Optimal FIR Pre- and Postfilters for Decimation and Interpolation of Random Signals

HENRIQUE S. MALVAR, MEMBER, IEEE, AND DAVID H. STAELIN, FELLOW, IEEE

Abstract—**A new technique for the design of finite impulse response (FIR) filters for decimation and interpolation in multirate systems is presented. With this technique, FIR pre- and postfilters that jointly minimize a frequency-weighted mean-square (MS) error between the original and reconstructed signals can be designed. Unlike most other FIR filter design methods, there is no need for ideal filter prototypes: the optimal pre-postfilter pair is determined from the signal and noise spectra and the up- and down-sampling factors. Some examples of image and speech processing show that the MS-optimal filter pair leads to typical SNR improvements of 2-6 dB, in comparison to other commonly used filters.**

I. Introduction

In multirate digital signal processing [1], [2] one is frequently faced with the design of pre- and postfilters for decimation (down-sampling) and interpolation (up-sampling). A typical system model is that of Fig. 1(a) where the signal $x(n)$ must be transmitted through a noisy channel whose sampling rate is K times lower than that of $x(n)$. Although the channel noise is physically added to the signal after it is decimated, we can always work with an equivalent noise $d(n)$ that precedes decimation, without loss of generality. That model could be applied, for example, to an image coding/decoding system in which the pre- and postfilters represent the interface between a low-resolution coder/decoder (codec) and high-resolution image acquisition and display subsystems; in such a case, the noise $d(n)$ would be due to the codec. Our model in Fig. 1(a) also includes an input noise source $u(n)$ since in some applications, the input signal may only be available through a noisy measurement, e.g., in some problems of telemetry and biomedical signal processing.

Although infinite impulse response (IIR) filters could be used in Fig. 1(a), in most applications FIR filters are preferred because of their inherent stability and also because they can easily be constrained to have a linear phase response. The design of suitable decimation and interpolation FIR filters for the system in Fig. 1 is commonly approached in two steps: first, ideal pre- and postfilter responses (usually low-pass) are determined, and then FIR responses that approximate the ideal ones are computed. Common techniques are windowing and equiripple-ripple Chebyshev approximation [2].

In this paper, we suggest a different approach for the design of FIR pre- and postfilters: under the assumption that a weighted mean-square (MS) error between the original and reconstructed signals is a reasonable performance measure, we directly optimize the FIR pre- and postfilter responses. An immediate advantage of this approach is the elimination of the ideal filter prototypes so that one has not to be concerned about transition bandwidths, ripple factors, etc. Another advantage is that, unlike other optimal FIR filter design methods such as Chebyshev approximation, our technique can easily be extended to multidimensional systems. Such extensions are reported elsewhere [3], [4]. The SNR gains of optimal pre- and postfilters over other commonly used FIR filters depend on the system under consideration; we report here some image and speech processing examples, for which the improvement is on the order of 2-6 dB.

In the next section, we examine the independent optimization of the pre- and postfilters. The joint optimization procedure is described in Section III. In Section IV, we compare the optimal pre- and postfilters to others designed under different criteria for image and speech processing.

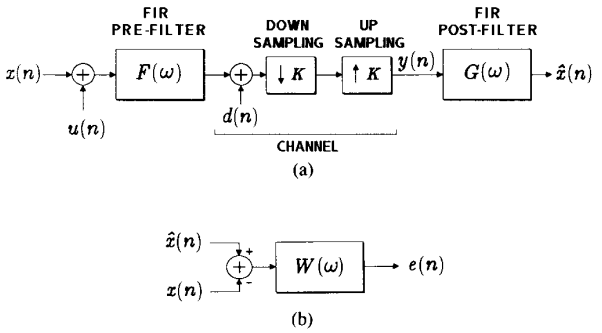

Fig. 1. Discrete-time communication system with down- and up-sampling. (a) System model where $u(n)$ and $d(n)$ are the input and channel noises, respectively; the down- and up-sampling are equivalent to periodic sampling by $\delta_K(n)$. (b) Error signal definition; the observer response $W(\omega)$ is a frequency weighting on the absolute reconstruction error.

II. Optimization of the Pre- or Postfilter

In this section, we are interested in the optimization of a single filter in Fig. 1(a), either the pre- or postfilter. We assume that the input signal $x(n)$ and the noise sources $u(n)$ and $d(n)$ are stationary random signals with known spectra. We further assume that the noise sequences are uncorrelated with the signals (in [3] we consider correlated channel noises and quantization noise, in particular). We note that the cascaded operations of down- and up-sampling by a factor of K are equivalent to multiplication by the standard periodic sampling sequence $\delta_K(n)$, defined as [1]

$$\delta_K(n) \equiv K \sum_{r=-\infty}^{\infty} \delta(n-rK) = \sum_{r=0}^{K-1} \exp\left(j\frac{2\pi rn}{K}\right). \quad (1)$$

Paper approved by the Editor for Digital Communications of the IEEE Communications Society. Manuscript received February 25, 1986; revised April 7, 1987. This work was supported in part by the Center for Advanced Television Studies. The work of H. S. Malvar was supported in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (Brazil) under Grant 200.832-82.

