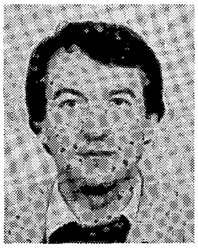
implement algorithms on single-chip digital signal processors," in *Proc. Second European Signal Proc. Conf.*, EUROSIP '83, pp. 851–854, (Erlangen, W. Germany), Sept. 1983.

[50] R. H. Cushman, "Signal-processing design awaits digital take-over," EDN Mag., pp. 119–128, June 24, 1981.

[51] L. Gazsi, "Single-path transmultiplexer scheme with multirate wave digital filters," in *Proc. IEEE Int. Conf. Communications* (ICC '84), pp. 675–678, (Amsterdam), May 1984.



Lajos Gazsi (M'83–SM'84) was born in Kaposvár, Hungary, on February 17, 1942. He received the Diploma and the Doctor Techn. degree in electrical engineering from the Technical University of Budapest, Hungary, in 1964 and 1974, respectively, and the Candidate of Technical Sciences degree in circuit theory from the Hungarian Academy of Sciences, Budapest, in 1974.

He worked in the Department of Measurement and Instrument, Technical University of Budapest, Hungary, from 1964. In 1974, he spent a year on leave, with A. Fettweis in the Department of Electrical Engineering, Ruhr-University Bochum, West Germany. He has been employed there since 1977, working mainly in the field of digital signal processing. In addition to the theoretical work he has been taking part in the design of a sophisticated digital signal processor that is being developed at Siemens Company in Munich, West Germany, in 1983. Furthermore, he is participating in the ESPRIT project "European Strategic Programme for Research and Development in Information Technology" sponsored by the European Community. The prime contractor is Professor H. DeMan from the ESAT Laboratory, Katholieke Universiteit, Leuven, Belgium.

Dr. Gazsi is a member of the European Association for Signal Processing (EURASIP).

Transactions Briefs

On Error-Spectrum Shaping in State-Space Digital Filters

P. P. VAIDYANATHAN

Abstract —**A new scheme for shaping the error spectrum in state-space digital filter structures is proposed. The scheme is based on the application of diagonal second-order error feedback, and can be used in any arbitrary state-space structure having arbitrary order. A method to obtain noise-optimal state-space structures for fixed error feedback coefficients, starting from noise optimal structures in absence of error feedback (the Mullis and Roberts Structures), is also outlined. This optimization is based on the theory of continuous equivalence for state-space structures.**

I. Introduction

The use of error spectrum shaping (ESS) for roundoff noise reduction in (narrow band) recursive digital filters is well known, and a number of interesting research contributions in this area have appeared [1]–[5] in the last few years. The application of this idea to state-space structures is mentioned in [2], and some studies in this connection have already been reported in [5]. In [7], Mullis and Roberts clarify the relation between error feedback (EFB) techniques and double precision implementations in state-space structures, among others.

The purpose of this paper is to outline a new procedure for choice of EFB coefficients in state-space structures. Specifically, we extend the feedback scheme proposed in [2] and [5] by incorporating an additional higher order matrix term. We do not consider error feedforward in this paper. We choose the EFB coefficients such that each noise source is "shaped" independent of others. In the resulting structures, each noise source is essentially replaced by an equivalent source which is no more white, but has zeros on the unit circle of the z-plane, at suitable locations. Consequently the major portion of noise power moves into the stopband. The overall noise is thus reduced, by the time it reaches the filter output. (It should be noticed that the idea of introducing zeros into the noise spectrum is itself not new, and is indeed the basis in [1].)

In such a structure, we essentially have "colored" noise sources, and the optimal state space structure for a given noise-spectral shaping is in general different from that for white noise sources. Based on the fundamental results on minimum-noise state-space structures [6] for uncorrelated white noise sources, we outline an iterative procedure for arriving at the minimum-noise structure with fixed EFB. The procedure is based on applying a sequence of similarity transformations in such a way that at each iteration there is an improvement in the objective function.

In Section II we deal with the shaping of error spectrum for a given state-space structure. In Section III, we outline the state-space optimization for a given ESS shaping.

