cascade, may result in block parameters with absolute values > 1. Accepting to some degree, non-optimal scaling and though a higher quantization noise an optimization under the restriction of parameters with absolute values <1 is possible. A substitution of some multipliers by shifters may also be

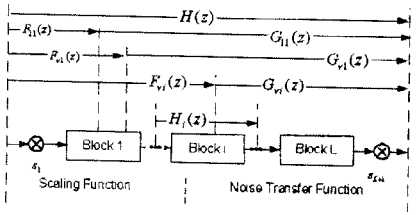


Fig. 8. Cascade scaling and noise transfer functions.

possible. In this case, like the direct forms, an underloaded scaling of the total filter has to be balanced by a post-multiplier s_{L+1}. However, we found that optimizing the allocation, of poles and zeros within the cascade mostly results in coefficient sets with absolute values < 1. Therefore, in general, a realization of the state space structure with fractional coefficients seems to be possible.

The noise behavior of filter realizations in an SSS cascade is shown below in Section 6 and the results are compared with cascade realization using SOS in direct forms.

5. Optimized allocation of poles and zeros

For the optimization of the cascade the scaling functions from the filter entrance to the individual scaling points within the cascade have to be calculated for the unscaled transfer function, see Fig. 8. With the transfer function $F_{Bvi}(z)$ from the block entrance to the ith scaling point within this block, we get for the total scaling function

$$F_{vi}(z) = F_{Bvi}(z) \prod_{k=1}^{i-1} H_k(z). \tag{20}$$

The $F_{Bvi}(z)$ are defined for the df1-form by (10), for the df2-form by (2) and (4) and for the sss-form by (14) and (16).

The scaling factors are calculated in a way that the scaled functions fulfill the used norm.

$$\hat{F}_{vi}(z) = c_r \frac{F_{vi}(z)}{\|F_{vi}\|_p} \leqslant 1. \tag{21}$$

The factor c_r is a correction factor for underloaded scaling due to special hardware restrictions for the numerator coefficients. The coefficients of the scaled functions are scaled in accordance to (6).

As a quality measure for the selected structure and the allocation of the poles and zeros in the cascade we use the mean or peak value of the power spectrum of the evoked quantization noise as an optimization criterion. The relative noise figure, defined in [34] and already introduced in (9) is calculated as the L_2-norm of the scaled power transfer function $\hat{G}_{vi}(z)$ from the noise sources to the filter output,

$$\hat{G}_{vi} = \hat{G}_{Bvi} \prod_{k=i+1}^{L} \hat{H}_k(z) \tag{22}$$

with the noise transfer function $G_{Bvi}(z)$ from the noise source to the output of the ith block.

For the noise figure we get

$$R_i = \frac{\bar{N}_{vi}}{\sigma_i^2} = \sum_{v,i} \|\hat{G}_{vi}\|_2^2. \tag{23}$$

Using the peak value of the noise power spectrum, we get the relative maximum PP_i:

$$PP_i = \max\left\{\frac{N_{vi}}{\sigma_i^2}\right\} = \max\left\{\sum_{v,i} |\hat{G}_{vi}|^2\right\}. \tag{24}$$

Both criteria will be used for the optimization and analysis of the allocations of poles and zeros in the cascade.

5.1. Simple heuristic methods

Although efficient optimization methods are available for the determination of the best allocation, in many cases a sufficiently good pairing and ordering of the poles and zeros can be done by simple heuristic methods. Avenhaus proposed in [1,2] a rule of thumb for the pairing. Further rules for ordering are given by Jackson [23].

Based on results received by the optimization procedures, described later on, three heuristic rules can be established:

Rule 1. *Starting with the poles closest to the unit circle, in a first step those poles and zeros are combined which are situated most closely to each other. Then, without taking into account already combined pairs, the procedure is repeated in the following steps.*

This first rule is based on the fact that poles close to the unit circle produce a high gain at the resonance frequency. This gain can be compensated best by the zeros closest to them.

Rule 2. *Having assigned pairs of poles and zeros, the pairs have to be allocated in the cascade in a way that both the corresponding numerator and denominator have an effect on one and the same quantization point.*

An allocation of poles and zeros according to this second rule causes the direct compensation of the pole resonance by the attenuation of the zeros. A smaller reduction of the signal in front of the quantizer means also a smaller gain of the corresponding noise source. Experiences have proved this course to be most effective.

This rule holds automatically for the df1-form, see Fig. 4. The corresponding zeros and poles are allocated each in one block and in front of the scaling and quantization point, marked by the quantizer symbol in the block diagrams. This rule is also true for the sss-form, as can be seen later.

As already mentioned in [9], this second rule is not directly observed for the df2-form, the second canonical form. Here numerator and denominator within one block have an effect on quantizers in different blocks, see Fig. 6.

To improve the performance for the df2-form, we can build virtual blocks out of the denominator of one block and the numerator of the block ahead. Allocating the selected pole and zero pairs in these virtual blocks both elements have again effect on one quantizer. Regarding the SOS the pairs of poles and zeros are in a "shifted allocation", as can be seen in Table 1. In many cases the shifted allocation leads to sufficiently good results, see curve "df2-up/sh" in Section 6, Fig. 12. Yet, the allocation of the first denominator may be still a problem. This will be discussed in the next rule.

Rule 3. *Poles are sorted according to their distance from the unit circle and consequently to their resonance behavior. Within the cascade the pairs of poles and zeros, assembled according to the 1st rule are either put in an ascendant ("up-ordering") or in a descendant order ("down-ordering" regarding their resonance behavior).*

Table 1
Pairing and Ordering by rules of thumb for an 8th order filter: pole no. 4 closest to the unit cycle

df1- /sss-form					
Up	1	2	3	4	Num
	1	2	3	4	Den
Down	4	3	2	1	Num
	4	3	2	1	Den
df2-form					
Up	2	3	4	1	Num
	1	2	3	4	Den
Down	4	3	2	1	Num
	1	4	3	2	Den

Both orderings lead to good results in general. However, for the df2-form the pairs have to be ordered into the virtual blocks. To avoid a large input attenuation by the pre-multiplier the most uncritical denominator has to be placed first.

The heuristic allocations are explained in Table 1 for an 8th order filter with four SOS. The numbering of the poles and zeros corresponds to ascending pole resonances.

In [22] Jackson analytically examines an adequate pairing and ordering of the cascade forms. The results are summarized in [23] and formulated as rules of thumb for minimizing the mean quantization noise by "up-ordering" or the peaks of the noise spectrum by "down-ordering". Further systematic investigations of a large number of examples show that these rules are met in many cases, but not in all. Examples are shown and discussed in Section 6, comments to Figs. 12 and 16.

5.2. Optimization methods

First optimization procedures were proposed among others in [2]. The results were close to an optimal solution, but inevitably *not* exact in the sense of the optimization criterion. The complete enumeration failed in most of the cases due to the large calculation effort. For a system of order 2^*L all possible $(L!)^2$ combinations of poles and zeros have to be verified. The computation time is approximately proportional to the number of combinations multiplied by the number

of blocks.

$$T_{en} \sim L(L1)^2. \tag{25}$$

The time consumption getting an exact solution can be extremely reduced by "Dynamic Programming", as e.g. proposed by Hwang in [20], Lüder et al. in [26] and in [7,9]. A precondition for "Dynamic Programming" is that the noise contribution of one step is only depending on this step itself and the preceding steps. This condition is fulfilled for the allocation problem. The procedure is evaluated in L steps, according to the L SOSs in the system. In a first step L^2 combinations of poles and zeros are checked as possible allocations for the first block. The noise contribution of this block, weighted by the power transfer function of the rest of poles and zeros, is calculated and intermediately stored. In the next step all possible $(\frac{L}{2})^2$ combinations of two numerator and denominator blocks with the 2^2 different permutations for each combination have to be proved. This is equivalent to the allocation of $(L-1)^2$ combinations of not yet used poles and zeros to each of the L^2 pre-calculated results of the first step. Now, only the best result for each new combination is stored intermediately for the next step. This procedure is continued till all poles and zeros are allocated in the last step. The allocation with the lowest noise figure (or with the lowest peak value) is the optimal result.

The computation time could be reduced to a value proportional to the total number of allocations times the number of blocks

$$T_{dp} \sim L \sum_{i=1}^{L} \binom{L}{i-1}^2 (L-i+1)^2. \tag{26}$$

The optimization procedure is part of a complete design program system, published in [10]. This system enables the optimization of filters in the two canonical forms for various hardware implementations including optimization of the coefficient word length for selective filters. A new and improved version of the optimization program, written in the MATLAB® Technical Language, is now available for cascade structures in the four direct forms and the state space form, including scaling for different hardware realizations of the basic blocks.

6. Examples and results

Optimal pairing and ordering of a cascade structure depends, as explained, on the implementation of the SOS in different basic forms, the special hardware realizations of the blocks, and the used scaling norm. The noise behavior of the chosen implementation is further depending on the pole and zero locations. To get an overview over the systems noise behavior the so-called "Allpass Test", introduced in [5] is performed for the cascade structures in df1-, df2- and sss-form. For this purpose, we design a normalized low pass with a pass-band edge at $\Omega_g = \frac{\pi}{2}$. Keeping the characteristic of the filter the edge frequency is moved over a region $\Omega_{min} \leqslant \Omega_g \leqslant \Omega_{max}$ by an allpass transformation of first order [34]. A schematic depiction of the transformation is given in Fig. 9.

In the following we show the noise figure R_t, (23), or the relative peak values PP_r, according to (24), for different implementations. Corresponding analysis can be done for band pass and stop band filters. Results for high-passes are equivalent to that of low-passes. The diagrams for the noise figures are mirrored to the frequency $\Omega = \pi/2$, compared to that of low-passes.

As an example we use a low-pass filter of 8th order with pass band and stop band ripples $\delta_P = 0.01$, $\delta_S = 0.001$ and edge frequencies $\Omega_P = 0.5$, $\Omega_S = 0.58$ for the normalized filter.

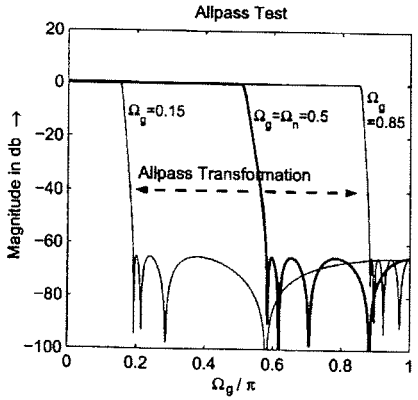


Fig. 9. Allpass test – modified frequency responses.

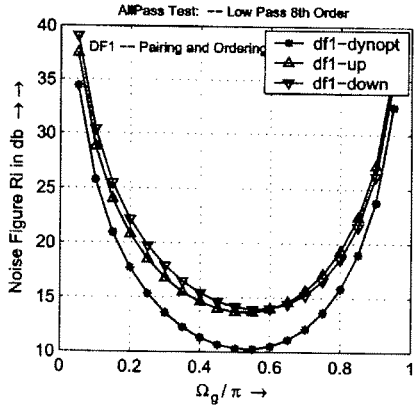


Fig. 10. Comparison of the Noise Figures for optimized ordering and the rules of thumb with df1-form.

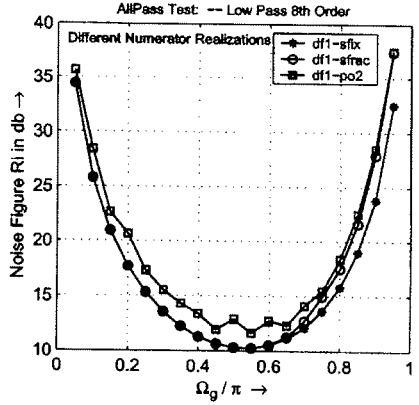


Fig. 11. Realization of numerator (df1-form) (a) by multipliers for non-fractional coefficients (sfix) (b) for fractional coefficients (sfrac) (c) with a shifter for the coefficient b_0^* (block gain by a power of two - p02).

In Fig. 10, the noise figures for a cascade of SOSs in df1-form are shown for optimal allocation and allocation by rules of thumb introduced in Section 5.1.

In any case the optimized ordering (by dynamic optimization) shows the lowest noise figures. But the results achieved with the simple rules (up-down-ordering) are only about 4 db worse. For the necessary additional word length Δw of the state variable[9] this means less than 1 bit (1 bit $\approx$ 6 db).

The influence of the numerator realizations (see Section 3) is given in Fig. 11, for optimized allocations of poles and zeros. Numerator realization with entire multipliers (sfix) or fractional multipliers (sfrac) show no difference in the noise figures for filters with low cutoff frequencies ($\Omega_g/\pi < 0.4$).

Due to the fact that in this region scaling leads to small block gains (realized by coefficient b_0^*) and therefore all numerator coefficients are fractional anyhow. For low-passes with a cutoff frequency > 0.6 pole positions are not so critical and therefore the block gain could be increased, and numerator coefficients of absolute values > 1 may occur. In this case an additional restriction to fractional coefficients may lead to underload and thus higher noise figures.

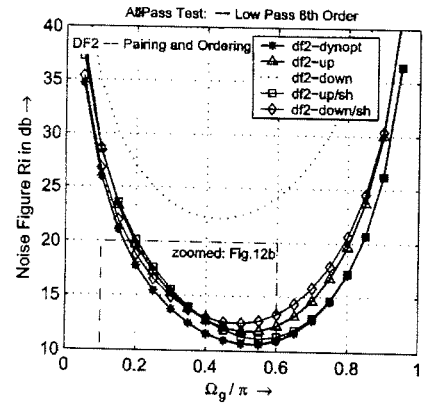
A similar effect can happen for a realization of the block gain factor (coefficient b_0^*) by a simple shift multiplier. Compared to the hardware realizations with entire multipliers (sfix) the increase of the noise figure is also small. Therefore, a realization on a low-cost fixed-point DSP without extra shifter in the ALU would show sufficient results for this example, as well.

The same filters are realized for cascades in the df2-form. Here, "up-ordering" without shift of the numerator blocks leads to reasonable results; see solid line -Δ- in Fig. 12.

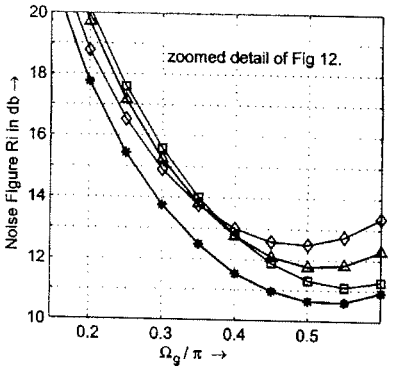
Using "up-ordering-shifted", according to rule 2, the results are close to the best solutions for wide band filters, $\Omega_g > 0.6\pi$. Allocations according to "down-ordering-shifted" show good results for cutoff frequencies $\Omega_g < 0.4\pi$ (small band), see zoomed details in Fig. 12a. Whereas "down-ordering" without shifting leads to a large extension of the output noise, see dotted line in Fig. 12a.

Comparing the df1-form (Fig. 10) and the df2-form (Fig. 12) in some respect we can see a contradictory

[9] $\Delta w = \left\lfloor \frac{1}{2} ld \frac{R_t}{R_e + 1} \right\rfloor$, R_t: noise figure of the filter; R_e: noise figure of the entrance quant.



(a) Overview



(b) Zoomed Detail

Fig. 12. Comparison of the noise figures for different ordering procedures for the df2-form: (a) Overview; (b) Zoomed detail.

behavior, using the simple rules of thumb. For both tests, not general, but for some regions of Ω_g, the rule "up-ordering" in conjunction with the mean power criterion [23] is hurt.

The optimal solutions are compared in Fig. 15 for both forms. The df1-form is showing slightly better results for wide band filters.

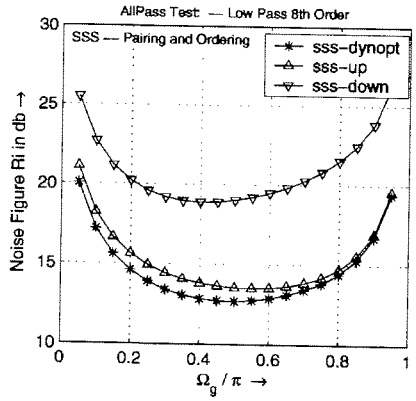


Fig. 13. Allpass test for the cascade in the sss-form.

For very narrow band filters with poles and zeros in the region of $z = \pm 1$ the quantization noise is growing for all filters in direct form. The noise behavior for filters with poles and zeros in these areas can be improved by the use of the sss-form, as pointed out in Section 4. The equivalent Allpass Test is performed again for the low-pass of 8th order.

In Fig. 13, we can see reduced noise figures for $\Omega_g < 0.2\pi$ and $> 0.8\pi$ (see the modified scale). Here an "up-ordering" of the poles is showing a noise behavior close to the optimized one. However, for the scaling by the L_∞-norm an increase of the noise figure in the fringe range cannot be overseen.