H. S. Malvar was with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139. He is now with the Departamento de Engenharia Elétrica, Universidade de Brasília, 70910 Brasília, Brazil.

D. H. Staelin is with the Department of Electrical Engineering and Computer Science and the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139.

IEEE Log Number 8717484.

We have included the scaling factor K in (1) to simplify the frequency-domain equations that follow.

The error signal $e(n)$ is defined in Fig. 1(b) as the result of passing $\hat{x}(n) - x(n)$ through an *observer* filter $w(n)$. The mean-square amplitude of $e(n)$ is the error performance measure that we seek to minimize. Although other error measures could be more appropriate for particular applications, the weighted mean-square error has been successfully adopted in many problems of communication theory [5].

Before we proceed with our analysis, it is important to note that the signals $y(n)$, $\hat{x}(n)$, and $e(n)$ in Fig. 1 are not wide-sense stationary because of the time-varying multiplication by $\delta_K(n)$, and so their spectra are not strictly defined. Nevertheless, those sequences are cyclostationary [6] processes, with periodic autocorrelation functions. For that class of processes, we can still define a meaningful spectrum as the Fourier transform of the autocorrelation function averaged over exactly one period [6]. Hence, the variance (or energy) of the error signal $e(n)$ can be obtained as

$$\xi = \frac{1}{K}\sum_{n=0}^{K-1} E[e^2(n)] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_{ee}(\omega)\, d\omega. \tag{2}$$

We can relate the error spectrum $\Phi_{ee}(\omega)$ to the signal and noise spectra and to the pre- and postfilter frequency responses by

$$\Phi_{ee}(\omega) = |W(\omega)|^2[\Phi_{xx}(\omega) + |G(\omega)|^2\Phi_{yy}(\omega)] - 2|W(\omega)|^2 \operatorname{Re}\{G(\omega)F(\omega)\}[\Phi_{xx}(\omega) + \Phi_{uu}(\omega)] \tag{3}$$

where the signal $y(n)$ is the channel output, that is, the input to the postfilter. The spectrum of $y(n)$ is given by [3]

$$\Phi_{yy}(\omega) = \tilde{\Phi}_{dd}(\omega) + \sum_{r=0}^{K-1} F^2(\omega + r\omega_K)[\Phi_{xx}(\omega + r\omega_K) + \Phi_{uu}(\omega + r\omega_K)] \tag{4}$$

where $\omega_K \equiv 2\pi/K$ is the Nyquist frequency and $\tilde{\Phi}_{dd}(\omega)$ is the spectrum of the sampled channel noise, i.e.,

$$\tilde{\Phi}_{dd}(\omega) \equiv \sum_{r=0}^{K-1} \Phi_{dd}(\omega + r\omega_K). \tag{5}$$

For any given magnitude responses for the filters $F(\omega)$ and $G(\omega)$, the term $\operatorname{Re}\{G(\omega)F(\omega)\}$ in (3) is maximized when the phases of $F(\omega)$ and $G(\omega)$ are both equal to zero for all ω (because we did not allow for a delay in the error signal definition). Therefore, we shall concentrate on zero phase filters. Specifically, we impose the following constraints on their impulse responses:

$$\begin{aligned} f(-n) &= f(n) \\ g(-n) &= g(n) \\ f(n) &= 0, \quad \text{if } |n| > L \\ g(n) &= 0, \quad \text{if } |n| > M. \end{aligned} \tag{6}$$

Under the above assumptions, we can rewrite the error spectrum as

$$\Phi_{ee}(\omega) = |W(\omega)|^2[\Phi_{xx}(\omega) + G(\omega)^2\Phi_{yy}(\omega)] - 2|W(\omega)|^2 G(\omega)F(\omega)[\Phi_{xx}(\omega) + \Phi_{uu}(\omega)]. \tag{7}$$

