$$\boxed{\begin{aligned} S_c \approx D_\infty\left(\frac{\delta_P}{2}, \delta_s\right) \cdot \frac{f_r}{\Delta f} \cdot \frac{1}{2^{P+1}} + \frac{1}{2} \cdot D_\infty(\delta_0, \delta_0) \\ \cdot \left[P + k \cdot \left(\frac{1}{1-k} + \frac{1}{2-k}\right)\right] \end{aligned}}$$

with $k = 2^P \cdot 2f_s/f_r$.

The number of storage locations for data is

$$\boxed{\begin{aligned} S_D \approx D_\infty\left(\frac{\delta_P}{2}, \delta_s\right) \cdot \frac{f_r}{\Delta f} \cdot \frac{1}{2^{P}} + 2 \cdot D_\infty(\delta_0, \delta_0) \\ \cdot \left[P + k \cdot \left(\frac{1}{1-k} + \frac{1}{2-k}\right)\right]. \end{aligned}}$$

These results will now be applied to examples.

VI. Examples

Two realistic examples presented in [3] will be considered to illustrate the preceding results.

Example 1: Let the specifications of a narrow-band lowpass filter be

$$f_r = 1; f_s = 0.05; \Delta f = 0.025; \delta_P = 0.01; \delta_s = 0.001.$$

Then the parameters take on the following values:

$$P = 3; \delta_0 = \min\left\{\frac{0.01}{12}, 0.001\right\} = 0.00083; D_\infty(\delta_0, \delta_0) = 3.3$$

$$D_\infty\left(\frac{\delta_P}{2}, \delta_s\right) = 2.76; k = 0.8; S(k) = 1.6.$$

The multiplication rate is

$$R_M = 3.3 \times 1.6 + \tfrac{1}{128} \times 40 \times 2.76 = 6.2 \text{ multiplications/s.}$$

The number of storage locations for coefficients is

$$\begin{aligned} S_c &= 2.76 \times 40 \times \tfrac{1}{16} + \tfrac{1}{2} \times 3.03 \times [3 + 0.8\,(\tfrac{1}{0.2} + \tfrac{1}{1.2})] \\ &\approx 20. \end{aligned}$$

The number of storage locations for data is

$$\begin{aligned} S_D &= 2.76 \times 40 \times \tfrac{1}{8} + 2 \times 3.3 \times [3 + 0.8\,(\tfrac{1}{0.2} + \tfrac{1}{1.2})] \\ &\approx 65. \end{aligned}$$

Straight direct form implementation for the filter would lead to 55 multiplications/s, 55 storage locations for coefficients, and 110 locations for data.

For the two-stage scheme of [3], implemented as in [4] but with the second decimation stage and first interpolation stage regrouped because the last reduction ratio is 2, the results are $R_M = 10.4$, $S_c = 26$, and $S_D = 57$.

Example 2: Let the specifications of a very narrow-band filter be

$$f_r = 1; f_s = 0.005; \Delta f = 0.00025; \delta_P = 0.001; \delta_s = 0.0001.$$

The parameters' values are

$$P = 6; \delta_0 = \min\left\{\frac{0.001}{24}, 0.0001\right\} = 0.0000416;$$

$$D_\infty(\delta_0, \delta_0) = 5.38; D_\infty\left(\frac{\delta_P}{2}, \delta_s\right) = 4.72; k = 0.64;$$

$$S(k) = 1.07.$$

The multiplication rate is $R_M = 7.76$ multiplications/s, the number of storage locations for coefficients is $S_c \approx 152$, the number of storage locations for data is $S_D \approx 350$.

The three-stage scheme of [2] has again a last reduction ratio of 2, and then it is an advantage to merge the last reduction stage with the first interpolation stage. The results are as follows: $R_M = 12.38$; $S_c = 225$; $S_D = 464$. As a remark on this second example, it must be pointed out that when P is large, the estimation of expression (2) can be significantly in excess for the first stages; thus, the estimation of R_M is on the high side for these cases.