II. Noise Shaping

Consider the standard state-space representation:

$$\boldsymbol{x}(n+1)=\boldsymbol{A}\boldsymbol{x}(n)+\boldsymbol{B}u(n) \tag{1a}$$

$$y(n)=\boldsymbol{C}x(n)+\boldsymbol{D}u(n). \tag{1b}$$

Here $\boldsymbol{A}$ is an $N\times N$ matrix, $\boldsymbol{B}$ is $N\times 1$, $\boldsymbol{C}$ is $1\times N$, and $\boldsymbol{D}$ is 1×1. We assume that the only quantization involved is in the implementation of $\boldsymbol{A}\boldsymbol{x}(n)$, as this is the only error that propagates through the feedback path. Fig. 1 shows the conventional EFB scheme, where the error vector due to the vector quantizer is fedback through a delay (to avoid delay free loops). The matrix

Manuscript received September 13, 1983; revised March 1, 1984. This work was supported in part by the National Science Foundation under Grant ECS82-18310 and in part by Caltech funds.

The author is with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125.

0098-4094/85/0100-0088$01.00 ©1985 IEEE

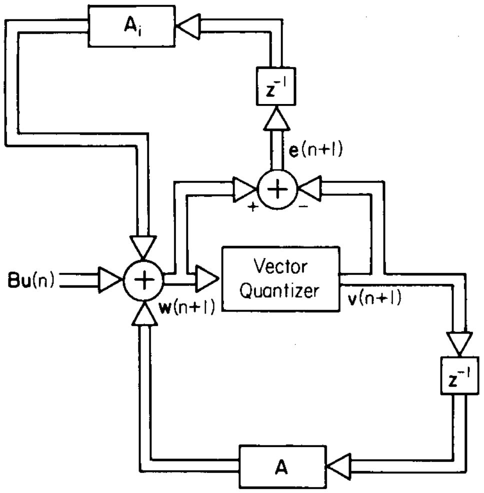


Fig. 1. Error feedback in a state-space structure.

A_i, is typically chosen to have integer elements, so that there is no additional roundoff error. Even if this restriction is partially relaxed, ESS schemes still lead to considerable noise-reduction, as the additional roundoff error is only at a secondary level.

Let us now consider the roundoff noise propagation under zero input conditions. If we regard the quantizer output $\boldsymbol{v}(n)$ to be the state vector, we obtain the following equations:

$$\boldsymbol{x}(n+1)=\boldsymbol{A}\boldsymbol{x}(n)+\boldsymbol{A}_i\boldsymbol{e}(n)-\boldsymbol{e}(n+1) \tag{2a}$$

$$y(n)=\boldsymbol{C}\boldsymbol{x}(n) \tag{2b}$$

whereas, if the quantizer input is regarded as the state vector, we get

$$\boldsymbol{x}(n+1)=\boldsymbol{A}[\boldsymbol{x}(n)-\boldsymbol{e}(n)]+\boldsymbol{A}_i\boldsymbol{e}(n) \tag{3a}$$

$$y(n)=\boldsymbol{C}\boldsymbol{x}(n). \tag{3b}$$

Applying z-transformation, (2a) and (3a) yield, respectively,

$$\boldsymbol{X}(z)=(z\boldsymbol{I}-\boldsymbol{A})^{-1}(\boldsymbol{A}_i-z\boldsymbol{I})\boldsymbol{E}(z) \tag{4}$$

$$\boldsymbol{X}(z)=(z\boldsymbol{I}-\boldsymbol{A})^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{E}(z). \tag{5}$$

The difference between the quantities $\boldsymbol{v}(n)$ and $\boldsymbol{w}(n)$ is negligible, being just equal to the basic quantizer error. However, (4) reveals explicitly the "shaping" of error spectrum, whereas (5) does not. The former is, therefore, more useful in judging the choice of EFB coefficients.

The conventional method for choice of $\boldsymbol{A}_i$ is to make each of its elements equal to an integer that is nearest to the corresponding element of $\boldsymbol{A}$. In this way, no secondary roundoff error is introduced, but the EFB circuit can get complicated, particularly with general full matrices, and the complexity is proportional to N^2. From (5), it appears that the choice $\boldsymbol{A}_i=\boldsymbol{A}$ leads to zero roundoff noise, but this is essentially equivalent to double precision arithmetic, in which case, the "secondary" error is the only source of error. The usefulness of EFB techniques stems from the fact that a tradeoff between single and double precisions can be achieved which considerably improves performance, and if the EFB network is chosen carefully, is simpler to implement than the double precision schemes.