Our objective in this paper is the minimization of (7) under the constraints in (6). We cannot work directly with $F(\omega)$ and $G(\omega)$ since we do not have enough degrees of freedom to arbitrarily set their values for all frequencies. One approach towards incorporating the FIR constraints into (3) is to convert to the time domain all terms in which $F(\omega)$ and $G(\omega)$ appear; this leads to

$$\begin{aligned} \xi = {} & \frac{1}{2\pi}\int_{-\pi}^{\pi} |W(\omega)|^2\Phi_{xx}(\omega)\, d\omega \\ & - 2\sum_{l=-L}^{L}\sum_{m=-M}^{M} f(l)g(m)a(l-m) \\ & + \sum_{l=-M}^{M}\sum_{m=-M}^{M} g(l)g(m)b(l-m) \\ & + \sum_{l=-L}^{L}\sum_{m=-L}^{L} f(l)f(m)\sum_{r=-M}^{M}\sum_{s=-M}^{M} g(r)g(s) \\ & \times \sum_{u=-\infty}^{\infty}\sum_{v=-\infty}^{\infty} w(u)w(v)\delta_K(r-s+u-v) \\ & \cdot c(l-m+r-s+u-v) \end{aligned} \tag{8}$$

where $w(n)$ is the observer impulse response and the sequences $a(n)$, $b(n)$, and $c(n)$ are defined by their Fourier transforms

$$\begin{aligned} A(\omega) &\equiv |W(\omega)|^2[\Phi_{xx}(\omega) + \Phi_{uu}(\omega)] \\ B(\omega) &\equiv |W(\omega)|^2\tilde{\Phi}_{dd}(\omega) \\ C(\omega) &\equiv \Phi_{xx}(\omega) + \Phi_{uu}(\omega). \end{aligned} \tag{9}$$

The optimization problem could be formulated as the minimization of (8) as a function of the vector of unknowns $[f(0)f(1)\ \cdots\ f(L)g(0)g(1)\ \cdots\ g(M)]$, but it would be virtually impossible to analyze such issues as convexity and convergence because (8) is a quartic form. However, if we fix the prefilter coefficients, then the error is a quadratic form on the postfilter coefficients, which is easier to minimize [3] (the error is also quadratic on the prefilter coefficients if we fix the postfilter). This suggests a simple approach towards the derivation of a jointly optimal filter pair: first obtain independent solutions for the pre- and postfilters, and then combine them in an iterative procedure that computes the jointly optimal pair. As discussed in [3], closed-form solutions for a jointly optimal filter pair cannot be obtained, except for trivial cases, e.g., $L = M = 1$, which will not be specifically considered.

A. The Optimal Postfilter

The design of the postfilter (or interpolator) has received much more attention in the literature than the prefilter design. Oetken *et al.* [7] derived the optimal interpolator without a prefilter for band-limited input signals and noiseless samples. Polydoros and Protonotarios [8] assumed a statistical description of the input signal, as in our work, and derived the optimal interpolator without a prefilter. As in [7], they have considered a noiseless system, but with the added restriction of zero intersymbol interference. Keys [9] used cubic convolution kernels, derived from cubic splines, to determine the impulse response of the interpolator; his main concern was the alleviation of sampling artifacts in image processing.

Interpolation of a stochastic signal from noisy samples with an FIR filter has been considered by Kay [10] and more recently by Radbel and Marks [11]. The solution in [11]

applies to the system in Fig. 1 for the case $F(\omega) \equiv 1$ and $u(n) \equiv 0$. Our results for the optimal interpolator here are essentially a generalization of [11] for any prefilter and input noise spectrum.

Our problem in this subsection is to solve (8) for the optimal $g(\cdot)$ for a fixed prefilter $f(\cdot)$. In this case, we can rewrite (8) explicitly as a function of the postfilter's coefficients in the form

$$\xi = \frac{1}{2\pi}\int_{-\pi}^{\pi} |W(\omega)|^2 \Phi_{xx}(\omega)\, d\omega + \sum_{l=-M}^{M}\sum_{m=-M}^{M} g(l)g(m)\psi(l-m) - 2\sum_{l=-M}^{M} g(l)\theta(l) \tag{10}$$

where $\psi(n)$ and $\theta(n)$ are the inverse Fourier transforms of $\Psi(\omega)$ and $\Theta(\omega)$, respectively, which are defined by

$$\Psi(\omega) \equiv |W(\omega)|^2 \Phi_{yy}(\omega) \tag{11}$$

and

$$\Theta(\omega) \equiv |W(\omega)|^2 F(\omega)[\Phi_{xx}(\omega) + \Phi_{uu}(\omega)]. \tag{12}$$

The first-order necessary condition for $g(n)$ to be an optimal postfilter is that $\partial\xi/\partial g(l) = 0$ for all l, which leads to the system of linear equations

$$\sum_{m=-M}^{M} g(m)\psi(l-m) = \theta(l) \qquad l = -M, -M+1, \cdots, M. \tag{13}$$

Since $\Psi(\omega)$ is a valid power spectrum, the matrix whose entries are $\psi(l-m)$ for $l, m = -M, \cdots, M$ is at least positive semidefinite [16]. With the mild assumption that $\Psi(\omega) > 0$ for all ω, the matrix is positive definite, and the error is then a strictly convex function of the postfilter coefficients. Thus, the unique solution to (13) globally minimizes the error for a fixed prefilter. We recognize (13) as a standard FIR Wiener filter equation [15].