VII. Discussion

Expressions have been derived which give simple and accurate estimates of the computation rate and the volume of storage in multirate filtering with half-band FIR filters. The results obtained on two typical examples show an advantage, particularly in computation rate, of the half-band filter approach over the general optimal design. This is due to the particular structure of the half-band filter.

However, in practice, after the arithmetic unit, coefficient and data memories, a fourth subset has to be considered, the control unit. It is, in general, more complicated in multistage than in straight direct implementations, and all the more complicated so that the number of stages is large. From this point of view, there can be a penalty in the half-band filter approach; to what extent? It is a difficult point because it depends on the kind of application, software or hardware, the specifications of the processing, and the environment.

After all, it is the user's task to determine the most efficient technique, taking all the facets of his particular problem into consideration; the results given in the present paper provide elements for the decision.

References

[1] M. G. Bellanger, J. L. Daguet, and G. P. Lepagnol, "Interpolation, extrapolation and reduction of computation speed in digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, Aug. 1974.

[2] R. E. Crochiere and L. R. Rabiner, "Optimum FIR digital filter implementation for decimation interpolation and narrow-band filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, Oct. 1975.

[3] L. R. Rabiner and R. E. Crochiere, "A novel implementation for narrow-band FIR digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, Oct. 1975.

[4] R. E. Crochiere and L. R. Rabiner, "Further considerations in the design of decimators and interpolators," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, Aug. 1976.

[5] D. J. Goodman and M. J. Carey, "Nine digital filters for decimation and interpolation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 121-126, Apr. 1977.

Constrained Least Squares Filtering

K. A. DINES AND A. C. KAK

Abstract—**In the following we: 1) present a frequency domain derivation of the constrained least squares filter, which is much simpler than previous approaches relying on the diagonalization of block circulant matrices; 2) derive a lower bound on the Lagrange multiplier appearing in the filter and obtain conditions under which the Lagrange multiplier may assume negative values; and 3) indicate how the frequency domain expression for the error energy can be used to reduce computation time in the iterative determination of the Lagrange multiplier.**

Manuscript received May 12, 1976; revised October 20, 1976 and February 10, 1977. This work was supported by NIH Grant N01-NS-3-2319.

The authors are with the School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

I. Introduction

One of the more recently proposed schemes for signal deconvolution is constrained least squares filtering [1]-[5], [7]. One of the attractive features of this technique is the small amount of *a priori* information that is required: only the total noise energy need be known. In comparison, techniques such as Wiener filtering require the assumption of stationarity and detailed knowledge of the correlation properties of the signal and noise. For a more detailed discussion of the comparative merits of the constrained least squares filtering, see [9].

For one-dimensional signals, the method of constrained least squares filtering was first formulated by Phillips [3] and later refined by Twomey [4], [5]. For long data streams, the implementation proposed by Phillips and Twomey is made difficult by the need to invert large matrices. This problem of implementation was successfully solved by Hunt [1], who recognized the special structure of degradation operators and made use of properties of circulant matrices.

In a recent paper, Hunt [3] derived the constrained least squares filter for the case of images. The basis of the derivation was that the two-dimensional discrete convolution can be expressed in vector-matrix form, in which degradation operators (point spread function of the degradation) are expressed as block-circulant matrices. Hunt then made use of the property that a block-circulant matrix is diagonalized by the two-dimensional discrete Fourier transform (DFT).

In this communication we present an alternate derivation of this filter. The derivation is much simpler and carries out the optimization directly in the frequency domain. This derivation also leads to a lower bound on the Lagrange multiplier in the filter. We will show that under certain conditions, the Lagrange multiplier may take negative values. Until now the Lagrange multiplier has been arbitrarily assumed to take only nonnegative values. Even though the derived conditions are of a highly restrictive nature (and, therefore, one may not run into them in everyday signal processing), we still feel one should at least be aware of them. Finally, we also show that the frequency domain expression for the error constraint can be used to reduce computation time for the iterative determination of the Lagrange multiplier appearing in the filter.

II. Frequency Domain Derivation of the Constrained Filter

In what follows we will present the derivation for the two-dimensional case, the reason being that the simplification, due to the optimization directly in the frequency domain, is most significant for this case.