Scaling by L_2-norm, the optimized allocations show an increase of the noise figures in dependence of the bandwidth of the low-pass. For scaling by L_∞-norm output noise is larger in general, but there is an additional increase for poles close to $z = \pm 1$. Furthermore it is remarkable that scaling by the absolute ℓ_1-norm, output noise is only less than 6 db higher and therefore the internal word length have to be extended by one bit. It has to be mentioned that only the absolute norm guarantees a prevention of any overflows.

The optimized solutions for all three investigated SOS-structures are compared in Fig. 15 again. Due to the fact that both direct forms need only one noise source per block, these structures show best behavior

Table 2
Allocations for optimized solutions

Low-pass 8th order, $\delta_1 = 0.01$, $\delta_2 = 0.001$, L_∞-norm			
Ω_g/π	SOS in	Noise figure R_i	Allocation numerator denominator
0.2	df1-form	58,69	3 4 1 2 2 4 1 3
	df2-form	59,69	1 4 3 2 2 1 4 3
	sss-form	**28,67**	2 3 1 4 2 3 1 4
0.5	df1-form	**10,77**	3 4 1 2 2 4 1 3
	df2-form	11,61	3 2 4 1 1 3 2 4
	sss-form	18,45	2 3 1 4 2 3 1 4
0.8	df1-form	38,38	3 4 1 2 2 4 1 3
	df2-form	51,21	3 4 2 1 1 2 4 3
	sss-form	**27,46**	1 3 2 4 1 3 2 4

Edge Frequency $\Omega_g = \pi/2$
Total Scaling Factor = 0.025703

No:	b0i	b1i	b2i	a0i	a1i	a2i
(1)	1.0	1.865889	1.0	1.0	−0.552721	0.137683
(2)	1.0	1.205603	1.0	1.0	−0.295469	0.423848
(3)	1.0	0.720260	1.0	1.0	−0.049283	0.715144
(4)	1.0	0.519901	1.0	1.0	0.075909	0.915918

for low-pass filters with Ω_g in a middle range. At the edges, the balanced scaling and noise behavior of the sss-forms predominates and may be the best choice.

To give an overview over the pole and zero allocations for the optimized solutions some of the results are listed in Table 2.

In the last column the ordering of numerator and denominator blocks are listed, the numbering is in accordance with the primary numbers of the coefficient set, shown below. The often stressed nearly constant noise behavior for the sss-form is obtained only for scaling by the L_2-norm, which is not very useful in practice. The optimized allocations for the sss-form

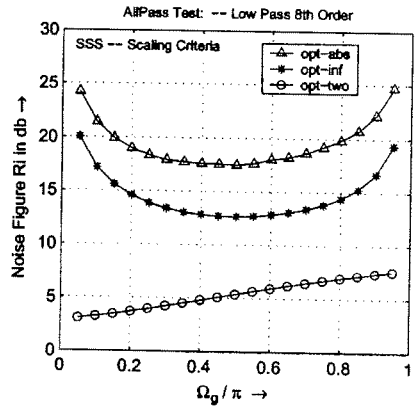


Fig. 14. Allpass Test for the sss-form, using different scaling norms: ℓ_1-norm (abs), L_∞-norm (inf), L_2-norm (two).

with different norms are shown in Fig. 14. For comparison, the total transfer function is adapted to the same overall gain for the different scaling norms.

In addition to the low-pass example many other selective filters, like band-passes and band-stops as well as some allpass filters and filters with group delay equalization have been tested. In most cases the optimal allocation was very close to that ones getting by the rules of thumb.

In special cases the peak value in the noise power spectrum may not be desirable. Therefore, the maximum of the power spectrum can be used as an alternative optimization criterion. Scaling can be done by any of the three norms. But, to compare the results with those, given in the literature, we use the L_2-norm for the following tests.

The peak values of the noise power spectrum are compared for the df1-form in Fig. 16 and for the df2-form in Fig. 17. In both cases good results can be achieved by optimization. For the df1-form the rule of thumb ("down-ordering") given in [23] leads to results very close to the optimum; whereas the df2-form "down-ordering" shows higher peak values. Here, again the new rule, according to Section 5, leads to much lower peak values PP_i.

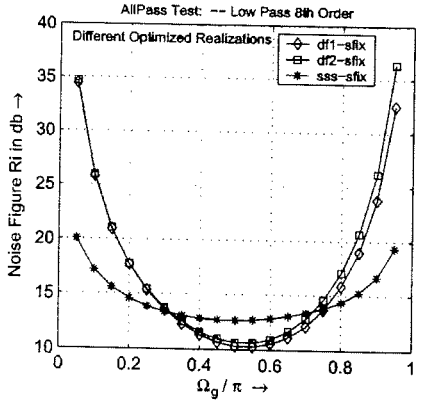


Fig. 15. Allpass test: optimized pairing and ordering for the df1-, df2- and sss-form.

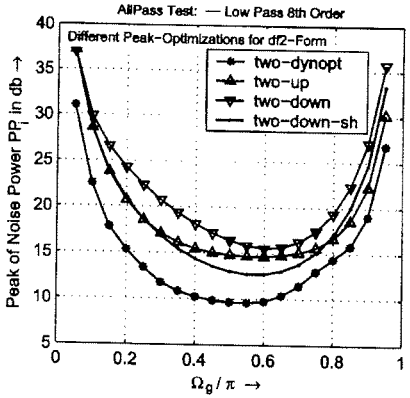


Fig. 17. Optimization of df2-form by peak noise power (scaled by L_2-norm).

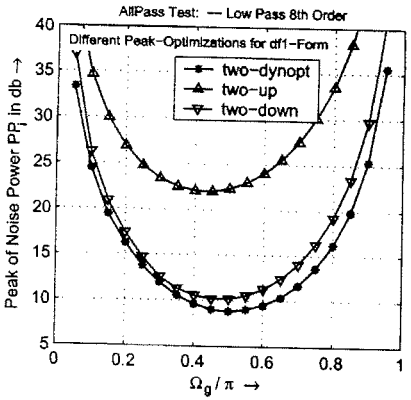


Fig. 16. Optimization of df1-form by peak noise power (scaled by L_2-norm).

7. Conclusion

Different aspects in designing noise optimized IIR filters have been reviewed. For the implementation of the transfer function some of the most popular cascade structures are further investigated due to the influence of special hardware implementations on their quantization noise behavior. Results are compared for cascades with SOS in df1-form, the canonic df2-form and the state space form.

Numerical problems by the transformation of the filter coefficients into state space parameters can be handled by a pre-scaling and noise optimization of each second order section in closed form. The calculation of the L_2-optimized SSS parameters is a convenient start for the final optimization of the cascade.

For all these different SOS an optimal allocation of the poles and zeros in the cascade can be determined by "Dynamic Programming". Yet it has to be considered that computing time is still growing nearly proportional to the power of the number of SOS's. Therefore, systems up to 16th order can be managed in a reasonable time (< 50 min) with a MATLAB® program running on a standard PC (1 GHz).

Based on the optimal solution, the well-known rules of thumb can be assessed and further improved. In many cases these heuristic allocations are very close to the best allocations getting by optimization. The noise figures may deviate less than 6 dB, which means the additional noise by the sub-optimal allocation can be compensated by one extra bit for the state

variables. The variations of the filter characteristic by the allpass transformation show that using only *one* of the rules may not always lead to sufficiently good results. Therefore, we propose to do always pairing according to the first rule and perform in parallel "up-ordering" and "down-ordering" for the df1-form, the df1t-form (not discussed in detail) and the sss-form. For the df2-form the shifted versions "up-ordering-shifted" and "down-ordering-shifted" should be proved in addition. The large number of performed tests have pointed out that in most cases a nearly optimal allocation can be achieved with these four tests only. The optimization program as well as the scaling procedures for special hardware, described above, are commercially available[10] and can be adapted for additional structures and new Hardware Designs.

Acknowledgements

The author would like to express his deep thanks to Professor Dr.-Ing. H.W. Schüssler for many discussions in the field of Quantized Digital Signal Processing and his support in restarting Design & Research Work in this field after a long time.

Appendix A. L_2-Scaled parameters for a second order section in scale space form

A.1. Transfer function

$$H(z) = (\beta_0 + \beta_1 z^{-1} + \beta_2 z^{-2})/(1 + \alpha_1 z^{-1} + \alpha_2 z^{-2})$$

and $\gamma_1 = \beta_1 - \beta_0\alpha_1$, $\gamma_2 = \beta_2 - \beta_0\alpha_2$

$$H(z) = \underline{d} + \underline{c}^t(z\mathbf{I} - \mathbf{A})^{-1} \text{ and}$$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \underline{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad \underline{c}^t = [c_1\ c_2], \quad \underline{d} = \beta_0.$$

[10] "Design and Optimization Procedures for SOS Digital Filters", IB Dehner*** qDSP Design & Application, 91058 Erlangen, Buckenhofer Weg 52, Germany.

A.2. Unscaled state space parameters

$$a_{11} = a_{22} = -\alpha_1/2,$$

$$a_{12} = \frac{(1+\gamma_2)(K_1+K_2)}{\gamma_1^2}$$

with

$$K_1 = \gamma_2 - \tfrac{1}{2}\alpha_1\gamma_1,$$

$$K_2 = (\gamma_2^2 - \gamma_1\gamma_2\alpha_1 + \gamma_1^2\alpha_2)^{1/2}$$

$$a_{21} = \frac{K_1 - K_2}{1+\gamma_2}$$

$$b_1 = \frac{1}{2}(1+\gamma_2), \quad b_2 = \frac{1}{2}\gamma_1$$

$$c_1 = \frac{\gamma_1}{1+\gamma_2}, \quad c_2 = 1.$$

A.3. L_2-scaled state space parameter

Substitutions:

$$N_{R+} = \left(1 + \left(\frac{K_1+K_2}{\gamma_1} + \frac{\alpha_1}{2}\right)^2\right)(1+\alpha_2) - \left(2\alpha_1\frac{K_1+K_2}{\gamma_1} + \alpha_1^2\right),$$

$$N_{R0} = \left(1 + \frac{\alpha_1^2}{4}\right)(1+\alpha_2) - \alpha_1^2,$$

$$N_{R-} = \left(1 + \left(\frac{K_1-K_2}{\gamma_1} + \frac{\alpha_1}{2}\right)^2\right)(1+\alpha_2) - \left(2\alpha_1\frac{K_1-K_2}{\gamma_1} + \alpha_1^2\right),$$

$$D_R = (1-\alpha_2)((1+\alpha_2)^2 - \alpha_1^2).$$

Scaled parameters:

$$\tilde{a}_{11} = a_{11},$$

$$\tilde{a}_{22} = a_{22},$$

$$\tilde{a}_{12} = \alpha_{12}\frac{\|F_2\|_2}{\|F_1\|_2} = \begin{cases} \dfrac{K_1+K_2}{|\gamma_1|}\sqrt{\dfrac{N_{R-}}{N_{R+}}} & \text{for } \gamma_1 \neq 0, \\ \sqrt{N_{R0}/(1+\alpha_2)} & \text{for } \gamma_1 = 0, \gamma_2 \geqslant 0 \\ \dfrac{\alpha_2 - \frac{\alpha_1^2}{4}}{\sqrt{N_{R0}/(1+\alpha_2)}} & \text{for } \gamma_1 = 0, \gamma_2 < 0, \end{cases}$$

$$\tilde{a}_{21} = a_{21}\frac{\|F_1\|_2}{\|F_2\|_2} = \begin{cases} \dfrac{K_1-K_2}{|\gamma_1|}\sqrt{\dfrac{N_{R+}}{N_{R-}}} & \text{for } \gamma_1 \neq 0, \\ -\dfrac{\alpha_2 - \frac{\alpha_1^2}{4}}{\sqrt{N_{R0}/(1+\alpha_2)}} & \text{for } \gamma_1 = 0, \gamma_2 \geqslant 0, \\ -\sqrt{N_{R0}/(1+\alpha_2)} & \text{for } \gamma_1 = 0, \gamma_2 < 0, \end{cases}$$

$$\tilde{b}_1 = \frac{1}{\|F_1\|_2}b_1 = \frac{1}{\|F_1\|_2}\frac{1+\gamma_2}{2} = \begin{cases} \sqrt{\dfrac{D_R}{N_{R1}}} & \text{for } \gamma_1 \neq 0, \\ 0 & \text{for } \gamma_1 = 0, \gamma_2 \geqslant 0, \\ \sqrt{\dfrac{D_R}{N_{R0}}} & \text{for } \gamma_1 = 0, \gamma_2 < 0. \end{cases}$$

$$\tilde{b}_2 = \frac{1}{\|F_2\|_2}b_2 = \frac{1}{\|F_2\|_2}\frac{\gamma_1}{2} = \begin{cases} \operatorname{sign}(\gamma_1)\sqrt{\dfrac{D_R}{N_R}} & \text{for } \gamma_1 \neq 0, \\ \sqrt{\dfrac{D_R}{N_{R0}}} & \text{for } \gamma_1 = 0, \gamma_2 > 0, \\ 0 & \text{for } \gamma_1 = 0, \gamma_2 \leqslant 0, \end{cases}$$

$$\tilde{c}_1 = \|F_1\|_2 \quad c_1 = \|F_1\|_2\frac{\gamma_1}{1+\gamma_2} = \begin{cases} \dfrac{\gamma_1}{2}\sqrt{\dfrac{N_{R+}}{D_R}} & \text{for } \gamma_1 \neq 0, \\ \gamma_2\sqrt{\dfrac{1+\alpha_2}{D_R}} & \text{for } \gamma_1 = 0, \gamma_2 > 0, \\ 0 & \text{for } \gamma_1 = 0, \gamma_2 \leqslant 0, \end{cases}$$

$$\tilde{c}_2 = \|F_2\|_2 \quad c_2 = \|F_2\|_2 = \begin{cases} \dfrac{|\gamma_1|}{2}\sqrt{\dfrac{D_R}{D_R}} & \text{for } \gamma_1 \neq 0, \\ 0 & \text{for } \gamma_1 = 0, \gamma_2 \geqslant 0, \\ -\gamma_2\sqrt{\dfrac{1+\alpha_2}{D_R}} & \text{for } \gamma_1 = 0, \gamma_2 < 0. \end{cases}$$

References

[1] E. Avenhaus, Zur Realisierung digitaler Filter mit günstigem Nutz-Störleistungsverhältnis, NTZ 23 (1970) 217–219.

[2] E. Avenhaus, Zum Entwurf digitaler Filter mit minimaler Speicherwortlänge für Koeffizienten und Zustandsgrössen, in: H.W. Schüssler (Ed.), Ausgewählte Arbeiten über Nachrichtensysteme, No. 13, Universität Erlangen-Nürnberg, 1971.

[3] E. Avenhaus, H.W. Schüssler, On the Approximation Problem in the Design of Digital Filters with Limited Wordlength, Archiv Elektronik Übertragungstechnik Electronics Communications 24 (1970) 571–572.

[4] M. Büttner, Untersuchungen über Grenzzyklen in digitalen Filtern, in: H.W. Schüssler (Ed.), Ausgewählte Arbeiten über Nachrichtensysteme, No. 27, Universität Erlangen-Nürnberg, 1977.

[5] R. Czarnach, Antwortstabile rekursive Digitalfilter in Festkomma-Arithmetik, in: H.W. Schüssler (Ed.), Ausgewählte Arbeiten über Nachrichtensysteme, No. 58, Universität Erlangen-Nürnberg, 1984.

[6] G. Dehner, On the Design of Digital Cauer Filters with Coefficients of Limited Wordlength, Archiv Elektronik Übertragungstechnik Electronics Communications 29 (1975) 165–168.

[7] G. Dehner, A Contribution to the Optimization of Roundoff-Noise in Recursive Digital Filters, Archiv Elektronik Übertragungstechnik Electronics Communications 29 (1975) 505–510.

[8] G. Dehner, On the Noise Behaviour of Digital Filters of Second Order, Archiv Elektronik Übertragungstechnik Electronics Communications 30 (1976) 394–398.

[9] G.F. Dehner, Ein Beitrag zum rechnergestützten Entwurf rekursiver digitaler Filter minimalen Aufwands, in: H.W. Schüssler (Ed.), Ausgewählte Arbeiten über Nachrichtensysteme, No. 23, Universität Erlangen-Nürnberg, 1976.

[10] G.F. Dehner, Program for the design of recursive digital filters, in: Digital Signal Processing Committee IEEE ASSP (Ed.), Programs for Digital Signal Processing, IEEE Press, New York, 1979, (Chapter 6.1).

[11] Documentation of the MATLAB simulation system including signal processing and filter design toolbox, The MathWorks Inc. Natick MA, USA, 1984–2002.

[12] B. Eckhardt, Untersuchungen des Multipliziererfehlers in digitalen Filtern, in: H.W. Schüssler (Ed.), Ausgewählte Arbeiten über Nachrichtensysteme, No. 21, Universität Erlangen-Nürnberg, 1975.