A different guideline for the choice of EFB coefficients is to make the $\boldsymbol{A}_i$ matrix diagonal:

$$\boldsymbol{A}_i=\operatorname{diag}[a_{kk}^{(i)}]. \tag{6}$$

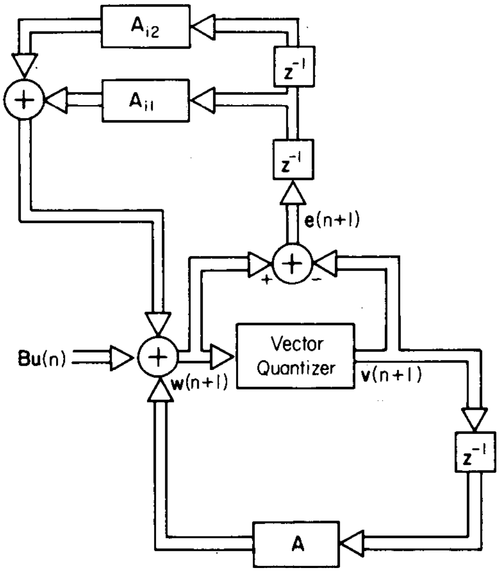


Fig. 2. Extension of scheme of Fig. 1.

Then the kth state equation becomes, from (2a):

$$x_k(n+1)=\sum_{j=1}^{N}a_{kj}x_j(n)+a_{kk}^{(i)}e_k(n)-e_k(n+1). \tag{7}$$

Thus the kth noise source is shaped by a transfer function with a real zero at $z=a_{kk}^{i}$. For low-pass filters, if $a_{kk}^{(i)}$ are chosen to be unity, this places zeros in the noise spectra at the passband frequency $\omega=0$. Thus for narrow-band lowpass filters the choice $\boldsymbol{A}_i=\boldsymbol{I}$ and similarly, for narrow-band highpass filters the choice $\boldsymbol{A}_i=-\boldsymbol{I}$, has the effect of reducing the overall noise. For band elimination filters, we could choose

$$\boldsymbol{A}_i=\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{0} & -\boldsymbol{I}\end{bmatrix}. \tag{8}$$

For filters that are not particularly narrow band, and for bandpass filters, it is more advantageous to introduce complex transmission zeros into the error spectrum. In order to do this, let us consider the EFB configuration of Fig. 2. If we regard the quantizer output $\boldsymbol{v}(n)$ to be the state vector, we obtain the following equations:

$$\boldsymbol{x}(n+1)=\boldsymbol{A}\boldsymbol{x}(n)+\boldsymbol{A}_{i1}\boldsymbol{e}(n)+\boldsymbol{A}_{i2}\boldsymbol{e}(n-1)-\boldsymbol{e}(n+1) \tag{9}$$

whereas, if the quantizer input $\boldsymbol{w}(n)$ is regarded as the state vector, we get

$$\boldsymbol{x}(n+1)=\boldsymbol{A}[\boldsymbol{x}(n)-\boldsymbol{e}(n)]+\boldsymbol{A}_{i1}\boldsymbol{e}(n)+\boldsymbol{A}_{i2}\boldsymbol{e}(n-1). \tag{10}$$

Applying z-transformation, (9) and (10) yield, respectively,

$$\boldsymbol{X}(z)=(z\boldsymbol{I}-\boldsymbol{A})^{-1}(\boldsymbol{A}_{i2}z^{-1}+\boldsymbol{A}_{i1}-z\boldsymbol{I})\boldsymbol{E}(z) \tag{11}$$

$$\boldsymbol{X}(z)=(z\boldsymbol{I}-\boldsymbol{A})^{-1}[(\boldsymbol{A}_{i1}-\boldsymbol{A})+z^{-1}\boldsymbol{A}_{i2}]\boldsymbol{E}(z). \tag{12}$$

From (11) it is clear that if we choose

$$\boldsymbol{A}_{i2}=-\boldsymbol{I} \quad\text{and}\quad \boldsymbol{A}_{i1}=\operatorname{diag}[\alpha_k] \tag{13}$$

then the effective "spectrally shaped" error source at the kth state is

$$E_k^{ss}(z)=E_k(z)[1-\alpha_k z^{-1}+z^{-2}]. \tag{14}$$

In order to introduce a transmission zero at $z=e^{\pm j\omega_k}$, α_k should be chosen as $\alpha_k=2\cos(\omega_k)$.