The principal signals (real-valued) involved in the constrained filter derivation are related as shown in Fig. 1. An ideal digital picture $s(m,n)$ serves as input to a linear space-invariant system with point spread function $p(m,n)$. The degraded picture $d(m,n)$ is corrupted by additive noise $e(m,n)$, resulting in the received picture $r(m,n)$. For the case of spatially invariant degradations, the received two-dimensional array $r(m,n)$ is related to $s(m,n)$, $p(m,n)$ and $e(m,n)$ by

$$r(m,n)=\sum_{k=0}^{M-1}\sum_{l=0}^{N-1} s(k,l)\,p(m-k,n-l)+e(m,n)$$

$$m=0,1,2,\cdots,M-1$$

$$n=0,1,2,\cdots,N-1$$

$m-k$ take integer values modulo M

$n-l$ take integer values modulo N (1).

where the convolution in (1) is recognized to be circular. This assumption is equivalent to making the matrices circulant [1]. In the usual manner, (1) can be used to compute the normal linear convolution if the arrays s and p have been padded with sufficient zeros to avoid foldover in the circular convolution.

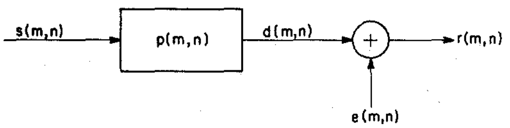

Fig. 1. Relationships of the principal signals involved in constrained estimation.

We shall leave M and N as variables for the time being, assuming that they will be selected to avoid the aliasing problem. After the filter has been derived, proper choices for M and N will be discussed.

Our basic aim is to estimate $s(m,n)$ from a knowledge of $r(m,n)$ and $p(m,n)$ in the presence of measurement errors $e(m,n)$. In the constrained solution to (1) it is assumed that information concerning the noise is limited to a knowledge of energy content defined as

$$\nu^2=\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} e^2(m,n). \tag{2}$$

We, therefore, seek an estimate of the picture $\hat{s}(m,n)$ that satisfies (1) subject to the constraint of (2). Since the summation constraint in (2) can be satisfied by many functions, we select the one that minimizes some property of the picture. This property is expressed in terms of the energy content in a filtered version of the estimated picture. Specifically,

$$\text{minimize} \sum_{m=0}^{M-1}\sum_{n=0}^{N-1}[c(m,n)*\hat{s}(m,n)]^2 \tag{3}$$

where $*$ denotes circular convolution, and $c(m,n)$ is a constraint array which we are free to select. For example, if we wish to obtain a "smooth" estimate, then $c(m,n)$ can be chosen as the Laplacian operator. In this case, one would be selecting the solution to (1) which satisfies (2) having minimum energy in its Laplacian.

The desired estimate $\hat{s}(m,n)$ can be obtained by formulating the problem in the frequency domain. By using a usual definition of the DFT [7] and Parseval's theorem that one can obtain from it, (1)-(3) can be expressed in the frequency domain as

$$R(k,l)=MN\,\hat{S}(k,l)P(k,l)+E(k,l) \tag{4}$$

$$\nu^2=MN\sum_{k=0}^{M-1}\sum_{l=0}^{N-1}|E(k,l)|^2 \tag{5}$$

$$\text{minimize } M^3N^3\sum_{k=0}^{M-1}\sum_{l=0}^{N-1}|C(k,l)\hat{S}(k,l)|^2 \tag{6}$$

$$\text{subject to } MN\sum_{k=0}^{M-1}\sum_{l=0}^{N-1}|R(k,l)-MN\,\hat{S}(k,l)P(k,l)|^2=\nu^2 \tag{7}$$

where we have used capital letters to denote the DFT's of corresponding sequences in lower case letters. Equation (7) is simply a restatement of (5), obtained by using (4).