The equations in (13) have a Toeplitz structure, and so they can be solved in $O[(2M+1)^2]$ operations by means of Levinson's recursion [12]. If M is very large, there are algorithms with $O(2M+1)[\log(2M+1)]^2$ complexity [13], [14], but these algorithms are considerably more difficult to implement than Levinson's recursion. It is interesting to note that the symmetry constraint imposed on the prefilter forces $\Theta(\omega)$ to be a real function, so that $\theta(n)$ is a symmetric sequence. Therefore, the solution to (13) necessarily leads to a symmetric sequence $g(n)$ that satisfies (6). We could exploit this symmetry to convert (13) to a Toeplitz-plus-Hankel system of only $M+1$ equations, which could also be efficiently solved, as discussed in [15].

With the optimal postfilter, (10) can be simplified to

$$\xi = \frac{1}{2\pi}\int_{-\pi}^{\pi} |W(\omega)|^2 \Phi_{xx}(\omega)\, d\omega - \sum_{l=-M}^{M} g(l)\theta(l). \tag{14}$$

It is not possible, however, to write the above equation in terms of the prefilter coefficients since Toeplitz forms are not, in general, analytically invertible [16].

B. The Optimal Prefilter

The design of optimal FIR prefilters has received little attention in the literature. Chevillat and Ungerboeck [21] derived optimal pre- and postfilters for a discrete-time input signal and a continuous-time band-limited channel. Their results apply directly to modem design, for example, but they cannot be used in our case since we have a discrete-time channel. Hummel [22] has considered the problem of designing an optimal prefilter when the interpolator is a spline function and the system is noiseless. He showed that the optimal prefilter in that case is also a spline function. Ratzel [23] has derived optimal Gaussian prefilters for digitized images, based on subjective experiments.

Recently, Faubert [24] has determined the optimal pre- and postfilters for a noiseless system for a performance criterion in which filtering and aliasing errors are independently weighted. If the system in Fig. 1 is noiseless and a flat frequency weight is considered, our results in the next section lead to filter pairs that are equivalent to those derived in [24]. Our work in this section can be viewed as a one-dimensional extension of Faubert's results for the noisy system in Fig. 1.

Under this assumption that the postfilter is fixed, the error expression in (8) can be simplified to

$$\xi = \frac{1}{2\pi}\int_{-\pi}^{\pi} |W(\omega)|^2 [\Phi_{xx}(\omega) + \tilde{\Phi}_{dd}(\omega)G^2(\omega)]\, d\omega + \sum_{l=-L}^{L}\sum_{m=-L}^{L} f(l)f(m)\gamma(l-m) - 2\sum_{l=-L}^{L} f(l)\vartheta(l) \tag{15}$$

where $\gamma(n)$ and $\vartheta(n)$ are the inverse Fourier transforms of $\Gamma(\omega)$ and $\Upsilon(\omega)$, respectively, which are defined by

$$\Gamma(\omega) \equiv [\Phi_{xx}(\omega) + \Phi_{uu}(\omega)] \sum_{r=0}^{K-1} |W(\omega - r\omega_K)|^2 G^2(\omega - r\omega_K) \tag{16}$$

and

$$\Upsilon(\omega) \equiv |W(\omega)|^2 G(\omega)\Phi_{xx}(\omega). \tag{17}$$

At this point, we introduce a power constraint on the prefilter output $v(n)$. The necessity of such a constraint is clear from (8); if we multiply all $f(n)$ by a constant α and divide all $g(n)$ by α with $|\alpha| > 1$, the error is reduced since the matrix formed by the elements $b(l-m)$ is at least positive semidefinite. Without loss of generality, we assume that the prefilter output power must be less than unity:

$$P \equiv E[v^2(n)] = \frac{1}{2\pi}\int_{-\pi}^{\pi} |F(\omega)|^2 [\Phi_{xx}(\omega) + \Phi_{uu}(\omega)]\, d\omega \le 1. \tag{18}$$

The above equation can also be written in the time domain as

$$P = \sum_{l=-L}^{L}\sum_{m=-L}^{L} f(l)f(m)[R_{xx}(l-m) + R_{uu}(l-m)] \le 1. \tag{19}$$

An optimal prefilter has to be a stationary point of the Lagrangian [17] corresponding to the objective function (15) and the constraint (19), i.e., there must exist a Lagrange multiplier λ such that

$$\frac{\partial\xi}{\partial f(l)} + \lambda\frac{\partial P}{\partial f(l)} = 0, \qquad \forall l. \tag{20}$$

The Lagrange multiplier also has the properties

$$\lambda \geq 0$$

$$\lambda(P-1)=0, \tag{21}$$

that is, if the power constraint is not satisfied by equality, then the value of the Lagrange multiplier is zero since the constraint is not binding. The Lagrange multiplier is nonnegative since the inequality is $P \leq 1$. A proof of (21) for the general nonlinear optimization problem can be found in [17].