[13] B. Eckhardt, On the Roundoff Error of a Multiplier, Archiv Elektronik Übertragungstechnik Electronics Communications 29 (1975) 162–164.

[14] B. Eckhardt, H.W. Schüssler, On the quantization error of a rounded sum of rounded products, Archiv Elektronik Übertragungstechnik Electronics Communications 29 (1975) 308–311.

[15] B. Eckhardt, H.W. Schüssler, On the quantization error of a multiplier, IEEE Proceedings ISCAS, 1976, pp. 634–637.

[16] B. Eckhardt, W. Winkelnkemper, Entwurf und Aufbau eines flexiblen rekursiven digitalen Filters, Tagungsband zur NTG-Tagung Signalverarbeitung, Erlangen, 1973, pp. 104–111.

[17] S. Hentschke, Untersuchungen und Entwürfe von hochintegrierten Echokompensionsverfahren zur Duplexübertragung, Frequenz 36, 1982, pp. 302ff.

[18] U. Heute, Über Realisierungsprobleme bei nichtrekursiven Digitalfiltern, in: H.W. Schüssler (Ed.), Ausgewählte Arbeiten über Nachrichtensysteme, No. 20, Universität Erlangen-Nürnberg, 1975.

[19] U. Heute, A Subroutine for finite wordlength FIR filter design, in: Digital Signal Processing Committee IEEE ASSP (Ed.), Programs for Digital Signal Processing, IEEE Press, New York, 1979.

[20] S.Y. Hwang, On optimization of cascade fixed-point filters, IEEE Trans. CS-21 (1974) 163–164.

[21] S.Y. Hwang, Roundoff noise in state-space digital filtering: a general analysis, IEEE Trans. ASSP-24 (1976) 256–262.

[22] L.B. Jackson, Roundoff-noise analysis for fixed-point digital filters realized in cascade or parallel form, IEEE Trans. AU18 (2) (1970) 107–122.

[23] L.B. Jackson, Digital Filters and Signal Processing, 3rd Edition with MATLAB Exercises. Kluwer Academic Publishers, Boston, 1996.

[24] L.B. Jackson, A.G. Lindgreen, Y. Kim, Optimal synthesis of second-order state space structures for digital filters, IEEE Trans. CAS-26 (3) (1979) 149–153.

[25] A. Kanbach, A. Koerber, ISDN Technik, Hüthig Verlag, Heidelberg, 1990.

[26] L. Lüder, H. Hug, W. Wolf, Minimization of round-off noise in digital filters by dynamic programming, Frequenz 29 (1975) 211–214.

[27] J.H. McClellan, C.S. Burrus, A.V. Oppenheim, T.W. Parks, R.W. Schafer, H.W. Schüssler, Computer-based exercises for signal processing, using MATLAB 5, Matlab Curriculum Series, Prentice-Hall, Englewood Cliffs, NJ, 1998.

[28] R. Meyer, Zur Realisierung von digitalen Systemen mit Festkomma-Signalprozessoren, in: H.W. Schüssler (Ed.), Ausgewählte Arbeiten über Nachrichtensysteme, No. 86, Universität Erlangen-Nürnberg, 1994.

[29] U. Meyer-Baese, Digital Signalprocessing with Field Programmable Gate Array, Springer, Berlin, 2001.

[30] S.K. Mitra, J.F. Kaiser (Eds.), Handbook for Digital Signal Processing. Wiley, New York, 1993.

[31] C.T. Mullis, R.A. Roberts, Synthesis of minimum roundoff noise fixed point digital filters, IEEE Trans. CAS-23 (9) (1976) 551–562.

[32] A.V. Oppenheim, R.W. Schafer, Digital Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1975.

[33] J.B.H. Peek, Communication aspects of the compact disc digital audio system, IEEE Commun. Mag. 23 (2) (1985) 7–20.

[34] H.W. Schüssler, Digitale Systeme zur Signalverarbeitung, Springer, Berlin, 1973.

[35] H.W. Schüssler, Digitale Signalverarbeitung 1, Analyse diskreter Signale und Systeme, Vierte Auflage, Springer, Berlin, 1994.

[36] K. Schwarz, Ein Beitrag zur Realisierung digitaler Filter, in: H.W. Schüssler (Ed.), Ausgewählte Arbeiten über Nachrichtensysteme, No. 87, Universität Erlangen-Nürnberg, 1994.

Optimal and Suboptimal Error Spectrum Shaping for Cascade-Form Digital Filters

WILLIAM E. HIGGINS, STUDENT MEMBER, IEEE, AND DAVID C. MUNSON, JR., MEMBER, IEEE

Abstract—Error spectrum shaping (ESS) can significantly reduce finite-wordlength error in recursive fixed-point digital filters. Most past work on ESS has focused on second-order filters. In this paper, expressions for the ESS coefficients minimizing the roundoff error are derived for high-order filters composed of cascaded second-order sections. A comparison of the noise gains is made between optimal and suboptimal ESS structures and the section-optimal structure; the ESS structures perform well in comparison. A heuristic section-ordering strategy, based on ideas underlying ESS, is proposed and deemed useful, particularly for elliptic filters.

I. Introduction

ERROR SPECTRUM shaping (ESS) is a technique that can help to alleviate finite-wordlength effects in digital systems. Spang and Schultheiss [1] first used ESS to reduce quantizing noise in a PCM rounding quantizer. Later work demonstrated how ESS could improve the accuracy of ΔM and DPCM systems [2], [3] and reduce roundoff error in DPCM digital filters [4].

ESS has also been successfully applied to recursive fixed-point digital filters [5]–[21]. (For the rest of this paper, we assume that a "digital filter" is a recursive fixed-point digital filter.) For digital filters using rounding quantizers, Claasen and Kristiansson [5] demonstrated that ESS could limit adder-overflow oscillations and Munson and Liu [17] derived bounds showing that ESS could reduce the maximum amplitude of zero-input limit cycles. For digital filters using magnitude-truncation quantizers, Chang [16] showed that ESS could suppress zero-input limit cycles.

Recently, many authors have used ESS to reduce roundoff error in direct-form implementations of digital filters [6]–[15], [17]–[21]. They have shown that ESS can greatly reduce roundoff error—particularly in narrow-band filters, when roundoff error can be especially large—without overly complicating the original filter.

Other authors have successfully reduced finite-wordlength effects in digital filters by using linear state-space (LSS) concepts [22]–[28]. (The normal form [29]–[32] is closely allied to the LSS structure. Incidentally, Barnes [32] has applied ESS to normal-form digital filters to reduce roundoff error.) Past research has shown that filters implemented with LSS ideas are free of overflow limit cycles [26], have low coefficients sensitivities [27], and exhibit small roundoff error [23]–[25], [27], [28]. To ease the hardware requirements, Mullis and Roberts [23], [28] devised the block-optimal structure and Jackson *et al.* [27] the section-optimal structure; both of these structures implement digital filters with LSS-form second-order sections. The block-optimal structure has the lowest roundoff error among structures implemented with LSS-form second-order sections (ignoring section ordering), but it requires a complicated synthesizing procedure. The section-optimal structure, however, is relatively easy to synthesize and reduces roundoff error almost as well as the block-optimal structure.

Nearly all previous ESS work has considered just second-order digital filters. Chang, however, has published two concise papers demonstrating the effectiveness of ESS for higher order filters and comparing noise reduction with the LSS approach [10], [12]. In this paper we expand on Chang's work. We consider high-order digital filters implemented as a cascade of second-order sections and examine how ESS can be applied most advantageously to reduce roundoff error.

Specifically, Section II gives a tutorial on the fundamentals of ESS.

Section III discusses the cascade-form ESS structure, including assumptions on the roundoff error model and scaling.

Section IV derives explicit expressions for the ESS coefficients that minimize the roundoff noise of the cascade-form ESS structure. (For parallel-form filters one need only optimize second-order sections individually. This problem has been solved in [21].)

Section V presents a comparison of the noise gains for several optimal and suboptimal ESS structures, the canonical structure, and the section-optimal structure. This comparison, made with several filter examples including five standard ones [10], [12], [27], reveals that the ESS structures significantly reduce roundoff error and compare favorably to the section-optimal structure. In particular, a simple ESS structure, which adds only an extra subtract and delay register to the original implementation, reduces roundoff error more than the section-optimal structure for all of the low-pass filters. A conclusion of this section is that ESS can significantly reduce roundoff error without overly complicating the hardware for narrow-band low-pass, high-pass, and bandpass cascade-form filters.

Manuscript received September 24, 1982; revised June 6, 1983. This work was supported by the Joint Services Electronics Program under Contract N00014-79-C-0424.

The authors are with the Coordinated Science Laboratory and the Department of Electrical Engineering, University of Illinois, Urbana, IL 61801.

0098-4094/84/0500-0429$01.00 ©1984 IEEE

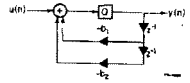


Fig. 1. A Second-order all-pole digital filter.

Section VI examines the section-ordering problem for digital filters that use ESS. A graphical example of a particular filter illustrates why appropriate section ordering is critical to the effectiveness of ESS; and a proposed heuristic section-ordering strategy for filters employing ESS proves useful, particularly for elliptic filters.

II. ESS Fundamentals

In order to set the stage for the subject of high-order cascade-form ESS filters, it is instructive to first consider just a simple second-order filter. Fig. 1 shows a second-order all-pole digital filter with transfer function

$$G(z) = \frac{1}{1 + b_1 z^{-1} + b_2 z^{-2}}.$$

We assume sign-magnitude arithmetic with B bits plus sign, a double-precision accumulator, and a rounding quantizer after the accumulator. Scaling will be disregarded for the present.

The quantizer makes the filter in Fig. 1 nonlinear. So, following standard practice, we model the nonlinear filter as a linear system by replacing the quantizer with an independent, additive, white error source having variance $\sigma_0^2 = 2^{-2B}/12$. It follows that the variance, σ^2, of the quantization error at the filter output is

$$\sigma^2 = \sigma_0^2 \|\mathscr{E}(z)\|_2^2 \tag{1}$$

where

$$\|\mathscr{E}(z)\|_2 = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} |\mathscr{E}(e^{j\lambda})|^2 d\lambda\right]^{1/2} \tag{2}$$

and $\mathscr{E}$ is the transfer function from the quantizer (error source) to the output. For the filter in Fig. 1, $\mathscr{E}(z) = G(z)$. The power spectral density of the output error is $\sigma_0^2|\mathscr{E}(e^{j\lambda})|^2$; hence, we will call $|\mathscr{E}(e^{j\lambda})|$ the *error spectrum*.

As an example, let $b_1 = -1.85$ and $b_2 = 0.9025$ in Fig. 1. The poles of $\mathscr{E}(z) = G(z)$ then lie in the z-plane at an angle $\theta = 0.23$ from the real axis and at a distance $\delta = 0.05$ from the unit circle. The dashed line in Fig. 2 shows the error spectrum, and as expected, since the poles are near the unit circle ($\delta = 0.05$) there is a large peak near $\lambda = \theta$. Thus the mean-squared error given by (1) will be large.

A standard method for reducing roundoff noise is to alter the filter structure so that the transfer function from the error source to the filter output is better behaved. Of course, this must be accomplished without changing the transfer function from the filter input to the filter output. Error spectrum shaping is a technique for doing this directly by feeding back (and possibly forward) the B-bit error sequence introduced at the quantizer.

Fig. 3 shows an ESS implementation of Fig. 1 where conceptually we again replace the quantizer with an additive error source, $e(n) = Q[w(n)] - w(n)$. Notice that by modeling the system in this fashion, the filter effectively has two independent inputs, $u(n)$ and $e(n)$. Hence, $y(n)$ is by superposition a sum of the responses to these two inputs. Therefore, the ESS circuitry does not affect the filter's input–output characteristic, $G(z)$.

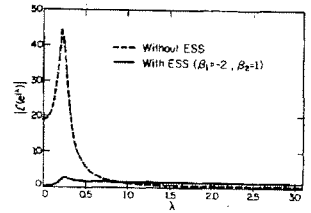


Fig. 2. Error spectra for two forms of a second-order digital filter ($b_1 = -1.85$, $b_2 = 0.9025$).

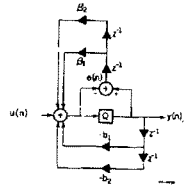


Fig. 3. A second-order digital filter with ESS.

TABLE I
Noise Gains for a Second-Order All-Pole Digital Filter with and Without ESS

Filter Structure	β_1	β_2	σ^2/σ_0^2 (dB)
Canonical	0	0	19.96
Optimal ESS	-1.85	.9025	0.0
Second-Order Suboptimal ESS	-2	1	1.30
First-Order Suboptimal ESS	-1	0	7.38

For the ESS filter in Fig. 3, the output roundoff error is again given by (1), but the transfer function from the error source to the filter output is now

$$\mathscr{E}(z) = \frac{1 + \beta_1 z^{-1} + \beta_2 z^{-2}}{1 + b_1 z^{-1} + b_2 z^{-2}}.$$

The error spectrum can therefore be written as

$$\mathscr{E}(e^{j\lambda}) = \left(1 + \beta_1 e^{-j\lambda} + \beta_2 e^{-j2\lambda}\right) G(e^{j\lambda}). \tag{3}$$

The coefficients cannot be chosen to make (3) arbitrarily small over the entire frequency band, but for many ESS filters the error spectrum is given by a formula similar to (3), where G is either an approximate low or high-pass characteristic. Thus the ESS coefficients can be chosen to provide attenuation of (3) within the passband of G, while G itself provides attenuation within the stopband.

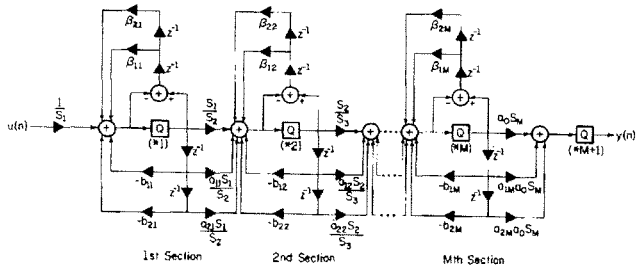


Fig. 4. Cascade-form realization of an Nth-order filter where $N = 2M$.

Returning to the example with $b_1 = -1.85$ and $b_2 = 0.9025$, it is possible to significantly reduce the peak at $\lambda \approx \theta$ in Fig. 3, while amplifying the noise for larger λ only moderately. Choosing $\beta_1 = -2$ and $\beta_2 = 1$ in Fig. 3 introduces a double zero in the error spectrum, (3), at $\lambda = 0$ and therefore reduces the nearby peak at $\lambda \approx \theta = 0.23$. Evaluating (3) at $\lambda = \pi$ shows that the spectrum will be no more than quadrupled elsewhere. The resulting error spectrum is shown as the solid curve in Fig. 2.

Table I lists the noise gains, σ^2/σ_0^2, for the filter without ESS and for three different forms of the filter with ESS. The optimal ESS structure minimizes the noise gain [21], but this form of ESS requires additional multiplications and is actually a disguised implementation of double-precision arithmetic [14], [21], [33]. The suboptimal ESS structure, chosen above, reduces the noise gain to a value near the optimal, and saves 3 bits per register (increasing the register length by 1 bit reduces the noise gain by approximately 6 dB). Even the simple first-order ESS structure, which adds only an additional subtract and delay register, reduces the noise gain considerably.

III. The Cascade-Form ESS Filter Structure

We now consider an Nth-order digital filter implemented in cascade-form as

$$H(z) = a_0 \prod_{i=1}^{M} H_i(z) \tag{4}$$

with

$$H_i(z) = \frac{1 + a_{1i}z^{-1} + a_{2i}z^{-2}}{1 + b_{1i}z^{-1} + b_{2i}z^{-2}}$$

where $H_i(z)$ is the transfer function of the ith section; $M = N/2$; and N is assumed even for simplicity. The following assumptions are made on the implementation of the filter:

(a) each section implemented in direct-form 2 (DF2) [34],
(b) L_2 scaling used,
(c) registers are B bits plus sign,
(d) sign-magnitude arithmetic used,
(e) each adder is double precision,
(f) rounding[1] is performed after each adder,
(g) all sources of roundoff error are independent, white, and have variance $\sigma_0^2 = 2^{-2B}/12$,
(h) second-order roundoff effects (magnitude proportional to 2^{-2B}) are ignored,
(i) $H(z)$ is stable.

We will study an implementation of (4) with second-order ESS applied to each $H_i(z)$. In general, the output roundoff noise can be further reduced by using higher order ESS (for an error source entering an Lth-order transfer function, the optimal ESS network is also Lth-order and essentially gives a double precision implementation [33]), but this prohibitively complicates the implementation. Second-order ESS networks that utilize both feedforward and feedback have been considered [10], [21]. We have found, though, that the inclusion of feedforward does not appreciably reduce the roundoff noise for most filters [21]. In addition, feedforward alone is generally ineffective; e.g., optimal second-order error feedforward can cancel only the first three terms in the unit pulse response of the error spectrum [21]. Therefore, we do not use feedforward in our implementation of (4).