The minimization can be carried out using Lagrange multipliers. We find $\hat{S}(k,l)$ that minimizes the functional U given by

$$U=\sum_{k=0}^{M-1}\sum_{l=0}^{N-1}\{M^3N^3|C(k,l)\hat{S}(k,l)|^2 + \lambda MN|R(k,l)-MN\hat{S}(k,l)P(k,l)|^2\} \tag{8}$$

where λ is a Lagrange multiplier. Since, in general, $\hat{S}(k,l)$ is complex, we write

$$\hat{S}(k,l)=A(k,l)+jB(k,l). \tag{9}$$

Substituting (9) into (8) and differentiating first with respect to each $A(k, l)$ and then with respect to each $B(k, l)$, we get

$$\frac{\partial U}{\partial A(k,l)} = M^3N^3\, 2\,|C(k,l)|^2 A(k,l) + 2\lambda M^3N^3\, |P(k,l)|^2 \cdot A(k,l) - \lambda M^2N^2 [R(k,l)P^*(k,l) + R^*(k,l)P(k,l)] \tag{10}$$

$$\frac{\partial U}{\partial B(k,l)} = M^3N^3\, 2\,|C(k,l)|^2 B(k,l) + 2\lambda M^3N^3\, |P(k,l)|^2 \cdot B(k,l) + j\lambda M^2N^2 [R(k,l)P^*(k,l) - R^*(k,l)P(k,l)]$$

$$k = 0, 1, 2, \cdots, M-1$$
$$l = 0, 1, 2, \cdots, N-1. \tag{11}$$

At the minimum, each of the $2MN$ derivatives in (10) and (11) must equal zero. From this we get

$$A(k,l) = \frac{1}{MN}\,\frac{\lambda\, \mathrm{Re}\,\{R(k,l)P^*(k,l)\}}{|C(k,l)|^2 + \lambda\,|P(k,l)|^2} \tag{12}$$

$$B(k,l) = \frac{1}{MN}\,\frac{\lambda\, \mathrm{Im}\,\{R(k,l)P^*(k,l)\}}{|C(k,l)|^2 + \lambda\,|P(k,l)|^2} \tag{13}$$

where Re $\{\cdot\}$ denotes the real part, Im $\{\cdot\}$ denotes the imaginary part; and * denotes the complex conjugate. From (9), (12), and (13) we obtain

$$\hat{S}(k,l) = \frac{1}{MN}\,\frac{P^*(k,l)R(k,l)}{\lambda\,|C(k,l)|^2 + |P(k,l)|^2} \tag{14}$$

where $\gamma = 1/\lambda$. Therefore, the restoration filter denoted by $H(k, l)$ is given by

$$H(k,l) = \frac{1}{MN}\,\frac{P^*(k,l)}{\lambda\,|C(k,l)|^2 + |P(k,l)|^2},$$
$$k = 0, 1, 2, \cdots, M-1, \quad l = 0, 1, 2, \cdots, N-1. \tag{15}$$

The form of the filter in (15) is the same as that derived by Hunt by using diagonalization properties of block circulant matrices.

The Lagrange multiplier λ (or $\gamma = 1/\lambda$) must be chosen to satisfy the constraint of (7). Substituting the expression for the estimate (14) into (7), we obtain an expression in the frequency domain that γ must satisfy

$$MN \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} \frac{\gamma^2\,|R(k,l)|^2\,|C(k,l)|^4}{(\gamma\,|C(k,l)|^2 + |P(k,l)|^2)^2} = \text{a known constant } \nu^2. \tag{16}$$

In order to determine γ for the restoration filter, one must solve (16) by an iterative procedure. The manner in which this can be done will be discussed later.

III. Lower Bound on the Lagrange Multiplier

We have yet to show the condition under which the filter of (15) results in a minimum rather than a maximum for the functional U. From the calculus of several variables [6] we have the following *sufficient* condition to guarantee a minimum:

$$\sum_{k=0}^{M-1} \sum_{l=0}^{N-1} \frac{\partial^2 U}{\partial A^2(k,l)} \mathrm{COS}^2\, \alpha(k,l) + \frac{\partial^2 U}{\partial B^2(k,l)} \mathrm{COS}^2\, \beta(k,l) + \text{all cross partials derivatives of order } 2 > 0 \tag{17}$$