From (20), we obtain

$$\sum_{m=-L}^{L} f(m)\{\gamma(l-m)+\lambda[R_{xx}(l-m)+R_{uu}(l-m)]\} = \vartheta(l)$$

$$l=-L,\ -L+1,\ \cdots,\ L. \tag{22}$$

We have again a symmetric Toeplitz system of linear equations to be solved. So, our discussion of fast algorithms for solving (22) also applies here. We note also that $\Gamma(\omega)$ is nonnegative for all ω, which means that $\gamma(n)$ is a valid autocorrelation function, and so the matrices formed by the elements $\gamma(l-m)$ and $\gamma(l-m)+\lambda[R_{xx}(l-m)+R_{uu}(l-m)]$ for $l, m = -L, \cdots, L$ are at least positive semi-definite. Thus, (15) is a convex function of the prefilter coefficients, and a solution to (22) is a global minimum.

There is still a problem in solving (22), which is the fact that the value of the Lagrange multiplier λ is not known *a priori*. There is a simple approach, however: first, we set $\lambda = 0$ and solve (22); if the solution satisfies $P < 1$, we are done; otherwise, the power constraint must be active, and we repeatedly solve (22) (λ must be updated by some technique for finding zeros of one-dimensional functions, e.g., Newton–Raphson's method [18]) until we obtain a solution for which $P \simeq 1$. Such a procedure is guaranteed to converge to an optimal prefilter [3].

III. Jointly Optimal Solution

In the previous section, we derived the optimal postfilter for any given prefilter and vice versa. The availability of those solutions suggests using them alternately until they converge to an optimal pair. Formally, this corresponds to the following.

Algorithm

Step 1: Set $i \leftarrow 0$ and $f_0(n) \leftarrow a\delta(n)$, with α chosen so that (19) is satisfied.

Step 2: Use (11)–(13) with $f(n) = f_i(n)$ and solve for the optimal postfilter $g(n)$. Set $g_i(n) = g(n)$.

Step 3: Set $\lambda = 0$ and use (16)–(22) to compute an optimal prefilter $f(n)$. Evaluate (19). If $P < 1$, go to Step 5; otherwise, go to the next step.

Step 4: Set λ to some positive value, solve (22), and update λ (by means of some technique for finding zeros of functions, e.g., Newton–Raphson's method [18]). Repeat the process until $P \simeq 1$.

Step 5: Compute Δ by

$$\Delta \equiv \max_{-L \leq l \leq L} |f_i(n)-f_{i-1}(n)|.$$

If Δ is sufficiently small, stop: the optimal pre- and postfilter are $f_i(n)$ and $g_i(n)$, respectively. Otherwise, set $i \leftarrow i+1$ and go back to Step 2. Alternatively, we could monitor the error level and stop whenever the error reduction from Step 4 is small enough.

The above algorithm is in the class of "coordinate descent" algorithms for minimization of functions of several variables [20], [19] since at each step it finds the unique global minimum of the error, with either the pre- or the postfilter coefficients kept fixed. Therefore, the algorithm necessarily converges to a stationary point of the Lagrangian [19], with a monotonic decrease in the error at each step. Unfortunately, there is no guarantee that the attained stationary point will be a global minimum; it could be a local minimum or a saddle point. However, as discussed in [3], our practical experience with the above algorithm has pointed out that stationary points tend to be well separated from each other, with large differences in their corresponding values of the error.

With the initial guess for the prefilter suggested in Step 1, we have never failed to obtain a correct solution for the optimal FIR filters with several different signal and noise spectra, but we did experience nonconvergence problems if the observer frequency response $W(\omega)$ got too close to zero for some frequency range since this leads to ill-conditioned matrices in (13) and (22).

The algorithm described above has a rate of convergence typical of coordinate descent methods, i.e., a weakly linear convergence [20] that is somewhat slower than that of the steepest descent algorithm [19]. Faster convergence, in terms of the number of iterations, could be obtained by using the steepest descent or Newton's methods. In either of these two alternative approaches, however, additional information would have to be computed, namely, the gradient of the error for the steepest descent method, and both the gradient and the Hessian for Newton's. For example, the number of operations required by the coordinate descent approach with $L = M = 8$ is approximately 6000 per iteration, whereas Newton's method requires about 200 000 operations per iteration (assuming, in both cases, that convolutions are performed by means of FFT's). Typically, the coordinate descent algorithm would have converged before a single iteration of Newton's method could be performed. Another advantage of the coordinate descent method besides its simplicity is that, at any iteration, we have at the end of Step 5 a "partially optimal" solution in the sense that at least the prefilter is optimal for the current postfilter, which is in turn optimal for the previous postfilter.