Fig. 4 depicts our cascade-form ESS implementation, where the β_{ki} are the ESS feedback coefficients. This structure employs the same scaling policy as Jackson's structure 1D in [36]. Hence, we implement

$$H(z) = \frac{1}{S_1} \prod_{i=1}^{M} \frac{\hat{a}_{0i} + \hat{a}_{1i}z^{-1} + \hat{a}_{2i}z^{-2}}{1 + b_{1i}z^{-1} + b_{2i}z^{-2}}$$

where,

(a) $\hat{a}_{ki} = \dfrac{S_i a_{ki}}{S_{i+1}}, \qquad k = 0,1,2$

(b) $S_i = \left\| \dfrac{1}{1 + b_{1i}z^{-1} + b_{2i}z^{-2}} \prod_{j=1}^{i-1} H_j(z) \right\|_2, \qquad i = 1, \cdots, M \quad \left(\prod_{j=1}^{0} = 1\right)$

[1] ESS can also be applied successfully to systems using sign-magnitude truncation, but in this case the correlation between the quantization error samples [35] makes a precise analysis difficult.

and

$$S_{M+1} = \frac{1}{a_0}.$$

In the following we will devote considerable attention to simple cases where the β_{ki} in Fig. 4 are powers of 2. However, for the general case when the β_{ki} are represented in B bits plus sign, the ESS multiplier outputs will require $(2B+1)$-bit representations with the least significant bits representing 2^{-3B}. These outputs must be rounded before accumulation with other adder inputs, resulting in a small error bounded in magnitude by $2^{-2B}/2$ at each adder. As assumed previously (see (h)), this error will be neglected; hence, the associated quantization is not shown in Fig. 4. Analyses including this error term are given in [21] for second-order filters. It should also be mentioned that we assume that the full B-bit roundoff error is fed back from the quantizers, even though most of the benefit from ESS can generally be gained by using just the higher order error bits [17].

IV. Optimization of ESS Coefficients

The expression for the mean-squared roundoff noise at the filter output for Fig. 4 is

$$\sigma^2 = \sigma_0^2 \left[1 + a_0^2 \sum_{i=1}^{M} S_i^2 \| F_i(z) G_i(z) \|_2^2 \right] \tag{5}$$

where

$$F_i(z) = 1 + \beta_{1i}z^{-1} + \beta_{2i}z^{-2}$$

and

$$G_i(z) = \prod_{j=i}^{M} H_j(z). \tag{6}$$

$F_i(z)H_i(z)$ is the unscaled transfer function from the first error source in the ith section (point (*i)) to the output of that section, and $F_i(z)G_i(z)$ is the unscaled transfer function from the first error source in the ith section to the filter output. The factors a_0 and S_i in (5) account for the scaling used in Fig. 4.

We wish to minimize σ^2 in (5) with respect to the error feedback coefficients β_{1i} and β_{2i}. Excluding the superfluous terms in (5) that do not enter into the optimization, a restatement of our problem is

$$\min_{\substack{\beta_{1i}, \beta_{2i} \\ i = 1, \cdots, M}} \left[\sum_{i=1}^{M} S_i^2 \| F_i(z) G_i(z) \|_2^2 \right]. \tag{7}$$

Since each term in (7) is a function of a separate pair of ESS coefficients, the single optimization problem can be broken up into M separate optimization problems; i.e., each set of error feedback gains can be optimized independently from the other sets. Each of the separate optimization problems has the form

$$\min_{\beta_{1i}, \beta_{2i}} \left\| \left(1 + \beta_{1i}z^{-1} + \beta_{2i}z^{-2}\right) \prod_{j=i}^{M} H_j(z) \right\|_2^2 \tag{8}$$

where F_i and G_i have been substituted for from (6). Equation (8) is equivalent to minimizing

$$I = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[1 + \beta_{1i}e^{-j\lambda} + \beta_{2i}e^{-j2\lambda}\right] \left[1 + \beta_{1i}e^{j\lambda} + \beta_{2i}e^{j2\lambda}\right] Q(\lambda)\, d\lambda$$

where

$$Q(\lambda) = \prod_{k=i}^{M} H_k(e^{j\lambda}) H_k(e^{-j\lambda}).$$

With algebra, the expression for I can be simplified to

$$I = J_1\beta_{1i}(1 + \beta_{2i}) + J_2\beta_{2i} + \frac{J_3}{2}\left(1 + \beta_{1i}^2 + \beta_{2i}^2\right) \tag{9}$$

where,

$$J_1 = \frac{1}{\pi} \int_{-\pi}^{\pi} Q(\lambda) \cos\lambda\, d\lambda$$

$$J_2 = \frac{1}{\pi} \int_{-\pi}^{\pi} Q(\lambda) \cos(2\lambda)\, d\lambda$$

$$J_3 = \frac{1}{\pi} \int_{-\pi}^{\pi} Q(\lambda)\, d\lambda.$$

First-order ESS, with $\beta_{2i} = 0$, merits attention for implementations requiring simple hardware. In this case, (9) reduces to

$$I = J_1\beta_{1i} + \frac{J_3}{2}\left(1 + \beta_{1i}^2\right).$$

The optimal first-order ESS feedback coefficient for the ith section is obtained by solving

$$\frac{\partial I}{\partial \beta_{1i}} = 0 = J_1 + J_3\beta_{1i}$$

giving

$$\beta_{1i} = \frac{-J_1}{J_3}. \tag{10}$$

The second derivative is

$$\frac{\partial I^2}{\partial \beta_{1i}^2} = J_3 > 0$$

so that the β_{1i}, given by (10), attain the minimum round-off noise for first-order ESS.

The optimal coefficients for second-order ESS are obtained by differentiating (9), giving

$$\frac{\partial I}{\partial \beta_{1i}} = 0 = J_1(1 + \beta_{2i}) + J_3\beta_{1i}$$

$$\frac{\partial I}{\partial \beta_{2i}} = 0 = J_1\beta_{1i} + J_2 + J_3\beta_{2i}.$$

Solving for β_{1i} and β_{2i} gives

$$\beta_{1i} = \frac{J_1J_2 - J_1J_3}{J_3^2 - J_1^2} \qquad \beta_{2i} = \frac{J_1^2 - J_2J_3}{J_3^2 - J_1^2}. \tag{11}$$

Furthermore, the determinant of the matrix with elements

$$d_{kl} = \frac{\partial}{\partial \beta_{ki}} \frac{\partial I}{\partial \beta_{li}} \text{ is } J_3^2 - J_1^2$$

TABLE II
σ^2/σ_0^2 IN DECIBELS FOR THE CASCADE-FORM DESIGNS

FILTER / STRUCTURE	BUTTERWORTH LPF, N=6 $\lambda_c = \pi/2$	BUTTERWORTH LPF, N=6 $\lambda_c = \pi/20$	CHEB-I LPF, N=10 $\lambda_c = \pi/4$	ELLIPTIC LPF, N=10 $\lambda_c = \pi/4$	ELLIPTIC BPF, N=8 BWIDTH=$\frac{\pi}{25}$ $\lambda_{cent} = \frac{\pi}{10}$	CHEB-II BRF, N=6 BWIDTH=$\frac{\pi}{10}$ $\lambda_{cent} = \frac{\pi}{6}$
SECTION-OPTIMAL	4.58	3.00	19.15	13.01	45.26	5.76
ESS1	1.84	0.08	2.95	3.79	35.77	16.58
ESS2	1.91	0.10	4.35	5.91	40.70	16.66
ESS3	2.71	0.99	10.94	11.20	46.78	16.60
ESS4	2.82	0.99	11.00	11.31	46.95	16.66
CANONICAL	4.11	14.09	23.69	20.77	49.19	16.66

which is strictly positive. Therefore, the choice given by (11) achieves the minimum roundoff noise for second-order ESS.

Equations (10) and (11) show that the optimal ESS coefficients for the ith section rely in a complicated way on all filter coefficients from the ith section to the filter output. Therefore, the resulting ESS structure is not just a simple approximation of a double-precision implementation with $\beta_{ki} = b_{ki}$. It is possible, however, to view the ESS filter as a double-precision structure with the B least significant bits of the numerator coefficients, a_{ki}, equal to zero, and the B least significant bits of the denominator coefficients, b_{ki}, equal to the ESS coefficients. This can, in turn, be viewed as a special case of a double-precision state-space structure that can be optimized using constrained optimization techniques to arrive at the coefficients given by (10) and (11). Other double-precision state-space structures might also offer attractive tradeoffs between hardware complexity and roundoff noise, but, unfortunately, there appears to be no organized way to search for these structures.

V. Examples

For the same five filter examples used in [10], [12], and [27], plus a narrow-band bandpass elliptic filter, we have computed the noise gain, σ^2/σ_0^2, for the following structures: the section-optimal LSS structure [27]; the ESS structure with two optimal ESS feedback gains (ESS1); the ESS structure with the two optimal ESS feedback gains rounded to $\{0, \pm 0.5, \pm 1, \pm 2\}$ (ESS2); the ESS structure with one optimal ESS feedback gain (ESS3); the ESS structure with one optimal ESS feedback gain rounded to $\{0, \pm 0.5, \pm 1, \pm 2\}$ (ESS4); and the canonical structure, included as a reference. The results are given in Table II. The second-order sections for each filter were ordered using an "ascending order" policy explained in the next section. Table III lists the ESS feedback coefficients for the filters in Table II. All of the assumptions of Section III are in force here. In particular, rounding is assumed after each adder, as is necessary for the ESS structures we consider. Therefore, our canonical and section-optimal results differ from those of Jackson *et al.* [27], where rounding was assumed after each multiplication (before each adder). In addition, Chang [10], [12], who gave the first comparison between ESS and LSS structures, compensated his results so he could make a comparison with [27]. Hence, our results differ from his also.

Notice from Table II that all of the ESS structures outperform the section-optimal structure for all of the low-pass filters. For the tenth-order Chebychev filter, the ESS structures do particularly well. For example, the structure ESS2, which adds only shifts, additions, and delay registers to the implementation of the filter, outperforms the section-optimal structure by nearly 15 dB. Considering that one bit provides a 6-dB increase in dynamic range, the wordlengths of the ESS filter could be reduced by at least 2 bits and still provide a noise-gain comparable to that offered by the section-optimal structure. Even ESS4, which adds just a single register and subtraction to each second-order section, performs from 1 to 8 dB better than the section-optimal structure.

For the elliptic bandpass filter in Table II, none of the low-noise structures reduce the noise gain to a low level. However, the ESS structures offer performance that is better than or comparable to the performance of the section-optimal structure. This example also shows that for narrow-band bandpass filters, optimal second-order ESS may perform significantly better than suboptimal ESS; this difference could warrant using optimal ESS.

For the band-reject filter in Table II, the ESS structures perform poorly; indeed, the noise gain is virtually identical to that provided by the canonical structure. This is because the band-reject filter, and hence the corresponding error spectra, have a significant response over most of the frequency band. As mentioned in Section II, the ESS circuitry cannot attenuate the error spectrum over the entire band. In fact, it can be shown analytically that decreasing the error spectrum in one part of the band requires increasing the spectrum in another part of the band [37]. Simple, second-order feedback ESS is unable to confine the spectrum increase to the neighborhood of the narrow reject-band where the error spectra are low. Thus ESS is not a good technique for reducing roundoff noise in narrow-band band-reject filters.

In Table II it should be noted that except for the bandpass filter with second-order ESS, suboptimal ESS performs nearly as well as optimal ESS. This is true for most filters having roundoff noise problems, especially for

TABLE III
ESS FEEDBACK GAINS

FILTER	ESS1		ESS2		ESS3	ESS4
BUTTERWORTH LPF, N=6 $\lambda_c = \pi/2$	β_{11}= -1.043,	β_{21}= 0.657	β_{11}= -1,	β_{21}= .5	β_{11}= -0.630	β_{11}= -.5
	-1.224,	0.648	-1,	.5	-0.743	-.5
	-0.990,	0.502	-1,	.5	-0.659	-.5
BUTTERWORTH LPF, N=6 $\lambda_c = \pi/20$	-1.987,	0.996	-2,	1	-0.995	-1
	-1.985	0.991	-2,	1	-0.997	-1
	-1.884,	0.903	-2,	1	-0.990	-1
CHEB-I LPF, N=10	-1.765,	0.930	-2,	1	-0.915	-1
	-1.877,	0.950	-2,	1	-0.962	-1
	-1.920,	0.965	-2,	1	-0.977	-1
	-1.943,	0.972	-2,	1	-0.985	-1
	-1.865,	0.896	-2,	1	-0.984	-1
ELLIPTIC LPF, N=10	-1.771,	0.933	-2,	1	-0.916	-1
	-1.828,	0.943	-2,	1	-0.941	-1
	-1.867,	0.951	-2,	1	-0.957	-1
	-1.898,	0.957	-2,	1	-0.970	-1
	-1.813,	0.871	-2,	1	-0.969	-1
ELLIPTIC BPF, N=8	-1.302,	0.979	-1,	1	-0.658	-.5
	-1.236,	0.982	-1,	1	-0.624	-.5
	-1.284,	0.958	-1,	1	-0.656	-.5
	-1.160,	0.772	-1,	1	-0.635	-.5
CHEB-II BRF, N=6	0.123,	0.065	0,	0	0.120	0
	0.175,	0.150	0,	0	0.158	0
	0.109,	0.054	0,	0	0.103	0

narrow-band low-pass and high-pass filters where suboptimal ESS places a null in the error spectrum at 0 or π and the original filter characteristic attenuates the error spectrum elsewhere. In addition, the sensitivity of the optimal structure, ESS1, is not improved over the canonical, whereas a low sensitivity implementation is available for ESS2 [17], [20] that is similar to a structure developed by Agarwal and Burrus [38]. For most filters, optimal ESS should therefore be considered only if the additional hardware/software cost is slight (e.g., when the cost of a multiplication is the same as the cost of a shift).

We next examine the section ordering problem and give further numerical results for the filters in Table II. Since section ordering is critical for ESS elliptic filters, we also consider two new elliptic designs.

VI. The Effect of Section Ordering

The roundoff error of a cascade-form digital filter is greatly influenced by the specific configuration of the cascade—i.e., how pole pairs are matched with zero pairs within sections and how the sections are ordered [36]. Jackson [36] proposed simple, heuristic rules for devising a configuration that has low roundoff error, but these rules do not guarantee a configuration with roundoff error near the optimal one [39].

Several researchers [39]–[42] have devised iterative algorithms that generate nearly optimal configurations. Butterworth, Chebychev, and Bessel filters, filters that have identical zero pairs, need not be designed with such an algorithm, because they only have $(N/2)!$ configurations for an Nth-order filter—all of the configurations could be checked. Elliptic filters, though, can be designed more quickly with such algorithms, since a typical Nth-order elliptic filter has $((N/2)!)^2$ configurations.

One such algorithm, the Liu–Peled algorithm [39], partially optimizes the configuration for a cascade-form digital filter. Basically, the algorithm works as follows. Generate a random ordering of zero pairs and a random ordering of pole pairs. Keeping the zero ordering fixed and examining all pairwise permutations of pole pairs, find the configuration that gives the smallest roundoff error—this is the local optimum for this random start. Liu and Peled showed that this algorithm can nearly optimize the configuration for a typical cascade-form filter and do so by considering only $(M(M-1)/2+1)$ separate configurations per random start, where M is the number of sections.

Suppose the transfer function (4) has poles

$$p_i = (1-\delta_i)e^{\pm j\theta_i}$$

and zeros

$$z_i = (1-\sigma_i)e^{\pm j\phi_i}$$

where δ_i is the distance of p_i from the unit circle and θ_i is the angle that p_i makes with the real axis and where σ_i and ϕ_i have interpretations analogous to δ_i and θ_i. Assume that the pole pair having the largest θ_i is paired with the zero pair having the largest ϕ_i, the pole pair having the second largest θ_i is paired with the zero pair having the second

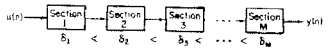


Fig. 5. Ascending ordering energy strage strategy for a cascade-form digital filter using ESS.

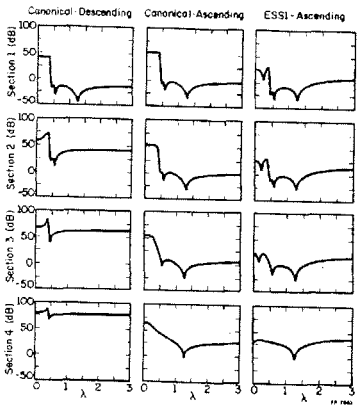


Fig. 6. Error spectra for three configurations of an eighth-order low-pass elliptic filter with cutoff frequency $\lambda c = /8$.

largest ϕ_i, and other pole pairs and zero pairs are grouped similarly.[2]

When a cascade-form filter incorporates ESS, the section-ordering problem gains a new wrinkle. Heuristically, a "good" strategy for configuring a cascade-form low-pass filter that uses ESS should—to give low roundoff error—force as many of the error spectra of the perspective configuration as possible to have low-pass magnitude characteristics; then, appropriate ESS can reduce the error in the passbands of the error spectra. (Similar lines of reasoning hold for bandpass and high-pass filters.)