for all $\alpha(k, l), \beta(k, l)$ such that

$$0 \leqslant \alpha(k,l), \beta(k,l) \leqslant 2\pi.$$

In (17), COS $\alpha(k, l)$ and COS $\beta(k, l)$ are the direction cosines of a vector that is located in the tangential hyperplane at the minimum in the $2NM$ dimensional space spanned by the $A(k, l)$'s and $B(k, l)$'s. From (10) and (11), we get

$$\frac{\partial^2 U}{\partial A^2(k,l)} = 2M^3N^3 [\,|C(k,l)|^2 + \lambda\,|P(k,l)|^2\,] \tag{18}$$

and

$$\frac{\partial^2 U}{\partial B^2(k,l)} = 2M^3N^3 [\,|C(k,l)|^2 + \lambda\,|P(k,l)|^2\,]. \tag{19}$$

Since in our case all cross partials of order 2 are zero, the condition in (17) can be expressed as

$$\sum_{k=0}^{M-1} \sum_{l=0}^{N-1} [\,|C(k,l)|^2 + \lambda\,|P(k,l)|^2] \, [\mathrm{COS}^2\, \alpha(k,l) + \mathrm{COS}^2\, \beta(k,l)] > 0, \quad 0 \leqslant \alpha(k,l),\ \beta(k,l) \leqslant 2\pi. \tag{20}$$

Since the strict inequality in (20) must hold with each direction cosine taking any value in the range (-1, +1), we must have

$$|C(k,l)|^2 + \lambda\,|P(k,l)|^2 > 0,$$
$$\text{for all} \quad k = 0, 1, 2, \cdots, M-1$$
$$\text{and} \quad l = 0, 1, 2, \cdots, N-1. \tag{21}$$

Therefore, the Lagrange multiplier must satisfy

$$\lambda > -\frac{|C(k,l)|^2}{|P(k,l)|^2}, \quad \text{for all} \quad k = 0, 1, 2, \cdots, M-1$$
$$\text{and} \quad l = 0, 1, 2, \cdots, N-1. \tag{22}$$

Equivalently,

$$\lambda > -\min_{k,l} \frac{|C(k,l)|^2}{|P(k,l)|^2} = \lambda_{\min}. \tag{23}$$

The implication of (23) is that negative values for the Lagrange multiplier are possible in the optimum filter realization. The lower bound in (23) depends on the ratio of the power spectra of the constraint sequence and the degradation. Previous workers [1]-[5] have assumed that the Lagrange multiplier was nonnegative. In the case of smoothness constraints, such as those used by Phillips [3], Hunt [1], [2], and Twomey [4], [5], $C(0, 0) = 0$. So, if $P(0, 0) \neq 0$, then (23) predicts that $\lambda > 0$. Thus, γ, the reciprocal of λ, must also be nonnegative.

In order to determine the circumstances which may lead to a negative value for γ, the error relationship in (16) will be explored in greater detail. It will be more convenient to work with the original Lagrange multiplier $\lambda = 1/\gamma$. When viewed as a function of λ, (16) becomes

$$\rho(\lambda) = MN \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} \frac{|R(k,l)|^2\,|C(k,l)|^4}{(|C(k,l)|^2 + \lambda\,|P(k,l)|^2)^2}. \tag{24}$$

Taking the derivative of (24), one has

$$\frac{d\rho(\lambda)}{d\lambda} = -MN \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} \frac{|R(k,l)|^2\,|C(k,l)|^4\,|P(k,l)|^2}{[\,|C(k,l)|^2 + \lambda\,|P(k,l)|^2]^3}. \tag{25}$$

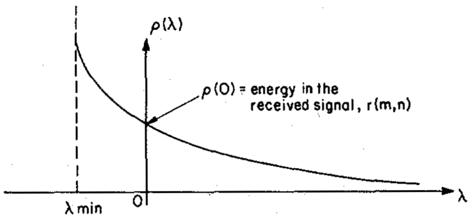

Fig. 2. Typical plot of the noise energy ρ as a function of the Lagrange multiplier λ with $\lambda_{\min}$ defined in (27).