We end this section by deriving an expression for the value of the Lagrange multiplier λ at a jointly optimal solution; knowledge of this value can accelerate λ's convergence in Step 4. Using (13), we can write the error as

$$\xi_g = \xi_0 - \sum_{l=-L}^{L} f(l) \sum_{m=-M}^{M} g(m)\zeta(l-m) \tag{23}$$

where

$$\xi_0 \equiv \frac{1}{2\pi}\int_{-\pi}^{\pi} |W(\omega)|^2 \Phi_{xx}(\omega)\, d\omega \tag{24}$$

and $\zeta(n)$ is the inverse Fourier transform of $|W(\omega)|^2\Phi_{xx}(\omega)$.

From (22), we obtain

$$\xi_f = \xi_g + \frac{1}{2\pi}\int_{-\pi}^{\pi} |W(\omega)|^2 G^2(\omega)\tilde{\Phi}_{dd}(\omega)\, d\omega - \lambda P \tag{25}$$

where P is the prefilter output power. Since a jointly optimal pair satisfies both (13) and (22), we must have $\xi_g = \xi_f$, and so

$$\lambda_{\mathrm{OPT}} = \frac{1}{2\pi P}\int_{-\pi}^{\pi} |W(\omega)|^2 G^2(\omega)\tilde{\Phi}_{dd}(\omega)\, d\omega. \tag{26}$$

Thus, the optimal value of the Lagrange multiplier has a noise-to-signal ratio interpretation: it is the ratio of the filtered channel noise at the interpolator output (weighted by the observer response) to the available prefilter power.