The ω_k's should be chosen at suitable points in the passband, such that the binary representations of α_k's are simple. We consider the number of nonzero bits in the code to be the complexity of a multiplier. (See Section V). We also assume SD code representation for EFB coefficients, so that, for example, "101111" has a complexity of 3, as it can be implemented as "11000-1". With two-bit complexities, a satisfactory zero-positioning can generally be achieved. A constrained optimization scheme for this purpose can be adopted, but the details are beyond the scope of this paper.

Note that, the implementation of A_{i2} does not require multipliers. For filters with passband centered around $\pi/2$ or $\pm\pi/3$, the entire EFB network is free from multipliers.

An obvious extension of the above scheme is the use of still higher order EFB. For example, the noise sources can be shaped by a transversal filter as follows:

$$E_k^{ss}(z) = E_k(z)[1+z^{-1}+z^{-2}+\cdots+z^{-L+1}]. \qquad (15)$$

Such an EFB network introduces zeros at frequencies $\omega_k = 2\pi k/L$, except at $k=0$. However, we do not always continue to get improvements in this fashion, because even though the EFB circuit introduces zeros at certain points, it begins to have a "gain" exceeding unity over a considerable region of the passband.

III. Performance Measure and Numerical Examples

Even though (9) gives us a guideline for choice of the EFB circuit, we find it more convenient to use (10) for deriving a performance measure. (Recall that, there is hardly any difference between $v(n)$ and $w(n)$ anyway). From (10) we get

$$x(n+1) = Ax(n) + (A_{i1} - A)e(n) + A_{i2}e(n-1). \qquad (16)$$

In (16), $e(n)$ is the quantizer error vector. We assume that each component is zero-mean white, and that the components are uncorrelated. We can then derive the covariance of the state vector in response to the quantizer error to be

$$E[x(n)x^t(n)] = \sigma_e^2[V_{11}+V_{12}+V_{21}+V_{22}] = \sigma_e^2 V \qquad (17)$$

where

$$V_{11} = \sum_{k=0}^{\infty} A^k(A_{i1}-A)(A_{i1}-A)^t(A^t)^k \qquad (18a)$$

$$V_{22} = \sum_{k=0}^{\infty} A^k(A_{i2})(A_{i2})^t(A^t)^k \qquad (18b)$$

$$V_{12} = \sum_{k=0}^{\infty} A^k(A)(A_{i1}-A)A_{i2}^t(A^t)^k \qquad (18c)$$

$$V_{21} = V_{12}^t. \qquad (18d)$$

The quantity CVC^t, which is proportional to the output noise power, will be taken as the performance measure, in the numerical example to follow. Clearly, smaller this quantity, the better the performance.

Examples

Consider implementing a state-space structure with

$$A = \begin{bmatrix} -6.24 & 5.76 & 5.2 \\ -6.76 & 6.24 & 4.8 \\ -1.483 & 1.282 & 1.97 \end{bmatrix} \qquad (19)$$

$$C = [0.0577 \quad -0.0578 \quad 0.0793]. \qquad (20)$$

The "all pole filter" corresponding to the matrix A is

$$H(z) = \frac{1}{1-1.97z^{-1}+1.56z^{-2}-0.454z^{-3}} \qquad (21)$$

and a plot of magnitude of $H(e^{j\omega})$ is shown in Fig. 3. It is clear, therefore, that if the EFB path introduces zeros in the range $(0, 0.25\pi)$, we can expect a considerable noise-reduction. For example, let us choose

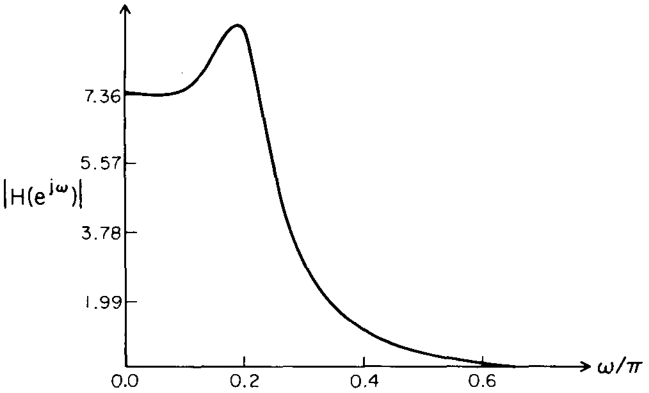


Fig. 3. Plot of magnitude of $H(e^{j\omega})$ of (21).