To accomplish this, the cascade should have the section with smallest δ_i appear first in the cascade, the section with the next smallest δ_i appear second, and remaining sections ordered similarly. Fig. 5 depicts this ordering strategy. Fig. 6 shows the error spectra for an eighth-order low-pass elliptic filter for three configurations: (a) canonical form that has its sections ordered by descending values of δ_i (Canonical-Descending); (b) canonical form that has its sections ordered by ascending values of δ_i (Canonical-Ascending), the configuration arrived at by applying Fig. 5; and (c) optimal ESS form with sections ordered by ascending values of δ_i (ESS1-Ascending).

For Canonical-Descending, none of the error spectra has a low-pass magnitude characteristic except the first, which

[2] This assumption does not have any significance for Butterworth, Chebychev, and Bessel filters. It is important, for our purposes, for elliptic filters

TABLE IV
NOISE GAINS FOR FOUR CONFIGURATIONS OF AN EIGHTH-ORDER LOW-PASS ELLIPTIC FILTER WITH $\lambda_c = \pi/8$

CONFIGURATION	σ^2/σ_0^2 (DB)
CANONICAL-DESCENDING	17.88
CANONICAL-ASCENDING	21.34
ESS1-DESCENDING	14.59
ESS1-ASCENDING	1.40

is the entire low-pass filter's response (minus a scale factor)—in fact, they have wide passbands; hence ESS1 will not be able to reduce roundoff error well. Notice that this problem is particularly acute for the error spectrum from the fourth section. This response is that of a single second-order elliptic section (the last in the cascade) with small δ_i. Unlike Butterworth, Chebychev, and Bessel filters, the zeroes of an elliptic filter are typically near the poles. For small δ_i the resulting response has a near-discontinuity at the pole/zero angle and is comparatively flat with a large amplitude elsewhere. In addition to making the descending ordering a poor one for ESS filters, this explains why second-order feedback-ESS is ineffective for DF2 second-order elliptic filters, as previously concluded by Higgins and Munson [21].

Although the ESS1 structure is not well suited to second-order elliptic filters with small δ_i, it can be extremely effective for higher order cascade-form filters by putting sections with smallest δ_i first in the cascade. The second column in Fig. 6 shows the error spectra for the Canonical-Ascending configuration suggested earlier. Note that Canonical-Ascending has all four of its error spectra being roughly low pass. Therefore, the ESS circuitry can attenuate the magnitude of the error in the "passbands" while the original error spectra keep the error low in the "stopbands." This effect is clearly demonstrated for ESS1-Ascending in the third column in Fig. 6.

Table IV gives the noise gains of the 8th-order elliptic filter for the configurations in Fig. 6, and also for the optimal ESS form with sections ordered by descending values of δ_i (ESS1-Descending). As expected, the noise gain for ESS1-Ascending is much lower than that for ESS1-Descending.

Using the five standard filters from Section V and a new elliptic low-pass filter with cutoff frequency $\lambda_c = \pi/20$, Table V gives the noise gains for three filter types (ESS1, canonical, and section-optimal) and for three configurations (sections arranged in order of ascending δ_i, sections arranged in order of descending δ_i, and the configuration that arises after applying the Liu–Peled algorithm to 25 random starts). From these results, we observe the following:

(1) Section ordering does not drastically influence the effectiveness of ESS on Chebychev and Butterworth filters, but, as we noted earlier, it may be feasible for filters of these types to examine all possible section orderings to derive the best ordering. Our

TABLE V
σ^2/σ_0^2 IN DECIBELS FOR SEVERAL CASCADE-FORM DESIGNS WITH SECTION ORDERING TAKEN INTO ACCOUNT

FILTER AND ORDERING POLICY	ESS1	CANONICAL	SECTION-OPTIMAL
BUTTERWORTH, LPF, N=6, $\lambda_c = \pi/2$			
ASCENDING	1.84	4.11	4.58
DESCENDING	1.87	3.48	4.46
LIU-PELED	1.77	3.46	4.22
BUTTERWORTH, LPF, N=6, $\lambda_c = \pi/20$			
ASCENDING	.079	14.09	3.06
DESCENDING	.084	14.07	3.06
LIU-PELED	.079	13.49	2.78
CHEB-I, LPF, N=10, $\lambda_c = \pi/4$			
ASCENDING	2.95	23.69	19.15
DESCENDING	4.27	22.87	19.15
LIU-PELED	1.15	13.11	8.40
ELLIPTIC, LPF, N=10, $\lambda_c = \pi/4$			
ASCENDING	3.79	20.77	13.01
DESCENDING	9.74	15.85	12.91
LIU-PELED	2.89	13.00	8.61
ELLIPTIC, LPF, N=10, $\lambda_c = \pi/20$			
ASCENDING	3.87	31.18	11.23
DESCENDING	28.42	30.38	11.20
LIU-PELED	15.53	25.50	7.88
CHEB-II, BRF, N=6, BWIDTH= $\pi/10$			
ASCENDING	16.58	16.66	5.76
DESCENDING	14.10	14.14	5.68
LIU-PELED	12.82	13.55	5.68

proposed ordering strategy, though, gives reasonable results.

(2) Our ordering strategy dramatically improves the performances of narrow-band elliptic filters that use ESS. For the elliptic low-pass filter with $\lambda_c = \pi/20$, for example, the strategy produced a better configuration that even the Liu–Peled algorithm.

VII. SUMMARY

A discussion of how ESS reduces roundoff error in a simple second-order digital filter was given. This discussion served both as a tutorial to designers wishing to apply ESS to their filters and as an introduction to an examination of ESS in cascade-form filters.

The ESS coefficients that minimize the roundoff error for recursive, fixed-point cascade-form digital filters implemented with second-order ESS were derived. The optimal ESS coefficients can be easily computed using numerical integration.

A comparison between optimal and suboptimal ESS structures, the section-optimal LSS structure, and the canonical structure, for several cascade-form digital filters, revealed that: (1) ESS reduces roundoff error well, even in comparison to the section-optimal structure; and (2) suboptimal forms of ESS, which are simple to implement, reduce roundoff error well for most filters of interest. The knowledge gained from this work indicates that ESS—particularly suboptimal forms—is an excellent technique for reducing roundoff error in narrow-band low-pass, high-pass, and bandpass filters.

How the cascade with configured—i.e., how pole pairs were matched with zero pairs within sections and how sections were ordered—was also found to significantly influence the roundoff error. For cascade-form filters using ESS, a heuristic section-ordering strategy was proposed where sections were cascaded in order of ascending δ_i, the distance of the poles of the ith section from the unit circle. This strategy was shown to be useful, especially for elliptic filters.

REFERENCES

[1] H. A. Spang and P. M. Schultheiss, "Reduction of quantizing noise by use of feedback," *IRE Trans. Commun.*, vol. CS-10, pp. 373–380, Dec. 1962.
[2] T. Ishiguro and H. Kaneko, "A nonlinear DPCM codec based on ΔM/DPCM code conversion with digital filter," in *Proc. IEEE Inter. Conf. on Communications*, Montreal, pp. 1-27–1-32, June 14–16, 1971.
[3] T. Ishiguro, T. Oshima, and H. Kaneko, "Digital DPCM codec for TV signals based on ΔM/DPCM digital conversions," *IEEE Trans. Commun.*, vol. COM-22, pp. 970–976, July 1974.
[4] Tran-Thong and B. Liu, "A recursive digital filter using DPCM," *IEEE Trans. Commun.*, vol. COM-24, pp. 2–11, Jan. 1976.
[5] T. Claasen and L. Kristiansson, "Improvement of overflow behavior of 2nd-order digital filters by means of error feedback," *Electron. Lett.*, vol. 10, pp. 240–241, June 13, 1974.
[6] Tran-Thong and B. Liu, "Error spectrum shaping in narrow-band recursive filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 200–203, Apr. 1977.
[7] T. L. Chang, "A low roundoff noise digital filter structure," in *Proc. IEEE Int. Symp. on Circuits and Systems*, New York, pp. 1004–1008, May 17–19, 1978.
[8] A. I. Abu-El-Haija and A. M. Peterson, "An approach to eliminate roundoff errors in digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 195–198, Apr. 1979.
[9] T. L. Chang, "Error-feedback digital filters," *Electron. Lett.*, vol. 15, pp. 348–349, June 7, 1979.
[10] T. L. Chang, "Error spectrum shaping structures for digital filters," in *Proc. 13th Asilomar Conf. on Circuits, Systems, and Computers*, (Pacific Grove, CA), pp. 279–283, Nov. 5–7, 1979.

[11] L. Mintzer, and J. P. Strauss, "A practical digital filter with near optimal rounding noise cancellation," in *Proc. 13th Asilomar Conf. on Circuits, Systems, and Computers*, (Pacific Grove, CA), pp. 263–266, Nov. 5–7, 1979.

[12] T. L. Chang, "Comparison of round-off noise variances in several low round-off noise digital filter structures," *Proc. IEEE*, vol. 68, pp. 173–174, Jan. 1980.

[13] T. L. Chang, "Comments on "An approach to eliminate roundoff errors in digital filters,"" *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, p. 244, Apr. 1980.

[14] A. I. Abu-El-Haija, "Author's reply to "An approach to eliminate roundoff errors in digital filters,"" *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 244–245, Apr. 1980.

[15] T. L. Chang, "Correction to "Comparison to round-off noise variances in several low round-off noise digital filter structures,"" *Proc. IEEE*, vol. 68, p. 1167, Sept. 1980.

[16] ——, "Suppression of limit cycles in digital filters designed with one magnitude-truncation quantizer," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 107–111, Feb. 1981.

[17] D. C. Munson, Jr., and B. Liu, "Narrowband recursive filters with error spectrum shaping," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 160–163, Feb. 1981.

[18] T. L. Chang, "A unified analysis of roundoff noise reduction in digital filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Atlanta, GA), pp. 1209–1212, Mar. 30–Apr. 1, 1981.

[19] T. L. Chang and S. A. White, "An error cancellation digital-filter structure and its distributed-arithmetic implementation," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 339–342, Apr. 1981.

[20] T. L. Chang, "On low-roundoff-noise and low-sensitivity digital-filter structures," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 1077–1080, Oct. 1981.

[21] W. E. Higgins and D. C. Munson, Jr., "Noise reduction strategies for digital filters: Error spectrum shaping versus the optimal linear state-space formulation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 963–973, Dec. 1982.

[22] S. Y. Hwang, "Roundoff noise in state-space digital filtering: A general analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 256–262, June 1976.

[23] 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.

[24] S. Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 273–281, Aug. 1977.

[25] C. T. Mullis and R. A. Roberts, "Roundoff noise in digital filters: Frequency transformations and invariants," *IEEE Trans. Acoust., Speech, Signal-Processing*, vol. ASSP-25, pp. 273–281, Aug. 1977.

[26] W. L. Mills, C. T. Mullis, and R. A. Roberts, "Digital filter realizations without overflow oscillations," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 334–338, Aug. 1978.

[27] 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.

[28] W. L. Mills, C. T. Mullis, and R. A. Roberts, "Low roundoff noise and normal realizations of fixed point IIR digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 893–903, Aug. 1981.

[29] C. M. Rader and B. Gold, "Effects of parameter quantization on the poles of a digital filter," *Proc. IEEE*, vol. 55, pp. 688–689, May 1967.

[30] C. W. Barnes and A. T. Fam, "Minimum norm recursive digital filters that are free of overflow limit cycles," *IEEE Trans. Circuits, Syst.*, vol. CAS-24, pp. 569–574, Oct. 1977.

[31] C. W. Barnes, "Roundoff noise and overflow in normal digital filters," *IEEE Trans. Circuits. Syst.*, vol. CAS-26, pp. 154–159, Mar. 1979.

[32] C. W. Barnes, "Error feedback in normal realization of recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 72–75, Jan. 1981.

[33] C. T. Mullis and R. A. Roberts, "An interpretation of error spectrum shaping in digital filters," vol. ASSP-30, pp. 1013–1015, Dec. 1982.

[34] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prenetice-Hall, 1975, pp. 150–151.

[35] T. A. C. M. Classen, W. F. G. Mecklenbrauker, and J. B. H. Peek, "Quantization noise analysis for fixed-point digital filters using magnitude truncation for quantization," *IEEE Trans. Circuits, Syst.*, vol. CAS-22, pp. 887–895, Nov. 1975.

[36] L. B. Jackson, "Roundoff-noise analysis for fixed-point digital filters realized in cascade or parallel form," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 107–122, June 1970.

[37] D. P. Looze, private communication.

[38] R. C. Agarwal and C. S. Burrus, "New recursive digital filter structures having very low sensitivity and roundoff noise," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 921–927, Dec. 1975.

[39] B. Liu and A. Peled, "Heuristic Optimization of the cascade realization of fixed-point digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 464–473, Oct. 1975.

[40] W. S. Lee, "Optimization of digital filters and low roundoff noise," in *Proc. 1973 IEEE Int. Symp. Circuit Theory*, (Toronto, Ont., Canada), pp. 381–383.

[41] S. Y. Hwang, "On optimization of cascade fixed-point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-21, pp. 163–166, Jan. 1974.

[42] E. Leuder, "Minimizing the roundoff noise of digital filters by dynamic programming," presented at the 1974 Arden House Workshop on Digital Signal Processing, (Harriman, NY), Jan. 14–17, 1974.

William E. Higgins (S'80) was born in Philadelphia, PA, on June 2, 1957. He received the B.S. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1979 and the M.S. degree in electrical engineering from the University of Illinois, Urbana, in 1981.

He is presently a Research Assistant at the University of Illinois, where he is working toward the Ph.D. degree in electrical engineering. His research interests are in digital signal processing.

Mr. Higgins is a member of Eta Kappa Nu and Tau Beta Pi.

David C. Munson, Jr. (S'75–M'79) was born in Red Oak, IA, on October 19, 1952. He received the B.S. degree in electrical engineering from the University of Delaware, Newark, DE, in 1975, and the M.S., M.A., and Ph.D. degrees in electrical engineering from Princeton University, Princeton, NJ, in 1977, 1977, and 1979, respectively.

Since August 1979, he has been with the University of Illinois, Urbana, IL, where he is currently an Associate Professor in the Department of Electrical Engineering and an Associate Research Professor in the Coordinated Science Laboratory.

His research interests include the effects of finite register length in digital signal processing, signal processing architectures, multidimensional and time-varying signal processing, and image reconstruction in sensor array processing systems such as synthetic aperture radar.

Dr. Munson is a member of Eta Kappa Nu and Tau Beta Pi.

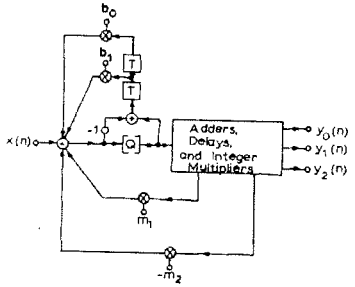


Fig. 6. Application of ESS.

TABLE VI
OPTIMUM STRUCTURES

Range of α_1	Structure	Condition for Optimality
$-2<\alpha_1<-1.5$	II-1	$-2<\alpha_1<-1.5$
$-1.5<\alpha_1<-0.5$	I-5	$-1.5<\alpha_1<-1.0$
	I-6	$-1.0<\alpha_1<-\alpha_2$
	II-2	$-\alpha_2<\alpha_1<-0.5$
$-0.5<\alpha_1<0.5$	I-7	$-0.5<\alpha_1<0.5$
$0.5<\alpha_1<1.5$	I-10	$1.0<\alpha_1<1.5$
	I-11	$\alpha_2<\alpha_1<1.0$
	II-5	$0.5<\alpha_1<\alpha_2$
$1.5<\alpha_1<2$	II-6	$1.5<\alpha_1<2$

The above approach has been applied for the remaining ranges of α_1. The optimum structures identified are summarized in Table VI.

It should be mentioned that the sensitivity analysis of this section, like that reported in [9], is more appropriate for digital-filter structures implemented in terms of floating-point arithmetic or in terms of fixed-point arithmetic but with the coefficients stored in normalized floating-point form. Nevertheless, the experimental results presented in Section VI show that the conclusions reached also hold in the case where standard fixed-point arithmetic is used in conjunction with signal scaling.