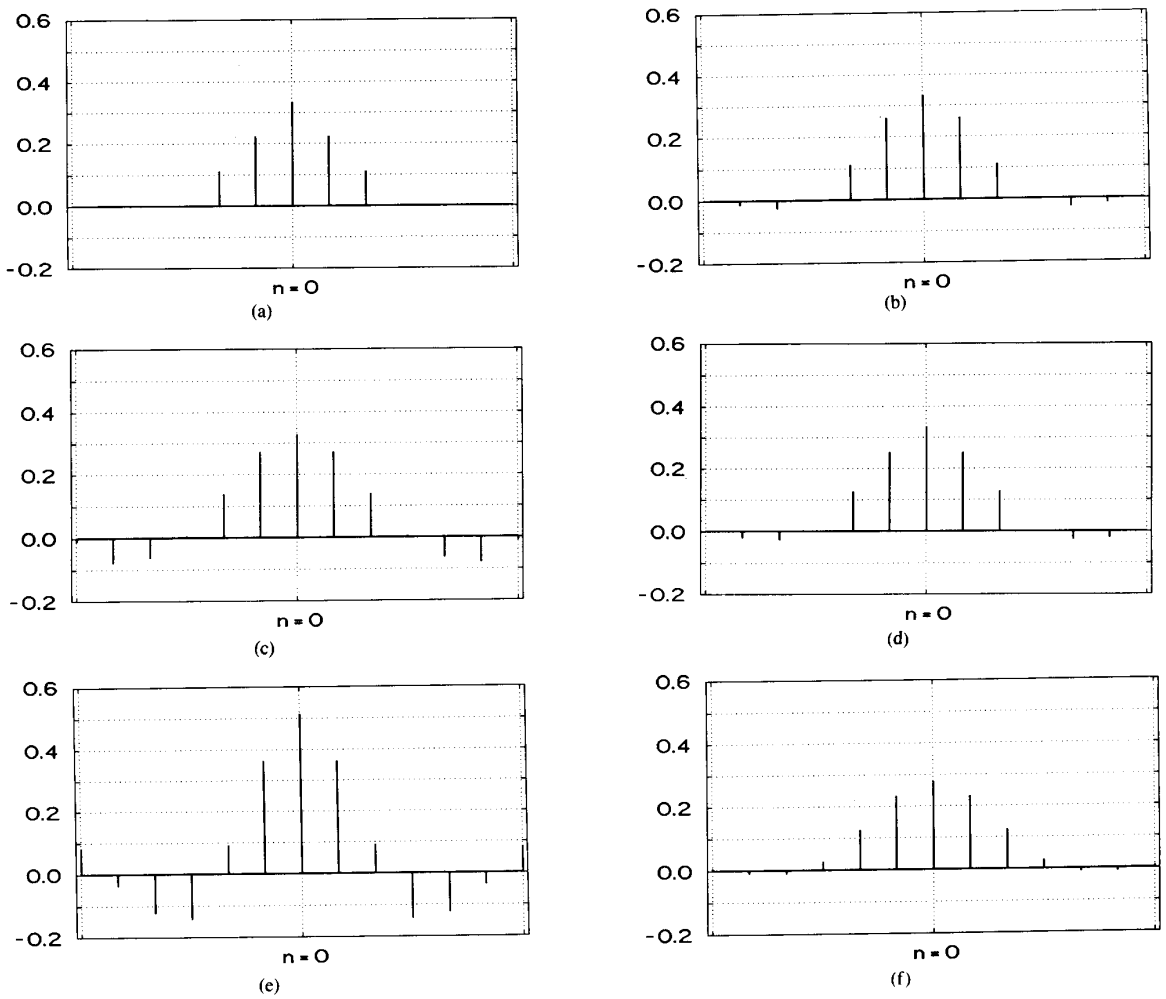

Fig. 2. Impulse responses of the filters evaluated for a system with a down-sampling factor $K = 3$: (a) linear prefilter, (b) cubic convolution postfilter, (c) Parks–McClellan prefilter, (d) Oetken, Parks, and Schüssler postfilter, (e), (f) jointly optimal pre- and postfilter for a flat observer ($W(\omega) \equiv 1$).

IV. Performance of Optimal FIR Filters

With the algorithm of the previous section, we can design a pair of jointly optimal pre- and postfilters for the system of Fig. 1. A natural question that arises at this point is: how much reduction in the MS reconstruction error can be achieved by using an optimal filter pair as compared to other commonly used pre- and postfilters? Although the answer to that question certainly depends on the particular system under consideration, we have performed a few image and speech processing experiments. A more detailed discussion of the performance of jointly optimal filter pairs can be found in [3].

Consider a system with a down- and up-sampling factor $K = 3$, and with pre- and postfilters of length 13, i.e., $L = M = 6$; furthermore, assume that the input and channel noise sources are white, with SNR's of 30 dB. In Fig. 2, we have the impulse responses of three pairs of pre- and postfilters that could be employed. The set in Fig. 2(a), (b) was chosen as one of the easiest to be designed: the prefilter is just a linear function, and the interpolator is a cubic convolution filter [9]. In Fig. 2(c), (d), we have a semi-optimal choice for the pre- and postfilters in the sense that each filter has been optimized under a certain criterion; the prefilter was designed using the Parks–McClellan algorithm [25] for equiripple approximation (passband ripple = 0.15, transition band from 0.267π to 0.4π), and the postfilter was obtained with the Oetken–Parks–Schüssler algorithm [7], [26] for optimal interpolator design. Finally, in Fig. 2(e), (f), the optimal pre- and postfilters for a flat observer ($W(\omega) \equiv 1$) are shown; they were computed by the algorithm of the previous section for an input signal with a first-order Gauss–Markov spectrum characterized by an inter-sample correlation coefficient $\rho = 0.95$.

We have processed the "KID" image of Fig. 3(a) with separable 2-D filters obtained from the filters in Fig. 2. The sampling grid was rectangular, and the down-sampling factor K was equal to three in both the horizontal and vertical directions. The results are shown in Fig. 3(b)–(d). The rms errors are indicated in Fig. 3 as a percentage of the signal rms value. The optimal filters led to an error improvement of 4.6 dB when compared to the linear–cubic convolution pair, and 1.7 dB when compared to the Parks–McClellan–Oetken pair. If we had chosen higher band-edge frequencies for the Parks–McClellan prefilter, for example, the mean-square error in Fig. 3(c) would have been higher. In general, a good choice for the parameters of the Parks–McClellan filter may require a trial-and-error approach.

Fig. 3. (a) Original "KID" image, 256 × 240 pixels, 8 bits/pixel. (b) "KID" processed with the linear prefilter and cubic convolution postfilter of Fig. 2(a) and (b), respectively; rms error = 19.3 percent. (c) "KID" processed with the Parks–McClellan prefilter and Oetken–Parks–Schüssler postfilter of Fig. 2(c) and (d), respectively; rms error = 13.7 percent. (d) "KID" processed with the mean-square-optimal pre- and postfilters of Fig. 2(e) and (f), respectively; rms error = 11.3 percent.

We have also processed a 120 ms speech segment with the filters in Fig. 2. The original segment, shown in Fig. 4(a), corresponds to the vowel "ah" spoken by a male person, sampled as 16 kHz. In Fig. 4(b)–(d), we have the error signals, magnified by a factor of six, due to processing the original segment with the pre- and postfilter pairs: linear-cubic convolution, Parks–McClellan–Oetken, and mean-square optimal, respectively. We note that the optimal filters led to an rms error about 6 dB below those of the other two filter pairs, the main reason for that being the virtual absence of low-frequency errors. For other segments, the improvement in MS error due to the optimal pre- and postfilters varied from 2 to 7 dB (typically, little improvement was obtained for unvoiced segments, for which most of the error is due to the loss of the high-frequency components). In [3], we show that the optimal FIR filter pair actually performs within 1–2 dB of the ideal IIR filters when $L, M \geq 2K$.