$$A_{i1} = \begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 1.875 & 0 \\ 0 & 0 & 2 \end{bmatrix} \qquad A_{i2} = -I. \qquad (22)$$

Thus the noise sources $e_1(n), e_2(n), e_3(n)$, face transmission zeros at $\omega = 0.23\pi$, 0.113π, and 0.0, respectively. The quantity CVC^t is calculated to be 0.10261 for this EFB scheme. With no error fedback at all, (i.e., with A_{i1} and $A_{i2} = 0$) the performance measure is 0.5905. Conventional EFB technique would suggest the use of

$$A_{i1} = \begin{bmatrix} -6 & 6 & 5 \\ -7 & 6 & 5 \\ -1 & 1 & 2 \end{bmatrix} \qquad A_{i2} = 0 \qquad (23)$$

which leads to a performance measure of 0.155. Note also that (23) is considerably more complex to implement, though it has the advantage of generating no secondary roundoff noise. If the total number of bits required to represent the EFB matrices in cannonical SD code is taken to be a complexity measure, the complexity of the new scheme (22) is 5 bits, whereas that of the conventional scheme (23) is 15 bits. Thus the performance measure is improved, at the same time reducing the complexity. In addition, note that in (23), most of the feedback gains are greater than unity, whereas in (22), this is not so, and this is an added advantage.

Next, the following EFB network was considered:

$$A_{i1} = 1.625I \qquad A_{i2} = -I \qquad (24)$$

The resulting performance measure is about 0.0805, which is nearly half as much as the conventional first-order EFB scheme. Each of the multipliers of value 1.625 introduces zeros at $\omega = 0.198106\pi$ for the corresponding noise source, and the complexity is only 9 bits (3 per multiplier).

Next, a third-order EFB scheme was tried. Three matrices:

$$A_{i1} = 2.625I \qquad A_{i2} = -2.625I \qquad A_{i3} = I \qquad (25)$$

were used, introducing zeros at $\omega = 0.198106\pi$ and $\omega = 0.0$. The resulting performance measure however was larger compared to the scheme of (24), indicating that we begin to get decreasing returns, as the EFB network has an overall gain exceeding unity, for a considerable region in the passband.

From (12) it appears that, if A_{i1} and A_{i2} are chosen, respectively, to be as close to $A+I$ and $-A$ as possible, the performance measure would improve dramatically. This is indeed true, but leads to complicated EFB networks.

IV. Noise Optimal State-Space Structures, with Fixed Error Feedback

Given a EFB circuit for a state-space configuration, such that the noise generated at the kth state is shaped by the function:

$$1-2\cos(\omega_k)z^{-1}+z^{-2} \tag{26}$$

we wish to find the optimal state-space structure. When the noise sources are white, i.e., when the function of (26) is replaced by unity, this leads to the well-known Mullis and Roberts structure [6]. In this section we outline an iterative procedure for arriving at the optimal structure, starting from a suitable initial configuration, such as the Mullis and Roberts form. Our method is based on the continuous equivalence approach, as adopted to state-space structures [8].

Given a state-space realization, let $f_i(n)$ denote the impulse response from the input to the ith register, and define $\boldsymbol{K}$ as in [6]:

$$\boldsymbol{K}=\sum_{k=0}^{\infty}(\boldsymbol{A}^k\boldsymbol{B})(\boldsymbol{A}^k\boldsymbol{B})^t. \tag{27}$$

The diagonal elements K_{ii} represent the l_2 norm of $f_i(n)$. Next, let $g_i(n)$ denote the impulse response from the ith register to the filter output, with $G_i(e^{j\omega})$ denoting its Fourier transform. Let $\boldsymbol{G}(e^{j\omega})$ denote the vector:

$$\boldsymbol{G}(e^{j\omega})=\left[G_1(e^{j\omega}),G_2(e^{j\omega}),\cdots,G_N(e^{j\omega})\right]^t. \tag{28}$$