Using the lower bound on λ (23), we can see from (25) that

$$\frac{d\rho(\lambda)}{d\lambda} < 0, \quad \text{for} \quad \lambda > \lambda_{\min}. \tag{26}$$

Therefore, the error energy is a monotonically decreasing function of the Lagrange multiplier λ. In order to gain further insight into the behavior of the errors as a function of λ, we can evaluate $\rho(\lambda)$ at some particular values of λ. As λ approaches zero, (24) yields

$$\rho(0) = MN \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} |R(k,l)|^2. \tag{27}$$

Equation (27) says that as $\lambda \to 0$, the error energy approaches the energy in the received signal $r(m,n)$. If λ is allowed to approach infinity ($\gamma \to 0$), then

$$\lim_{\lambda \to \infty} \rho(\lambda) = 0. \tag{28}$$

From (26)-(28), we can sketch a typical $\rho(\lambda)$ versus λ curve as in Fig. 2. It is noted that the monotonicity property (26) and the condition of (27) imply that λ will assume negative values if and only if: 1) negative values are possible from (23), and 2) the error energy is greater than the energy in the received signal, i.e.,

$$\nu^2 > MN \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} |R(k,l)|^2. \tag{29}$$

The first condition occurs if all of the frequencies are constrained so that

$$|C(k,l)| \neq 0, \quad \text{for any} \quad k, l. \tag{30}$$

The second condition occurs when the signal-to-noise ratio in the received signal $r(m,n)$ is less than one.

IV. Practical Implementation

First we make some comments about the proper values of M and N. As indicated previously, proper values for M and N must be chosen to avoid aliasing in the circular convolutions implied by the frequency domain estimate (14). Specifically, the required convolutions are indicated by multiplication in the frequency domain as

$$MN\hat{S}(k,l)\,[\gamma\,|C(k,l)|^2 + |P(k,l)|^2] = P^*(k,l)R(k,l). \tag{31}$$

Therefore, we must incorporate sufficient zero padding such that the circular convolutions given by

$$\hat{s}(m,n) * \gamma c(m,n) * c(-m,n) + \hat{s}(m,n) * p(m,n) * p(-m,-n) = p(-m,-n) * r(m,n) \tag{32}$$

are equivalent to normal linear convolution. Thus, suppose the original two-dimensional sequences were of the following sizes:

$$\hat{s}(m,n){:}\ M_s \times N_s \quad \text{(samples)}$$
$$c(m,n){:}\ M_c \times N_c$$
$$p(m,n){:}\ M_p \times N_p$$
$$r(m,n){:}\ M_r \times N_r. \tag{33}$$

The left-hand side of (32) implies that

$$M \geqslant \text{maximum}\ [M_s + 2M_c - 2; M_s + 2M_p - 2] \tag{34}$$

and

$$N \geqslant \text{maximum}\ [N_s + 2N_c - 2; N_s + 2N_p - 2] \tag{35}$$

to avoid aliasing. However, normally we do not know the size of the estimated sequence $\hat{s}(m,n)$. A logical assumption for practical implementation is that

$$M_s \leqslant M_r$$
$$N_s \leqslant N_r \tag{36}$$

because $r(m,n)$ is generally a smeared version of $s(m,n)$ and hence occupies a larger area in the (m,n) plane. Therefore, one may replace M_s and N_s in (34) and (35) by M_r and N_r, respectively.

A practical implementation of constrained least squares estimation has been described elsewhere [1], [2]. Our implementation differs from Hunt's [2] in that we evaluate the residual in the frequency domain by (24) during the iterative determination of γ instead of obtaining trial solutions in the space domain. This approach eliminates the need for inverse DFT's during the iterations, requiring only the evaluation of the sum in (24) for each choice of γ. Specifically, one can compute the DFT's of r, c, and p once; use (16) to find γ; and estimate $s(m,n)$ by inverse transformation of (14). Finally, since all the space-domain sequences are real, the symmetry in their DFT's can be exploited so that the summation in (16) need include only half the terms.