V. APPLICATION OF ESS

ESS can be applied in the structures of Figs. 2 to 3 by including a quantizer Q and an appropriate substructure at the output of the input adder, as illustrated in Fig. 6. In this way, the power-spectral density of the output noise can properly be shaped. In effect, error feedback is applied, which can be adjusted to force zeros in the power-spectral density of the output noise and by choosing coefficients b_0 and b_1 using the formulas in [8], the output noise can be reduced or minimized.

The application of ESS requires double-precision arithmetic for the input adder as well as for multipliers m_1 and m_2. However, single-precision arithmetic is entirely satisfactory for ESS multipliers b_0 and b_1. It should be mentioned that the precision of coefficients b_0 and b_1 can be reduced and sometimes b_0 and b_1 can even be chosen to be powers of two so as to replace the corresponding multiplications by data shifts. This reduces the complexity of the structure but our ability to reduce or minimize output noise is diminished.

TABLE VII
FILTER SPECIFICATIONS

Parameter	Elliptic Lowpass Filter	Chebyshev Highpass Filter	Elliptic Bandpass Filter	Butterworth Bandstop Filter
A_p, dB	1.0	0.6	0.5	2.0
A_a, dB	72.9	48.6	66.2	33.9
ω_{p1}, rad/s	250.0	4000.0	1180.0	2300.0
ω_{p2}, rad/s	-	-	1220.0	2700.0
ω_{a1}, rad/s	400.0	3300.0	1050.0	2450.0
ω_{a2}, rad/s	-		1350.0	2550.0
ω_s, rad/s	10,000.0			

VI. COMPARISON

For the sake of comparison, the cascade approach was used to design four 6th-order filters which included an elliptic low pass, a Chebyshev high pass, an elliptic bandpass, and a Butterworth bandstop filter. The specifications of the various filters are given in Table VII where

A_p Passband ripple, dB
A_a Minimum stopband attenuation, dB
ω_{p1}, ω_{p2} Passband edges, rad/s
ω_{a1}, ω_{a2} Stopband edges, rad/s
ω_s sampling frequency, rad/s.

Designs were obtained with the proposed structures [1], with the direct canonic structure, and with the section-optimal structure described in [15]. Signal scaling, based on the L_∞ norm, was applied in all designs, and the section ordering was chosen to minimize the output noise in each case. The implementation was assumed to be in terms of fixed-point arithmetic, and quantization of coefficients and products was assumed to be by rounding. The various designs were compared on the basis of actual amplitude responses with the coefficients quantized, and on the basis of computed output-noise spectra.

The results of the sensitivity analysis are depicted in Fig. 7(a)–(d). As can be seen, for the lowpass filter the section-optimal structure leads to better results than the proposed structure; for the high-pass and bandpass filters, the best results are achieved by using proposed structures II-6 and I-5, respectively; and for the bandstop filter the best results are achieved by using either proposed structure I-7 or the canonic structure. The most sensitive structure is the canonic structure for the low-pass and high-pass filters, and the section-optimal structure for the bandpass and bandstop filters.

The roundoff-noise analysis involved the computation of the relative power spectral density (RPSD) [1] in the various filter designs. In designs based on the proposed structures, ESS has been applied using the formulas in [8]. The quantization of products was carried out after the adders for all sections in the cascade except for the numerator multipliers of the output section where the quantization was carried out before the adder. ESS has not been applied in the corresponding designs based on the section-optimal structure since the number of multiplications would then become excessive. It should be mentioned that the proposed structures with ESS incorporated require between five and seven multipliers per section whereas the section-optimal structure without ESS requires nine multipliers per section.

The results of the roundoff-noise analysis are depicted in Fig. 8(a)–(d). As can be seen, for the low-pass, high-pass, and bandpass filters, the proposed structures with ESS incorporated lead to superior results, in particular, if second-order ESS is applied. For the bandstop filter, on the other hand, the passband average of the RPSD is lower in the section-optimal structure whereas the passband maximum of the RPSD is lower in the proposed structure. It should be mentioned that a similar reduction in output roundoff noise can be achieved by applying ESS to direct canonic structures. However, as was pointed out earlier, these structures are almost always more sensitive relative to the proposed structures.

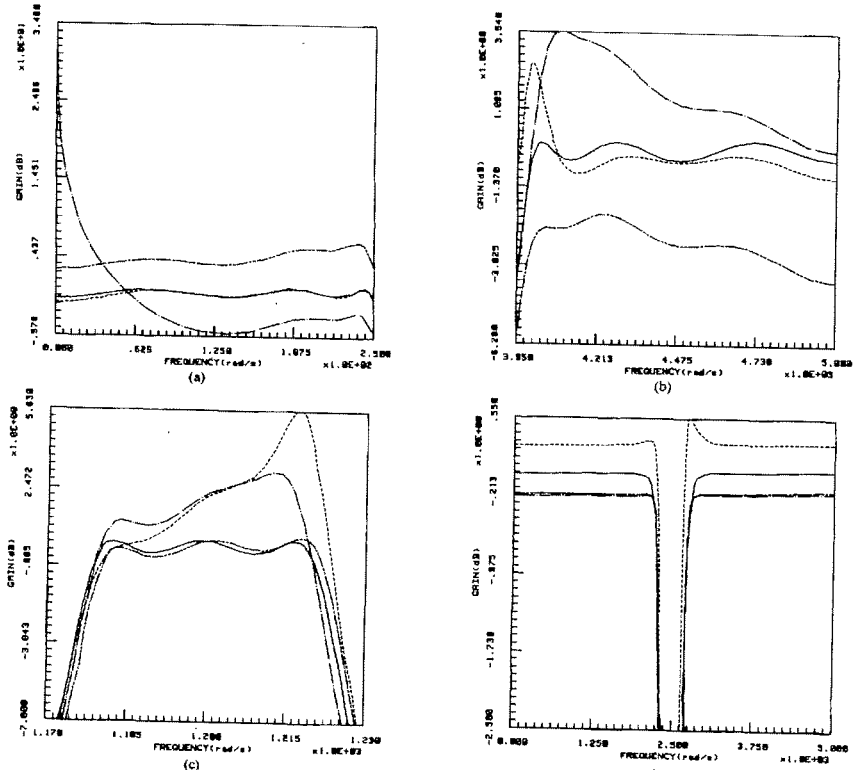


Fig. 7. Amplitude responses. (a) Low-pass filter, 8 bits. (b) High-pass filter, 5 bits. (c) Bandpass filter, 7 bits. (d) Bandstop filter, 5 bits. ——— Ideal amplitude response, ----- Section-optimal structure, — · — · — Direct-canonic structure, — — — — New structure (II-1 for lowpass filter, II-6 for high-pass filter, I-5 for bandpass filter, I-7 for bandstop filter).

VII. CONCLUSIONS

A systematic and exhaustive procedure has been described for the generation of economical low-sensitivity digital-filter structures which are amenable to the application of ESS. The procedure has then been used to generate the structures of Tables I and II.

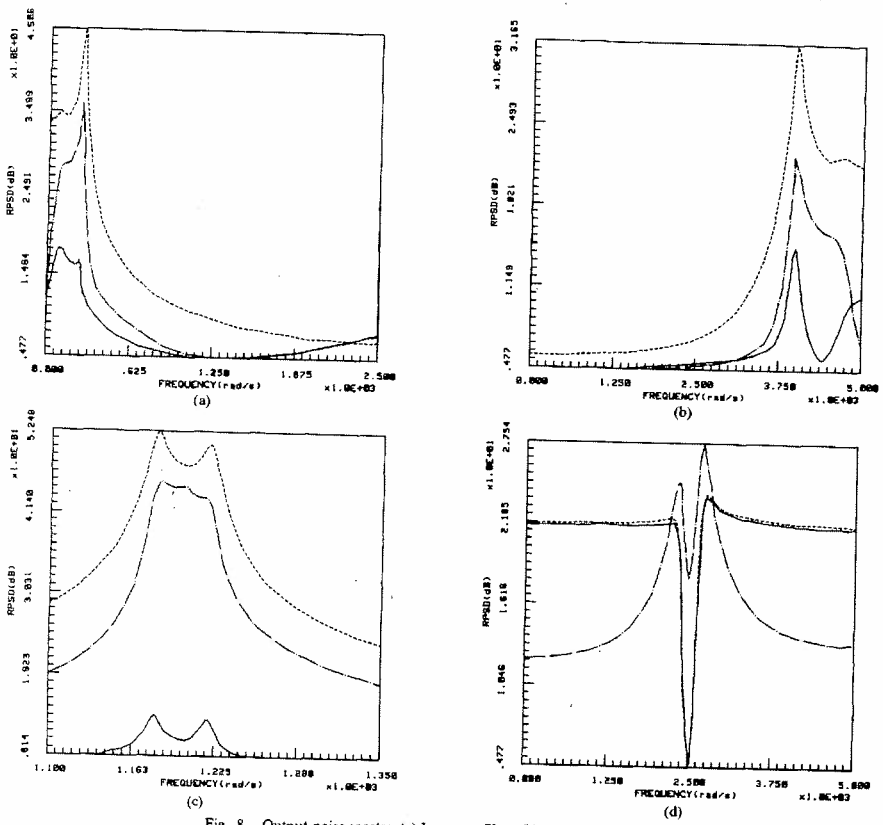


Fig. 8. Output-noise spectra. (a) Low-pass filter. (b) High-pass filter. (c) Bandpass filter. (d) Bandstop filter: ----- Section-optimal structure, — · — · — New structure with 1st-order ESS, ——— New structure with 2nd-order ESS (II-1 for low-pass filter, II-6 for high-pass filter, I-5 for bandpass filter, I-7 for bandstop filter).

In all, 21 structures were obtained and except for structures I-1, I-2, and II-1, which were reported in [9], [10], they are thought to be new. For each range of α_1, at least two distinct structures are possible but, through the sensitivity analysis of Section IV, the optimum structure for the transfer function under consideration can easily be chosen, as shown in Table VI.

A sensitivity comparison has shown that the proposed structures are usually superior relative to corresponding direct canonic structures and also relative to the section-optimal structure.

A roundoff noise comparison has shown the proposed structures with ESS incorporated to be superior relative to the section-optimal structure, except in the case of bandstop filters where a lower passband average of the RPSD can be achieved by using the section-optimal structure. Further, despite the additional multipliers needed for ESS, these structures require fewer multipliers per second-order section.

Acknowledgment

The authors are grateful to Coordenação do Aperfeiçoamento de Pessoal de Nivel Superior, Brazil, and to the Natural Sciences and Engineering Research Council, Canada for supporting this research.

References

[1] A. Antoniou, *Digital Filters: Analysis and Design*. New York: McGraw-Hill, 1979.

[2] A. Fettweis, "Digital filter structures related to classical filter networks," *Arch. Elek. Übertragung.*, vol. 25, pp. 79–89, Feb. 1971.

[3] A. Antoniou and M. G. Rezk, "A comparison of cascade and wave fixed-point digital filter structures," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 1184–1194, Dec. 1980.

[4] T. Thong and B. Liu, "Error spectrum shaping in narrowband recursive digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 200–203, Apr. 1977.

[5] T. L. Chang, "Error spectrum shaping structures for digital filters," in *Proc. 13th Asilomar Conf. Circuits Syst. Comput.*, Pacific Grove, CA, pp. 279–283, Nov. 1979.

[6] D. C. Munson and B. Liu, "Narrow-band recursive filters with error spectrum shaping," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 160–163, Feb. 1981.

[7] W. E. Higgins and D. C. Munson, "Noise reduction strategies for digital filters: Error spectrum shaping versus the optimal linear state-space formulation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 963–973, Dec. 1982.

[8] ——, "Optimal and suboptimal error-spectrum shaping for cascade-form digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 429–437, May 1984.

[9] R. C. Agarwal and C. S. Burrus, "New recursive digital filter structures having very low sensitivity and roundoff noise," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 921–927, Dec. 1975.

[10] S. Nishimura, K. Hirano, and R. N. Pal, "A new class of very low sensitivity and low roundoff noise recursive digital filter structures," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 1152–1158, Dec. 1981.

[11] K. Haug and E. Lüder, "Determination of all equivalent and canonic second order digital filter structures," *Arch. Elek. Übertragung.*, vol. 36, pp. 436–442, Nov.–Dec. 1982.

[12] D. C. Munson and B. Liu, "ROM/ACC realization of digital filters for poles near the unit circle," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 147–151, Feb. 1982.

[13] ——, "Low-noise realization for narrow-band recursive digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 41–54, Feb. 1980.

[14] O. Monkewich and W. Steenaart, "Stored product digital filtering with nonlinear quantization," in *Proc. IEEE Int. Symp. Circuits Syst.*, pp. 157–160, 1976.

[15] 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.

Paulo S. R. Diniz (S'80–M'84) was born in Niteroi, Brazil. He received the B.Sc. degree from Universidade Federal do Rio de Janeiro (UFRJ), in 1978, the M.Sc. degree from COPPE/UFRJ in 1981, and the Ph.D. degree from Concordia University, Montreal, Canada, in 1984, all in electrical engineering.

Since 1979, he has been with the Department of Electronic Engineering, UFRJ. He has also been with the Department of Electrical Engineering of COPPE/UFRJ, since 1984, where he is associate professor. His teaching and research interests include passive and active network synthesis, electronic circuits, digital filters and digital signal processing.

Andreas Antoniou (M'69–SM'79–F'82) received the B.Sc.(Eng.) and Ph.D degrees in Electrical Engineering from London University in 1963 and 1966, respectively.

From 1966 to 1969 he was Senior Scientific Officer at the Post Office Research Department, London, England, and from 1969 to 1970, he was a member of the Scientific Staff at the R&D Laboratories of Northern Electric Company Ltd., Ottawa, Ontario, Canada. From 1970 to 1983 he served in the Department of Electrical Engineering, Concordia University, Montreal, Quebec, Canada, as Professor from June 1973 and as Chairman from December 1977. On July 1, 1983 he was appointed founding Chairman of the Department of Electrical Engineering, University of Victoria, Victoria, B.C., Canada.

His teaching and research interests are in the areas of electronics, network synthesis, digital system design, active and digital filters, and digital signal processing. He has published a number of papers on electronic circuits, active filters, and digital filters. He has authored *Digital Filters: Analysis and Design*, (New York: McGraw-Hill). One of his papers on gyrator circuits was awarded the Ambrose Fleming Premium by the Institution of Electrical Engineers, UK.

Dr. Antoniou is a Member of the Order of Engineers of Quebec, and a Fellow of the Institution of Electrical Engineers. He was Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS during the period June 1983 to May 1985. He is now serving as Editor of the same TRANSACTIONS.

References

[1] A. Antoniou, *Digital Filter, Analysis and Design*. New York: McGraw-Hill, 1979.

[2] A. Sedlmeyer and A. Fettweis, "Digital filters with true ladder configuration," *Int. J. Circuit Theory and Applications*, vol. 1, pp. 5–10, Mar. 1973.

[3] A. Antoniou and M. G. Rezk, "Digital-filter synthesis using concept of generalized immittance converter," *IEE J. Electron. Circuits Syst.*, vol. 1, pp. 207–216, Nov. 1977.

[4] M. N. S. Swamy and K. S. Thyagarajan, "A new type of wave digital filter," *J. Franklin Inst.*, vol. 300, pp. 41–58, July 1975.

[5] L. B. Jackson, "Roundoff noise analysis for fixed-point digital filters realized in cascade or parallel form," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 107–122, June 1970.

[6] J. W. K. Lam, "Computer-aided analysis of wave digital filters and comparison of the effects of quantization on digital filters," M. Eng. Thesis, Concordia Univ., Montreal, Canada, 1980.

[7] A. Antoniou and M. G. Rezk, "A comparison of cascade and wave fixed-point digital filter structures," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 1184–1194, Dec. 1980.

An Interpretation of Error Spectrum Shaping in Digital Filters

C. T. MULLIS AND R. A. ROBERTS

Abstract—Error spectrum shaping is used in digital filters to reduce roundoff noise. There are several other filter structures which can also reduce roundoff noise. Are these methods of noise reduction in any way related? This paper shows that optimal error spectrum shaping is a method of increasing the word length of accumulators and internal variables, i.e., it is, in its optimum form, extended precision arithmetic. Furthermore there is no hardware advantage to be gained with optimum ESS as compared to traditional methods of implementing extended precision arithmetic.

I. Introduction

In digital systems one is generally forced to consider the effects of quantization noise. Since this noise places a fundamental limit on the accuracy of digital systems, it has received a great deal of study dating back to the late 1940's [1]–[3]. One technique originally proposed within the context of converting analog signals into digital signals for transmission over discrete channels employs the use of error feedback in order to reduce quantizer noise [4]. More recently this technique has been proposed by several authors as a method of reducing roundoff noise in fixed point digital filters [5]–[14]. In this application the technique is commonly termed error spectrum shaping (ESS). ESS has been shown to significantly reduce the mean square output roundoff noise and the amplitude of zero-input limit cycles in fixed point filters [14]–[16].