Let us define

$$P_{ii}=\int_0^{2Gp}\left[\boldsymbol{G}(e^{j\omega})\boldsymbol{G}^t(e^{-j\omega})\right]_{ii}S_i(\omega)\frac{d\omega}{2\pi}. \tag{29}$$

Here $[\boldsymbol{G}(e^{j\omega})\boldsymbol{G}^t(e^{-j\omega}]_{ii}$ denotes the ith diagonal element of the $N\times N$ matrix $[\boldsymbol{G}(e^{j\omega})\boldsymbol{G}^t(e^{-j\omega})]$ and $S_i(\omega)$ is given by

$$S_i(\omega)=4\left[\cos^2(\omega)+\cos^2(\omega_i)-2\cos(\omega)\cos(\omega_i)\right]. \tag{30}$$

Clearly the quantity $\Sigma_{i=1}^N P_{ii}$ represents a measure of the total output noise variance, assuming that each noise component of the vector quantizer has same variance. Now consider a similarity transformation of the form

$$\boldsymbol{A}^*=\boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T},\quad \boldsymbol{B}^*=\boldsymbol{T}^{-1}\boldsymbol{B},\quad \boldsymbol{C}^*=\boldsymbol{C}\boldsymbol{T},\quad \boldsymbol{K}^*=\boldsymbol{T}^{-1}\boldsymbol{K}\boldsymbol{T}^{-t}. \tag{31}$$

In particular, as the $\boldsymbol{K}$ matrix is transformed as shown above, the output noise variance of the scaled filter in terms of unscaled parameters is measured by

$$\Phi=\sum_{i=1}^{N}P_{ii}K_{ii}. \tag{32}$$

Now P_{ii} can be rewritten as

$$P_{ii}=4\left[\boldsymbol{M}+\cos^2(\omega_i)\boldsymbol{R}-2\cos(\omega_i)\boldsymbol{S}\right]_{ii} \tag{33}$$

where the $N\times N$ symmetric matrices $\boldsymbol{M},\boldsymbol{R},\boldsymbol{S}$, are defined by

$$\boldsymbol{M}=\frac{1}{2\pi}\int_0^{2\pi}\boldsymbol{G}(e^{j\omega})\boldsymbol{G}(e^{-j\omega})^t\cos^2\omega\,d\omega \tag{34a}$$

$$\boldsymbol{R}=\frac{1}{2\pi}\int_0^{2\pi}\boldsymbol{G}(e^{j\omega})\boldsymbol{G}(e^{-j\omega})^t\,d\omega \tag{34b}$$

$$\boldsymbol{S}=\frac{1}{2\pi}\int_0^{2\pi}\boldsymbol{G}(e^{j\omega})\boldsymbol{G}(e^{-j\omega})^t\cos\omega\,d\omega. \tag{34c}$$

Note that $\boldsymbol{R}$ is the $\boldsymbol{W}$-matrix defined in [6]. Note also that all these matrices are symmetric. It is easily verified that the similarity transformation affects the matrices as follows:

$$\boldsymbol{M}^*=\boldsymbol{T}^t\boldsymbol{M}\boldsymbol{T},\quad \boldsymbol{R}^*=\boldsymbol{T}^t\boldsymbol{R}\boldsymbol{T},\quad \boldsymbol{S}^*=\boldsymbol{T}^t\boldsymbol{S}\boldsymbol{T}. \tag{35}$$

The proposed iterative scheme applies a transformation of the form

$$\boldsymbol{T}=\boldsymbol{I}+\Delta x\boldsymbol{Q} \tag{36}$$

at each stage of the iteration. Here x is a dummy continuous variable of iteration, and $\boldsymbol{Q}$ is to be chosen such that the objective function Φ decreases in a "steepest descent fashion." The gradient of the objective function is

$$\frac{d\Phi}{dx}=\sum_{i=1}^{N}\frac{dP_{ii}}{dx}K_{ii}+\frac{dK_{ii}}{dx}P_{ii}. \tag{37}$$

This can be evaluated by noting that the derivatives involved can be found in the following manner: from (33),

$$\frac{dP_{ii}}{dx}=4\left[\frac{dM_{ii}}{dx}+\cos^2(\omega_i)\frac{dR_{ii}}{dx}-2\cos(\omega_i)\frac{dS_{ii}}{dx}\right]. \tag{38}$$