V. Summary

In this communication we first presented a derivation for the constrained deconvolution filter that carries out the optimization directly in the frequency domain. The simplification due to this approach is noteworthy for the two-dimensional case. That is why the two-dimensional case was presented here. The derivation for the one-dimensional case is completely analogous.

Using the frequency domain optimization, we also derived the lower bound for the Lagrange multiplier in the filter and derived the conditions under which it could take negative values. Until now, the Lagrange multiplier has been arbitrarily assumed to take only nonnegative values. We also showed how, by iterating directly in the frequency domain, one can calculate the Lagrange multiplier much faster than previous approaches. During the review of this paper, one of the referees brought to our attention a currently pending publication [8] that also brings out this faster way to calculate the Lagrange multiplier. Finally, for the sake of completion, the authors would like to draw the attention of the reader to other types of constrained minimization techniques in image restoration [10], [11]. In these techniques, the restoration algorithm is obtained by solving a boundary value problem.

References

[1] B. R. Hunt, "Deconvolution of linear systems by constrained regression and its relationship to Wiener theory," *IEEE Trans. Automat. Contr.*, vol. AC-17, pp. 703-705, Oct. 1972.

[2] ——, "The application of constrained least squares estimation to image restoration by digital computer," *IEEE Trans. Comput.*, vol. C-22, pp. 805-812, Sept. 1973.

[3] D. L. Phillips, "A technique for the numerical solution of certain integral equations of the first kind," *J. Ass. Comput. Mach.*, vol. 9, pp. 84-97, 1962.

[4] S. Twomey, "On the numerical solution of the Fredholm equations of the first kind by the inversion of the linear system produced by quadrature," *J. Ass. Comput. Mach.*, vol. 10, pp. 97-101, 1963.

[5] —, "The application of numerical filtering to the solution of integral equations encountered in indirect sensing measurements," *J. Franklin Inst.*, vol. 279, pp. 95-109, Feb. 1965.

[6] W. Kaplan, *Advanced Calculus*. New York: Addison-Wesley, 1959.

[7] A. Rosenfeld and A. C. Kak, *Digital Picture Processing*. New York: Academic, 1976.

[8] S. S. Reddi, "An improved computational scheme for constrained least squares image restoration," submitted for publication to *IEEE Trans. Comput*.

[9] B. R. Hunt, "Digital image processing," *Proc. IEEE*, vol. 63, pp. 693-708, Apr. 1975.

[10] A. K. Jain and E. Angel, "Image restoration, modelling, and reduction of dimensionality," *IEEE Trans. Comput.*, vol. C-23, pp. 470-476, May 1974.

[11] A. K. Jain and D. Sworder, "Constrained linear filtering as a multistage process," *J. Math. Anal. Appl.*, vol. 44, pp. 48-56, Oct. 1973.

A Symmetry Relationship for "Between" Scaling in Cascade Digital Filters

B. P. GAFFNEY AND J. N. GOWDY

Abstract—**This correspondence examines a symmetry relationship for fixed-point cascade digital filters having no error sources in the numerator and scaled with L_2 norm factors which are placed between the filter's subsections. It is shown that the noise gain of any ordering is the same as its inverse. It is also shown that the symmetry property holds for filters implemented in $1D$ and $2D$ form.**

In order to prevent the nonlinear effects of adder overflow in cascade fixed-point digital filters, the signal must be scaled either internally or externally to the filter. Jackson [1] has developed a method of scaling which attenuates or amplifies the signal at internal points (nodes) of the filter. Internal scaling allows the signal to fully use the available register lengths at the internal nodes of the filter. Jackson examined implementations which altered the numerator coefficients of the canonical subsections of the filter.

In certain types of filters, a scaling implementation that does not modify the numerator coefficients reduces the number of multipliers by approximately a third. In the analysis below, the scaling factors are considered as separate numerator elements between the filter's subsections. The scaling multipliers do not alter any of the filter's coefficients.

Consider the digital filter represented by Fig. 1. Let

$$H(z) = a_0 \prod_{K=1}^{n} H_K(z) \tag{1}$$

be the transfer function of the digital filter where $H_K(z)$ is a subsection and a_0 is a constant. The transfer function from the input node 0 to node i is

$$F_i(z) = \prod_{j=1}^{i} H_j(z). \tag{2}$$

Manuscript received September 13, 1976; revised January 19, 1977.