Roundoff noise is often of critical importance in the design of fixed point filters. Thus, techniques for the reduction of this noise are of practical significance in the hardware implementation of filters. About the time the first papers in ESS applied to roundoff noise reduction appeared in the literature, a series of papers concerning the synthesis of minimum roundoff noise structures were published [17]–[20]. These papers detail a theory of minimizing roundoff noise that is based on a state space model of digital filters. These new structures are optimized assuming word length constraints, an l_2 scaling rule, and a model that replaces the quantizer with an additive white noise source. This is the same quantizer model employed in ESS.

Recently some comparisons have been made between ESS and state-space, low-noise, and second-order structures [11], [12]. Comparisons of this sort are meaningful only if both performance and hardware complexity are included in the comparison. Performance in this case includes such effects as roundoff noise, overflow oscillations, coefficient sensitivity, etc. Using direct forms with ESS, for example, will entail more conservative scaling rules because direct forms (with ESS) will support overflow oscillations. This is not true of low-noise, state-space structures. The more conservative the scaling rule, the more roundoff noise is present, all other parameters being equal [21].

In order to help gain an alternate perspective of ESS we have written an analysis of ESS which shows that optimal ESS is nothing more than a method of performing extended precision arithmetic, i.e., arithmetic with longer word lengths. In other words, optimal ESS is exactly equivalent to extending the word length of the internal variables of the filter. Moreover, there is *no hardware savings* in using optimal ESS to accomplish this extended precision arithmetic. This simple interpretation seems to have been overlooked in previous discussions of ESS.

II. Quantization Error and Error Feedback

For purposes of analysis, quantization errors occur at accumulators or "summing junctions." These errors are noise-like and the noise sources due to different accumulators are assumed to be independent. We wish to analyze the output noise due to quantization at a particular accumulator. To this end, let us isolate an accumulator, replace it with an appropriate noise model and calculate the output noise variance.

Error spectrum shaping feedback requires extended precision accumulators in order to "measure" the error to be fed back. Two descriptions of such an accumulator are found in Fig. 1. In the first of these an ideal accumulator is followed by an extended precision quantizer which gives the most significant, and second most significant parts of the input variable. These have m_1 and m_2 bit representations, respectively. This description is *nonlinear*. The second description is useful for analysis based on *linear* system theory, but assumes that the lower order parts of x are noiselike and depend on x only negligibly. This description represents a splitting of an ideal accumulator into two accumulators.

Taking a classical approach, let us reduce the entire filter to a signal flow graph with five nodes: the input, output, internal accumulator, and two noise sources. The result is found in Fig. 2. (We have shortened notation by writing $q_1(t)$ for $q_1(x(t))$ etc.)

The output noise term due to the quantization error $q_2(t)$ is given by the following expression:

$$\sigma^2 = c_0 2^{-2m_1} \|g\|^2, \qquad c_0 \text{ a constant.} \tag{1}$$

Here,

$$\|g\|^2 = \sum_{k=0}^{\infty} g_k^2 \tag{2}$$

where

$$g(z) = \sum_{k=0}^{\infty} g_k z^{-k}.$$

Note that $(-g(z))$ is the transfer function from $q_2(t)$ to $y(t)$. The variance of $q_2(t)$ is $c_0 2^{-2m}$. The variance of $\xi(t)$ is $c_0 2^{-2(m_1+m_2)}$.

Manuscript received August 17, 1981; revised June 8, 1982.
The authors are with the Department of Electrical Engineering, University of Colorado, Boulder, CO 80309.

0096-3518/82/1200-1013$00.75 © 1982 IEEE

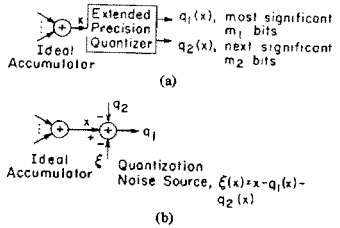


Fig. 1. Quantization models. (a) Nonlinear model for an accumulator. (b) Linear model.

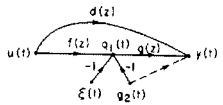


Fig. 2. A reduced signal flow graph.

The motivation for error spectrum shaping begins at this point. First of all, the output noise variance σ^2 is proportional to the square of the norm of the impulse response from noise source to output. Second, if extended precision accumulators are used, the most significant m_2 bits of the error are measurable. This is $q_2(t)$. Therefore, one can modify the impulse response and therefore σ^2 by feeding $q_2(t)$ through a filter $e(z)$ back into one (or more) accumulators inside the digital filter. This will modify the noise transfer function $g(z)$ only; it will not alter the transfer function from $u(t)$ to $y(t)$.

But where should the filtered error be fed? There may be many accumulators to choose from. An optimal solution is obvious but expensive. Simply feed $q_2(t)$ forward to $y(t)$ through a filter with transfer function $e(z) = g(z)$ at the dotted line in Fig. 2. Since this reduces the transfer function from error to output to zero, the output variance due to $q_2(t)$ goes to zero. This may be surprising but is easily explained. It is exactly equivalent to using $m_1 + m_2$ bit internal registers rather than m_1 bit registers. However, the entire filter is "bit-sliced." In other words it processes the two parts of the $m_1 + m_2$ bit word separately, and knits them back together at the extended precision accumulators.

The price, of course, lies in the extra resources necessary to realize $e(z)$. One seeks to avoid the expense of an extra filter for each noisy accumulator. Presumably, this is the reason for feeding the error back into internal accumulators rather than the output. In the following section such an approach is followed.

III. A Direct Form Example

Second order direct form structures are very common in the literature. Fig. 3 shows a direct form structure with error shaping feedback which is reasonably typical of the papers [5]–[14]. Part (a) of the figure contains the details of the structure, but part (b) is the description which is used in the design problem.

The ESS feedback structure is postulated as in the figure. The design problem is to choose a_1' and a_2' to minimize the norm squared of the transfer function

$$T_{y,q_2}(z) = -\left(\frac{1 + a_1' z^{-1} + a_2' z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}}\right). \tag{3}$$

The solution is trivial; since the impulse response sequence associated with the transfer function must begin with unity, the norm squared must be bounded below by one. But this is attainable by setting $a_k' = a_k$. (*This makes the top half of Fig. 3(a) identical with the bottom half.* The extended precision accumulator knits the two together.)

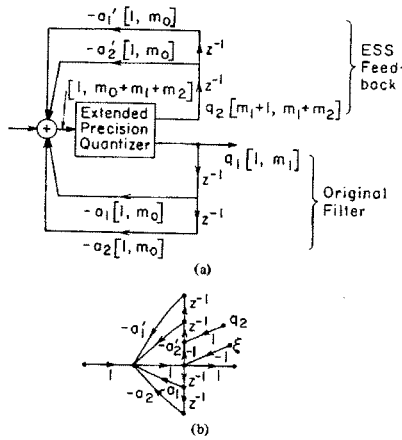


Fig. 3. A direct form ESS realization for $H(z) = 1/1 + a_1 z^{-1} + a_2 z^{-2}$. (a) Block diagram with nonlinear quantization. (b) Linear signal flow graph with noise sources.

Therefore, the optimal ESS structure is a bit sliced version of extended precision arithmetic except that the output itself is quantized to m_1 bits. Obviously one should add q_1 and q_2 to get the output. If this were done the whole structure would be double precision. The output noise variance is then of the order $2^{-2(m_1+m_2)}$ rather than 2^{-2m_1}.

In summary, optimal ESS feedback structures are equivalent to 1) extending the length of internal filter variables, 2) without increasing the length of stored coefficients. The resources needed to do the one are exactly the same as the resources needed to do the other.

IV. State Variable Structures and Error Feedback

Every digital filter has a state variable description, although this description may not contain as much detail as a signal flow graph or block diagram. The state variable equations are given in (4).

$$\begin{aligned} x(t+1) &= Ax(t) + bu(t) \\ y(t) &= cx(t) + du(t). \end{aligned} \tag{4}$$

The state $x(t)$ is vector valued. These equations relate sequences of real numbers without quantization. In other words they are "infinite precision." In Fig. 4 the state variable structure is shown, including extended precision quantization and possible ESS feedback. An equivalent "noisy" but linear description is also given. (The outputs have been deleted for simplicity.) The ESS feedback filter $z^{-1}E(z)$ must contain a delay in order to avoid a delay free loop.

The equations describing the system in Fig. 4(b) follow.

$$x_1(t+1) = Ax_1(t) + bu(t) + \psi(t) - \xi(t+1) - x_2(t+1)$$

$$\psi(t) = \sum_{k=0}^{\infty} b'_{k+1} x_2(t-k). \tag{5}$$

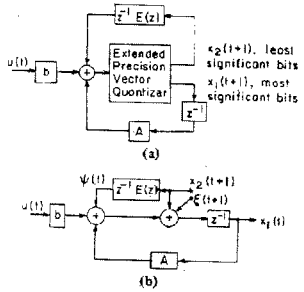


Fig. 4. A state variable realization with ESS feedback. (a) Block diagram with nonlinear quantizer. (b) Block diagram with equivalent noise sources.

Again referring to Fig. 4(b), we compute the matrix transfer function from the "measurable quantization noise" $x_2(t+1)$ to $x_1(t)$ to be

$$G(z) = -z(zI - A)^{-1}(zI - E(z))$$
$$= -z^{-1}I - z^{-2}(A - E_0) - z^{-3}(\cdots). \tag{6}$$

The covariance of the error in $x_1(t)$ due to $x_2(t)$ is therefore

$$R = c_0 2^{-2m} \sum_{k=0}^{\infty} G_k G_k^T. \tag{7}$$

Since $G_0 = 0$ and $G_1 = -I$, $R - I$ is positive semidefinite for all $E(z)$. However this bound is attainable by setting $E(z) = A$.

Therefore the optimal ESS feedback filter is given by $E(z) = A$. This makes the top half of Fig. 4(a) exactly the same as the bottom half of Fig. 4(a). In other words we again see that optimal ESS feedback structures are equivalent to doing extended precision arithmetic with a bit sliced processor. With $E(z) = A$, we have $\psi(t) = Ax_2(t)$ and (5) reduces to (8).

$$[x_1(t+1) + x_2(t+1)] = A[x_1(t) + x_2(t)] + bu(t) + \xi(t). \tag{8}$$

One can write $\hat{x}(t) = x_1(t) + x_2(t)$, and substitute into (8) to get the equation for the double precision filter.

V. Suboptimal ESS

The remarks at the end of Section II indicate that there is a great deal of freedom in employing ESS feedback. One can choose the accumulators which receive the feedback as well as the coefficients in the feedback filter. The goal is to get the maximum benefit (reduced error variance) from the minimum expenditure of resources.

In suboptimal ESS, the hardware resources are minimized by limiting the order of the feedback filter and/or choosing coefficients which are integers or powers of two. Suboptimal ESS structures trade performance for a decrease in complexity.

VI. Conclusions

Optimal ESS feedback structures are equivalent to increasing the word lengths of internal filter variables without increasing the coefficient word lengths. The hardware resources needed to do either are identical.

Comparisons of ESS structures and other low-noise structures should be made on the basis of performance and hardware complexity. Performance should include such factors as roundoff noise, overflow oscillations, and coefficient sensitivity. Equivalent scaling rules will be difficult to obtain in comparing direct form ESS structures and state-space, low-noise structures. However, as Barnes [11] has pointed out, the ESS concepts can be applied to any structure. The analysis and comparisons of suboptimal ESS structures would appear to be on a case by case basis.

References

[1] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, pp. 446–472, July 1948.
[2] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inform. Theory*, vol. IT-6, pp. 7–12, Mar. 1960.
[3] C. C. Cutler, "Transmission systems employing quantization," U.S. Patent 2 927 962, 1960.
[4] H. A. Spang III and P. M. Schultheiss, "Reduction of quantizing noise by use of feedback," *IRE Trans. Commun. Syst.*, vol. CS-10, pp. 373–380, Dec. 1962.
[5] T. Thong and B. Liu, "Error spectrum shaping in narrow band recursive digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 200–203, Apr. 1977.
[6] T. L. Chang, "A low roundoff noise digital filter structure," in *Proc. IEEE Int. Symp. Circuits Syst.*, New York, May 1978, pp. 1004–1008.
[7] A. I. Abu-El-Haija and A. M. Peterson, "An approach to eliminate roundoff errors in digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, Apr. 1979.
[8] T. L. Chang, "Error spectrum shaping structures for digital filters," in *Proc. 13th Asilomar Conf. Circuits, Syst., Comput.*, Pacific Grove, CA, Nov. 1979, pp. 279–283.
[9] L. Mintzer and J. P. Strauss, "A practical digital filter with near optimal rounding noise cancellation," in *Proc. 13th Asilomar Conf. Circuits, Syst., Comput.*, Pacific Grove, CA, Nov. 1979, pp. 263–266.
[10] T. L. Chang, "Comparison of roundoff noise variance in several low roundoff noise digital filter structures," *Proc. IEEE*, vol. 68, pp. 173–174, Jan. 1980 and correction in vol. 68, pp. 1167, Sept. 1980.
[11] C. W. Barnes, "Error feedback in normal realizations of recursive digital filters," *IEEE Trans. Circuits Syst.*, pp. 72–75, Jan. 1981.
[12] T. L. Chang, "A unified analysis of roundoff noise reduction in digital filters," in *Proc. ICASSP*, 1980, pp. 1209–1212.
[13] D. C. Munson, Jr. and B. Liu, "Narrowband recursive filters with error spectrum shaping," in *Proc. ICASSP*, 1979, pp. 367–370.
[14] —, "Narrowband recursive filters with error spectrum shaping, *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 160–163, Feb. 1981.
[15] A. I. Abu-El-Haija, "On limit cycle amplitudes in error-feedback digital filters," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1981, pp. 1227–1230.
[16] T. L. Chang, "Suppression of limit cycles in digital filters designed with one magnitude-truncation quantizer," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 107–112, Feb. 1981.
[17] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, no. 9, pp. 551–562, Sept. 1976.
[18] —, "Roundoff noise in digital filters: Frequency transformations and invariants," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 538–550, Dec. 1976.
[19] S. Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 273–281, Aug. 1977.
[20] 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. 154–159, Mar. 1979.
[21] L. B. Jackson, "On the interaction of roundoff noise and dynamic range in digital filters," *Bell Syst. Tech. J.*, vol. 49, pp. 159–184, Feb. 1970.

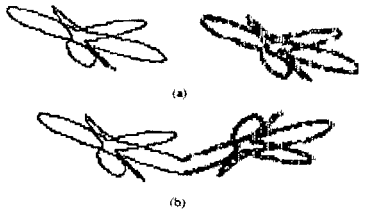


Fig. 3. Trajectories obtained for the connected sum of two (3,4)-torus knots using a computer program. (a) Separate knots before moving for connection. (b) Connected sum

subscripted by 2, and follow K_2 by again solving the semistate equations until we again reach P_1.

For a (3,4)-torus knot since $m_1 = 3$ and $m_2 = 4$, $X_0 = X(t_1)$ is found numerically using the method described in part a) and δ is given by (22), which is a function of time; we choose $\mu = 3$ for both Van der Pol oscillators. We point out that the connected sum of two torus knots may or may not be a torus knot [3].

IV. Discussion

Here we have given the semistate equations for the realization of the connected sum of two identical torus knots using the technique discussed in [1]; this assumes that the semistate equations of the individual knots are available. A numerical example using two (3,4)-torus knots illustrates the theory. We comment that we used identical knots so that the concepts would not be confused by notational difficulties; the ideas however readily apply to nonidentical subknots. Using the technique in [1], we notice that in the semistate equations defining the connected sum, we evaluate $S_1 X'$ and $S_2 X'$ at the connection points P_2 and P_1. This necessitates the use of impulses which serve to reset the pertinent portion of S from 0 to the proper initial value at the connection points to continue smoothly on the new connected sum knot. Of course this involves the assumption that there are trajectories for which we can choose our semistate variables and develop $A_0(S)$. But this existence assumption is implicit in our main assumption that two knots K_1 and K_2 and their semistate realizations are already on hand at the start. In short by applying the method discussed in [1] we find the semistate equations for the connected sum of two identical torus knots with the final equations agreeing with the ones obtained in [2].

References

[1] R. Rassai and R. W. Newcomb, "Realization of the connected sum of two knots," in *Proc. 1986 ISCAS*, San Jose, CA, pp. 483–485, May 1986.

[2] R. Rassai and R. W. Newcomb, "A two-torus oscillator," in *Proc. 28th Midwest Symp. on Circuits and Systems*, Louisville, KY, pp. 685–688, Aug. 1985.

[3] I. H. Kaufman, *Formal Knot Theory*. Princeton, NJ: Princeton Univ. Press, 1983.

[4] R. H. Crowell and R. H. Fox, *Introduction to Knot Theory*. Boston, MA: Ginn, 1963.

[5] S. L. Campbell, *Singular Systems of Differential Equations*. London: Pitman, vol. I, 1980, and vol. II, 1983.