Next,

$$\boldsymbol{M}(x+\Delta x)=\boldsymbol{T}^t\boldsymbol{M}(x)\boldsymbol{T}=(\boldsymbol{I}+\Delta x\boldsymbol{Q})^t\boldsymbol{M}(x)(\boldsymbol{I}+\Delta x\boldsymbol{Q}). \tag{39}$$

This leads to

$$\frac{d\boldsymbol{M}}{dx}=\boldsymbol{Q}^t\boldsymbol{M}+\boldsymbol{M}\boldsymbol{Q} \tag{40}$$

whence

$$\frac{dM_{ii}}{dx}=2\sum_{j=1}^{N}Q_{ji}M_{ji}. \tag{41}$$

The derivatives of $\boldsymbol{R}$ and $\boldsymbol{S}$ take similar forms. Eventually we get

$$\frac{dP_{ii}}{dx}=8\left[\sum_{j=1}^{N}Q_{ji}M_{ji}+\cos^2\omega_i\sum_{j=1}^{N}Q_{ji}R_{ji}-2\cos\omega_i\sum_{j=1}^{N}Q_{ji}S_{ji}\right]. \tag{42}$$

Similarly, we can show

$$\frac{dK_{ii}}{dx}=-2\sum_{j=1}^{N}Q_{ij}K_{ij} \tag{43}$$

whence

$$\frac{d\Phi}{dx}=-8\sum_{i=1}^{N}\sum_{j=1}^{N}Q_{ij}\Theta_{ij} \tag{44}$$

where Θ_{ij} is given by

$$\Theta_{ij}=M_{ii}K_{ij}-M_{ij}K_{jj}+(R_{ii}K_{ij}-R_{ij}K_{jj})\cos^2(\omega_i)-2(S_{ii}K_{ij}-S_{ij}S_{jj})\cos(\omega_i). \tag{45}$$

The quantity $\boldsymbol{Q}$ is chosen at each iteration according to the following rule:

$$Q_{ij}=\operatorname{sign}\left[\Theta_{ij}\right]. \tag{46}$$

This ensures that the derivative of the objective function (44) is as negative as possible, for a given Δx.

Summarizing, given an initial state-space realization, we calculate the matrices $\boldsymbol{K}$ and $\boldsymbol{R}$. As these matrices satisfy the Lyapunov matrix equation, they can be obtained by efficient algorithms as described in [6]. It can be shown that the matrices $\boldsymbol{M}$ and $\boldsymbol{S}$ satisfy the following equations which look essentially like the Lyapunov equation:

$$\boldsymbol{M}=\boldsymbol{A}^t\boldsymbol{M}\boldsymbol{A}+0.5\left[\boldsymbol{C}^t\boldsymbol{C}+0.5\boldsymbol{C}^t\boldsymbol{C}\boldsymbol{A}^2+0.5(\boldsymbol{A}^2)^t\boldsymbol{C}^t\boldsymbol{C}\right] \tag{47}$$

$$\boldsymbol{S}=\boldsymbol{A}^t\boldsymbol{S}\boldsymbol{A}+0.5[\boldsymbol{C}^t\boldsymbol{C}\boldsymbol{A}+\boldsymbol{A}^t\boldsymbol{C}^t\boldsymbol{C}]. \tag{48}$$

Once these matrices are calculated for an initial state-space realization, we can then evaluate the objective function Φ, and the quantity Θ_{ij}. This then determines the right choice of $\boldsymbol{Q}$ according to (46), and the transformation is now made. This corresponds to updating the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{K}$ according to (31), and updating $\boldsymbol{M}$ according to

$$\boldsymbol{M}(x+\Delta x) = \boldsymbol{M}(x) + \Delta x\left[\boldsymbol{Q}^t\boldsymbol{M}(x) + \boldsymbol{M}^t\boldsymbol{Q}(x)\right] + (\Delta x)^2 \boldsymbol{Q}^t\boldsymbol{M}\boldsymbol{Q}. \quad (49)$$

$\boldsymbol{R}$ and $\boldsymbol{S}$ can be updated in a similar fashion as $\boldsymbol{M}$. The process can be repeated until there is negligible change in the objective function. Note that local minima are possible, but if the initial configuration is chosen as the optimal structure for "white noise" [6], the iteration definitely improves the performance because Φ decreases at each iteration if Δx is sufficiently small. The choice of Δx can be made according to standard intuitive guidelines.