B. P. Gaffney is with RCA/Government and Commercial Systems, Moorestown, NJ 08057.

J. N. Gowdy is with the Department of Electrical and Computer Engineering, Clemson University, Clemson, SC 26932.

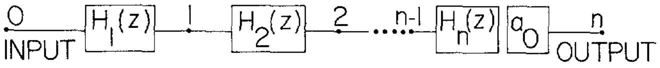

Fig. 1. Subsections for a cascade digital filter.

The filter's subsection $H_K(z)$ can be either a numerator or a denominator section, that is, a subsection containing either all zeros or all poles, but not zeros and poles together.

The scaled configuration of (1) is represented in Fig. 2 where for L_2 norm scaling, the scaling factors are given by

$$\begin{aligned} S_0 &= 1/\|F_1\|_2 \\ S_j &= \|F_j\|_2/\|F_{j+1}\|_2, \qquad 1 \leqslant j \leqslant n-1 \\ S_n &= \|F_n\| a_0 \end{aligned} \tag{3}$$

where $\|F_i\|_2$ is the L_2 norm.

The noise gain of the filter represented by Fig. 2 can be shown to be

$$\begin{aligned} V = &\sum_{i=1}^{n} (1+d_i)a_0^2 \left\| \prod_{j=i}^{n} H_j \right\|_2^2 \left\| \prod_{j=1}^{i} H_j \right\|_2^2 \\ &+ \sum_{i=1}^{n} n_i\, a_0^2 \left\| \prod_{j=1}^{i} H_j \right\|_2^2 \left\| \prod_{j=i+1}^{n} H_j \right\|_2^2 + 1 \end{aligned} \tag{4}$$

where

$$d_i = \begin{cases} >0, & \text{for denominator subsection} \\ =0, & \text{for numerator subsection} \end{cases} \tag{5}$$

and

$$n_i = \begin{cases} \geqslant 0, & \text{for numerator subsection} \\ =0, & \text{for denominator subsection.} \end{cases} \tag{6}$$

That is, the n_i and d_i are the number of noninteger coefficients of the subsections.

Theorem

The "between" scaling has a symmetry property which requires the calculation of only one-half the possible permutations of a fixed-point cascade digital filter for the L_2 norm scaling and no error sources in the numerator ($n_i = 0$ for all i). The symmetry property is that the gain of the roundoff noise of any permutation is the same as its inverse permutation.

Proof:

For $n = 2$, (3) and (4)

$$\begin{aligned} S_0 &= 1/\|H_1\|_2 \\ S_1 &= \|H_1\|_2/\|H\|_2 \\ S_2 &= \|H\|_2\, a_0 \end{aligned} \tag{7}$$

$$\begin{aligned} V = &(1+d_1)\|H\|_2^2 \|H_1\|_2^2 a_0^2 + (1+d_2)\|H_2\|_2^2 \|H\|_2^2 a_0^2 \\ &+ n_1 \|H_1\|_2^2 \|H_2\|_2^2 a_0^2 + n_2 \|H\|_2^2 a_0^2 + 1. \end{aligned} \tag{8}$$

The interchanging of 1 and 2 in (8) corresponds to the noise gain for the inverse permutation, i.e.,

$$\begin{aligned} V = &(1+d_2)\|H\|_2^2 \|H_2\|_2^2 a_0^2 + (1+d_1)\|H_1\|_2^2 \|H\|_2^2 a_0^2 \\ &+ n_2 \|H_2\|_2^2 \|H_1\|_2^2 a_0^2 + n_1 \|H\|_2^2 a_0^2 + 1. \end{aligned} \tag{9}$$

For $n_1 = n_2$, the orderings have the same noise gain. Consider the general case. The scale factors and noise gain of the filter represented by Fig. 2 are given by (3) and (4). The scale factors and noise gain of the inverse permutation are

$$S_0 = 1/\|H_1\|_2$$