[6] F. L. Lewis and R. W. Newcomb, Eds. "Special Issue: Semistate Systems," *Circuits, Syst., Signal Processing*, vol. 5, no. 1, 1986.

[7] R. W. Newcomb, "The semistate description of nonlinear time-variable circuits," *IEEE Trans. Circuits Syst.*, Vol. CAS-28, pp. 62–71, Jan. 1981.

[8] R. Parris "A three dimensional system with knotted trajectories," *American Math. Mthly.*, vol. 84, no. 6, pp. 468–469, June–July 1977.

[9] R. Rassai, "Theory and electronic circuit design of multi-torus knot oscillators," Ph.D. dissertation, Univ. of Maryland, Dec. 1985.

[10] R. Rassai and R. W. Newcomb, "Van der Pol realization of Torus Knot Oscillators", *IEEE Trans. Circuits Syst.*, vol. 35, pp. 213–220, Feb. 1988.

[11] M. Waite and C. L. Morgan, "*Graphics Primer for the IBM PC*. Berkeley, CA: McGraw-Hill, 1983, pp. 468–469.

A Pole-Sensitivity Based Method for the Design of Digital Filters for Error-Spectrum Shaping

Y. V. RAMANA RAO AND C. ESWARAN

Abstract — A method for realizing second-order digital filter structures which are amenable to error-spectrum shaping (ESS) was proposed recently by Diniz and Antoniou. This method minimizes the non-integer multiplier values for identifying the low-sensitivity structures. An alternative approach based on the concept of pole-sensitivity is presented in this paper for identifying the optimum structures. The proposed method is applicable for any pole location and it does not require a separate sensitivity analysis.

I. Introduction

A pole-sensitivity based procedure is presented in this paper for obtaining low-sensitivity second-order digital filter structures which are amenable to error-spectrum shaping (ESS). Two types of second-order structures which are amenable to ESS were proposed recently by Diniz and Antoniou [1]. Low-sensitivity is achieved in these structures by introducing integer multiplier constants and choosing the optimum values of these constants such that they result in low values of the non-integer multiplier constants. In effect, the procedure of [1] makes use of the following steps for selecting the optimum structure.

1) Generation of low-sensitivity structures by forcing the non-integer multiplier constants to be low.

2) Selection of the optimum structure by a separate sensitivity analysis.

The proposed pole-sensitivity based procedure yields the optimum structure directly and hence the need for a separate sensitivity analysis is eliminated. Further, the proposed method is applicable for transfer functions with arbitrary pole locations. In general, low-sensitivity is desired for any pole location, though it is particularly important for structures with poles close to the unit circle as the sensitivity for such structures is high. We present in this paper a z-plane map containing several regions depending on the optimum set of integer multiplier constants. This map enables one to realize low-sensitivity second-order digital filter structures which are amenable to ESS with poles anywhere inside the unit circle.

A sensitivity comparison of the proposed method and the one given in [1] is also carried out. For a specific transfer function, the optimum structures are obtained by both the methods. The sensitivity performance of both these structures is compared by computing the error in the amplitude characteristics for different wordlengths of the multiplier constants.

Manuscript received December 3, 1987; revised May 31, 1989. This paper was recommended by Associate Editor Y. C. Jenq.

The authors are with the Department of Electrical Engineering, Indian Institute of Technology, Madras 600 036, India.

IEEE Log Number 8927742.

0098-4094/89/0700-1017$01.00 ©1989 IEEE

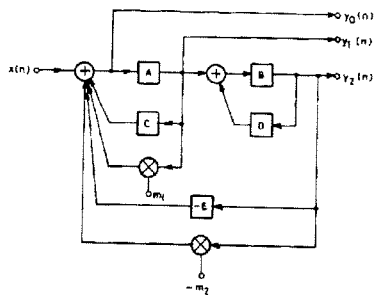


Fig. 1. Second-order structure which is amenable to ESS.

II. Pole-Sensitivity Based Method

Fig. 1 shows the general second-order structure given in [1] which is amenable to ESS. The branches A, B, C, D, and E represent unit delays or integer multipliers whose multiplier constants are restricted to 0, ± 1, ± 2, and m_1 and m_2 represent non-integer multipliers. Imposing the constraints required for avoiding delay free loops and maintaining the number of delays to the minimum of two, we obtain from Fig. 1 the following two possible structures [1]: Structure-I with $A = B = z^{-1}$ and Structure-II with $A = D = z^{-1}$.

The remaining branches, namely, C, D, and E for Structure-I and B, C and E for Structure-II represent the integer constants. The optimum set of these integer constants for any specific pole location $p = re^{\pm j\theta}$ is obtained by minimizing with respect to the non-integer multiplier constants m_1 and m_2, the pole-sensitivity function S defined as [2], [3]

$$\begin{aligned} S &= |s^p_{m_1}|^2 + |s^p_{m_2}|^2 \\ &= \left(s^r_{m_1}\right)^2 + r^2\left(s^\theta_{m_1}\right)^2 + \left(s^r_{m_2}\right)^2 + r^2\left(s^\theta_{m_2}\right)^2 \end{aligned} \tag{1}$$

where

$$s^r_{m_i} = m_i \frac{\partial r}{\partial m_i} \quad \text{and} \quad s^\theta_{m_i} = m_i \frac{\partial \theta}{\partial m_i}, \quad \text{for} \quad i = 1, 2. \tag{2}$$

The characteristic polynomial $D(z)$ of Structure-I is given by [1]

$$\begin{aligned} D(z) &= z^2 - z(C + D + m_1) + CD + m_1 D + m_2 + E \\ &= z^2 + a_1 z + a_2 \\ &= z^2 - 2r\cos\theta + r^2. \end{aligned} \tag{3}$$

From (3), we get

$$\left.\begin{aligned} 2r\cos\theta &= C + D + m_1 \\ \text{and} \quad r^2 &= CD + m_1 D + m_2 + E. \end{aligned}\right\} \tag{4}$$

Using (2) and (4), we obtain

$$\left.\begin{aligned} s^r_{m_1} &= \frac{m_1 D}{2r}; & s^\theta_{m_1} &= \frac{m_1(D\cos\theta - r)}{2r^2\sin\theta} \\ s^r_{m_2} &= \frac{m_2}{2r}; & s^\theta_{m_2} &= \frac{m_2\cos\theta}{2r^2\sin\theta}. \end{aligned}\right\} \tag{5}$$

The pole-sensitivity function for Structure-I denoted as S_1 can be obtained from (1) and (5). For any specific pole location $p = re^{\pm j\theta}$ (which can be anywhere inside the unit circle), the optimum set of integer constants C, D, and E is obtained using an exhaustive search algorithm which comprises the following steps:

Step 1: From the allowed values of 0, ± 1, ± 2, choose a particular combination for the integer constants C, D and E and obtain the values of the corresponding non-integer multiplier constants m_1 and m_2 from (4).

Step 2: Obtain the values of $s^r_{m_1}$, $s^\theta_{m_1}$, $s^r_{m_2}$, and $s^\theta_{m_2}$ from (5) and evaluate the pole-sensitivity function S_I using (1).

Step 3: Repeat Steps 1 and 2 for all possible combinations of C, D and E and identify the optimum set as the one which yields the minimum value for S_I.

It can be shown that for Structure-II we have,

$$\left.\begin{aligned} 2r\cos\theta &= B + C + m_1 - m_2 B - BE \\ \text{and} \quad r^2 &= BC + m_1 B. \end{aligned}\right\} \tag{6}$$

From (2) and (6), we get

$$\left.\begin{aligned} s^r_{m_1} &= \frac{Bm_1}{2r}; & s^\theta_{m_1} &= \frac{-m_1(r - B\cos\theta)}{2r^2\sin\theta} \\ s^r_{m_2} &= 0; & s^\theta_{m_2} &= \frac{Bm_2}{2r\sin\theta}. \end{aligned}\right\} \tag{7}$$

The pole-sensitivity function for Structure-II denoted as S_{II} is obtained from (1) and (7). For any pole location $p = re^{\pm j\theta}$, the optimum set of the integer constants B, C, and E is identified using an exhaustive search algorithm as described earlier to minimize S_{II}.

Since for every pole location, we have a choice of two structures, namely, Structures-I and II, the optimum structure among these two is selected as the one having a lower value for its pole-sensitivity function.

III. Classification Map

The selection of the optimum structure is completely dependent on the pole location and hence the z-plane can be partitioned into several regions such that for each region, a single optimum structure is chosen. Such a classification map is shown in Fig. 2. The other quadrants can be derived from symmetry. Table I gives the optimum structure along with the values of the integer constants for the different regions of the z-map.

IV. Sensitivity Comparison

Consider the following band-pass transfer function:

$$H(z) = \frac{0.0024731(z^2 - 1)}{z^2 - 0.379127z + 0.9950537} \tag{8}$$

with $r = 0.9975237$ and $\theta = 79.045$ deg. The optimum structure obtained for $H(z)$ using tables I and VI of [1] is as follows:

$$\text{Structure-I with } C = 0,\ D = 0, \quad \text{and} \quad E = 1. \tag{9}$$

From (4) and (9), we get

$$m_1 = 0.379127 \quad \text{and} \quad m_2 = 0.0049463 \tag{10}$$

For the pole-sensitivity approach, using the exhaustive search algorithm, the optimum sets of Structures-I and II are identified

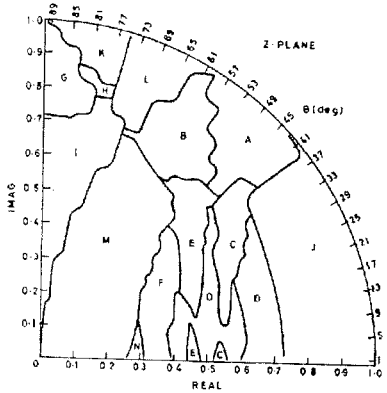


Fig. 2. Classification map.

TABLE I
VALUES OF INTEGER CONSTANTS FOR DIFFERENT REGIONS

Region	Structure	Values of Integer Constants C / B	D / C	E for Structure I / F for Structure II
A	I	1	0	1
B	I	0	1	1
C	I	−1	2	2
D	I	0	1	0
E	I	2	1	2
F	I	1	0	0
G	I	1	−1	2
H	I	0	0	1
I	I	−1	1	1
J	II	1	1	0
K	II	−1	−1	2
L	II	1	1	1
M	II	1	0	1
N	II	1	0	2
O	II	1	0	0

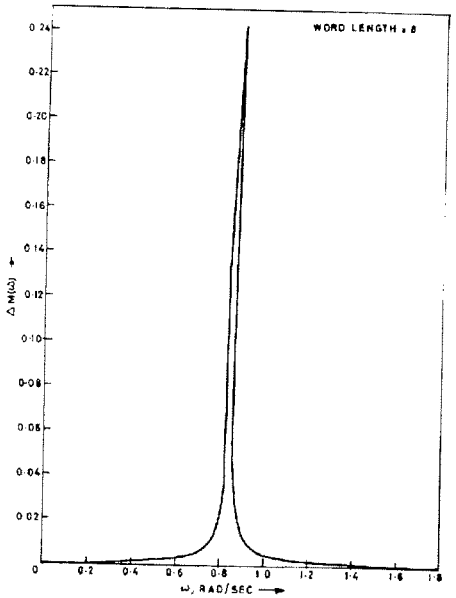


Fig. 3. Error curve for wordlength 8.

as follows.

For Structure-I: $C = 0$, $D = 0$ and $E = 1$

and

$$m_1 = 0.379127 \quad \text{and} \quad m_2 = -0.0049463 \tag{11a}$$

For Structure-II:

$$B = -1,\ C = -1, \quad \text{and} \quad E = 2$$

and

$$m_1 = 0.0049464 \quad \text{and} \quad m_2 = 0.37418064 \tag{11b}$$

The values of the pole–sensitivity function for the above cases are obtained as

$$S_{\mathrm{I}} = 0.0374723 \quad \text{and} \quad S_{\mathrm{II}} = 0.0365099. \tag{12}$$

Thus Structure-II qualifies to be the optimum structure in the pole sensitivity approach. The rejected structure in this approach viz. Structure-I of (11a) is the same as the one obtained using the procedure of [1]. Note that one can obtain the optimum structure in the pole–sensitivity approach directly from the classification map of Fig. 2. Since the given values of r and θ correspond to the region K, we get the optimum structure of equation (11b) from Table I directly.

The error in the amplitude response $\Delta M(\omega)$ defined as in (13) is computed for both these structures for different wordlengths.

$$\Delta M(\omega) = M(\omega) - M_Q(\omega) \tag{13}$$

where $M_Q(\omega)$ and $M(\omega)$ are the amplitude responses of the structure with and without coefficient quantization, respectively. The plot of $\Delta M(\omega)$ versus ω shown in Fig. 3 represents the error curves of both the structures for wordlength 8. It is found that even for wordlengths 10 and 12, the error curves of both the structures are identical. This is because of the fact that the values of S_{I} and S_{II} in (12) do not differ much.

V. CONCLUSIONS

A pole–sensitivity based approach for identifying the optimum second-order digital filter structures which are amenable to FSS has been described. This approach eliminates the need for a separate sensitivity analysis. A classification map containing several regions has been presented which enables one to identify easily the optimum structure corresponding to any given pole location. The sensitivity comparison carried out for a specific transfer function yielded an interesting result, namely that though the optimum structures obtained by the proposed method and that of [1] are completely different, the sensitivity characteristics are found to be identical.

REFERENCES

[1] Paulo S. R. Diniz and Andreas Antoniou, "Low-sensitivity digital filter structures which are amenable to error-spectrum shaping," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 1000–1007, Oct. 1985.

[2] Akinori Nishihara and Kazunori Sugahara, "A synthesis of digital filters with minimum pole sensitivity," *Trans. IECE Japan*, vol. E 65, no. 5, pp. 234–240, May 1982.

[3] Akinori Nishihara, "Low-sensitivity second-order digital filters – Analysis and design in terms of frequency sensitivity," *Trans. IECE Japan*, vol. E 67, no. 8, pp. 433–439, Aug. 1984.

Polynomial Reduction in the Analysis of Active-Filter Parasitics

P. BOWRON, A. P. O'CARROLL, AND A. A. DAABOUL

Abstract — **The exact Tarmy–Ghausi technique for polynomial reduction is extended to any order in terms of both combinatorial and recursive formulae. Several practical advantages accrue including an accurate method for predistortion. These features are verified by numerical root finding and by SPICE for a higher-frequency design example of a single-amplifier active filter.**

I. INTRODUCTION

Active filters are customarily designed as second-order sections for incorporation in multi-stage implementations. However, the presence of active and passive parasitics in practical circuits causes the voltage transfer function to be of higher order. As a result, response performance can deviate from the nominal specification, especially in higher frequency designs. In order to gauge the influence on the standard second-order parameters ω_0 (center/critical radian frequency) and Q (selectivity), it is necessary to reduce the denominator to quadratic form.

Several means for deriving the effective coefficients and consequent perturbed parameters have been suggested [1] usually with respect to the finite gain bandwidth product (ω_t) of the operational amplifiers. The most popular techniques have been based on extracting the dominant second-order factor and comparing factorized coefficients. This has usually been carried out only for third-order denominators [2]–[4], sometimes merely by disregarding higher powers and replacing the s^3 term by an approximately equivalent s term [5], [6]. Fourth-order reductions have also been reported [7]–[9] to cater for two-amplifier configurations and amplifier second pole (ω_2). The approach appears [9] to stem from Friedman's method [10] but has usually been applied under approximate first-order conditions to give results valid only for high Q and small perturbations [11], [12]. Only the Tarmy–Ghausi relations [1] are exact and convenient to apply [13]–[15] in higher frequency situations where parameter deviations are more pronounced.

Alternative techniques include directly evaluating classical sensitivities [16], locating undesired poles by zeroing desired pole derivatives [17]–[19] and considering shifts in dominant singularities [20]. However, none are so analytically tractable as the Tarmy–Ghausi method which is now extended from its original fifth-order statement [1] to the general order.

II. ALGEBRAIC FORMULATION

Consider an active network subject to many linear parasitic effects so that the denominator of its transfer function can be written as an nth-order polynomial:

$$D(s) = b_n s^n + b_{n-1} s^{n-1} + \cdots + b_1 s + 1. \tag{1}$$

If the effective perturbed second-order parameters are center

Manuscript received December 3, 1987; revised August 8, 1988. This paper was recommended by Associate Editor J. Mavor.

P. Bowron is with the Department of Electrical Engineering, University of Bradford, Bradford, West Yorkshire BD7 1DP, U.K.

A. P. O'Carroll is with Luptronics Ltd., Pudsey, West Yorkshire LS28 6HR, U.K.

A. A. Daaboul is with Kemo Ltd., Beckenham, Kent BR3 3QZ, U.K.

IEEE Log Number 8927743.

0098-4094/89/0700-1020$01.00 ©1989 IEEE