V. Concluding Remarks

Certain implementations of digital filters involve an architecture, where the number of bits per multiplier (and signal) primarily determines the complexity. This is so, for example if the multiplier is implemented by coding shift/add operations, as done in a typical INTEL 2920 type of filter implementation. For such applications, EFB techniques are much easier to incorporate than incorporating double precision, and the techniques described in this paper are most relevant. In addition, the number of nonzero bits in an EFB coefficient can be taken as a reasonable complexity measure for the EFB network, and, therefore, the use of cannonic SD code seems appropriate. On the other hand, if the filter architecture is such that a number of parallel multipliers (16-bit standard multipliers, for example) form the major building blocks, the incorporation of EFB might increase the overall architectural complexity, unless the EFB coefficients are very simple, as in [1]. For such an architecture, it might even be better to go for a double precision scheme.

It should be noted that the performance measure introduced in Section III ignores the secondary quantization errors, and can be misleading if the EFB network is highly complex. However, we trust that the choice of simple diagonal matrices as indicated in the examples, makes the measure reasonably accurate. The use of diagonal EFB networks introduced in this paper can be attractive in case of state matrices that are not sparse. The noise minimization scheme outlined in Section IV can be easily extended to arbitrary noise source spectra, that may arise by using still higher order diagonal EFB networks. Also, note that an obvious application of the ideas introduced here is in the implementation of block-digital filter structures [9], which involve block-quantizers.

References

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

[2] 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.

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

[4] 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. Acoustics, Speech, Signal Processing*, vol. ASSP-30, pp. 963–973, Dec. 1982.

[5] M. Renfors, "Roundoff noise in error-feedback state space filters," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 619–622, Boston, MA, Apr. 1983.

[6] 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.

[7] ——, "An interpretation of error spectrum shaping in digital filter structures," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-30, pp. 1013–1015, Dec. 1982.

[8] R. Devanathan, "Continuously equivalent networks in state space," *Electron. Lett.*, vol. 9, no. 16, Aug. 1973.

[9] S. K. Mitra and R. Gnanasekaran, "Block implementation of Recursive Digital filters—New structures and properties," *IEEE Trans. Circuits Syst.*, vol. CAS-25, Apr. 1978.

A Hybrid Floating-Point Logarithmic Number System Processor

FRED J. TAYLOR

Abstract —**The attributes of the traditional floating-point processor and the logarithmic number system are combined. The result is a hybrid system which offers some advantages over the familiar floating-point system. The new system, called the (FU)^2, does not require exponent alignment during addition, supports high-speed addition and multiplication, has an efficient accumulator structure, and admits a simple VLSI realization.**

Introduction

When a large dynamic range and high precision are required, a floating-point representation is often adopted. Although standardization is on the way, there are many floating-point formats currently in use today. However, the problem with any of these choices is that compared to fixed point, floating-point arithmetic operations are slow and complex. Furthermore, the time it takes to perform a floating-point addition can vary markedly depending upon the relative values of the data to be added. As a result, developing efficient real-time code, in floating point, can result in temporal inefficiency. In addition, the multiplier and addition floating paths are sufficiently different so as to demand (in most commercial hardware realizations) two or more separate hardware units. As a result the utilization of a hardware floating point unit can be as low as 50 percent.

In this work, a variation of the floating-point theme is presented. It possesses the precision and dynamic range of the floating-point system without the high overhead and reduced throughput (due to the exponent alignment requirement) of floating point addition. This new concept, shall be referred to as the *Fl*orida *Un*iversity *Fl*oating (Point) *Un*it or (FU)^2.

Floating-Point Format

In a floating-point system, a real number X can be approximated by

$$X = m_x r^{e_x} \quad (1)$$

where r is the radix, m_x is the signed $(M+1)$-bit mantissa, and e_x is the signed $(E+1)$-bit exponent. In this form, the precision

Manuscript received June 10, 1983; revised April 30, 1984. This work was supported by the National Science Foundation.

The author was with the University of Cincinnati, Cincinnati, OH. He is now with the Department of Electrical Engineering, University of Florida, Gainesville, FL 32611.

0098-4094/85/0100-0092$01.00 ©1985 IEEE