The above example also verifies the robustness of the optimal filters with respect to the input spectrum since a first-order Gauss–Markov process with a correlation coefficient of 0.95 is a good model for images but not for speech [27]. In fact, we have used optimal pre- and postfilters computed from an estimate of the input speech spectrum, and the rms error was only 0.3 dB below that of Fig. 3(d).

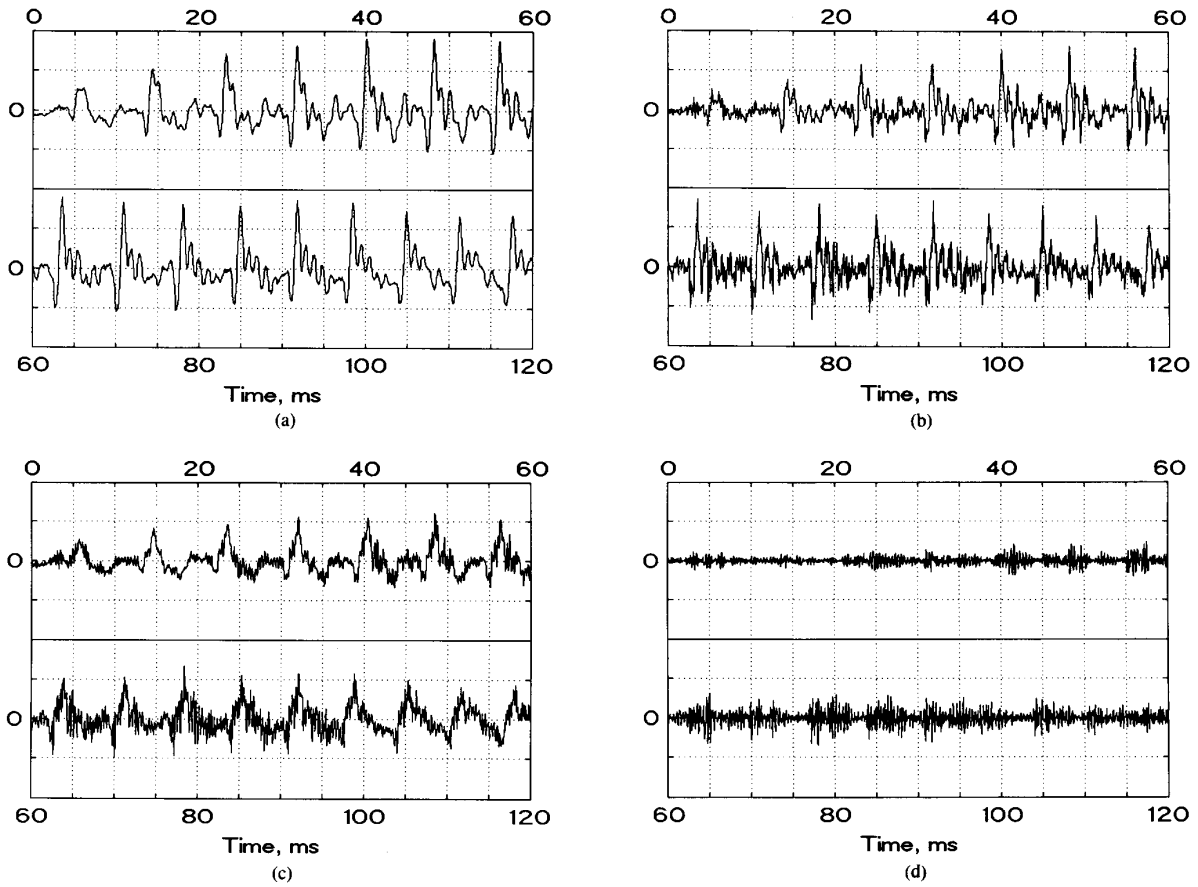

Fig. 4. (a) 120 ms speech segment for the vowel /a/, male speaker; the bottom trace is a continuation of the top one. (b) Error signals ($\times 6$) for the filter pairs in: Fig. 2(a), (b), rms amplitude = 12.6 percent; (c) Fig. 2(c), (d), rms amplitude = 11.5 percent; and (d) optimal filters of Fig. 2(e), (f), rms amplitude = 5.5 percent.

V. Concluding Remarks

We have presented in this paper an iterative algorithm for the design of jointly optimal FIR pre- and postfilters for a noisy communication/storage system under a weighted mean-square error criterion. As a byproduct, we have also derived the independent solutions for the optimal pre- or postfilters, which can be applied to systems in which one of those filters is predetermined by other factors. Although the algorithm is only guaranteed to converge to a local minimum of the error measure, in practice we have always obtained the correct solution. The good practical performance of the optimal FIR filters has been verified by means of speech and image processing examples.

The optimal FIR filters described in this paper will probably be most useful for sampling and interpolation systems where hardware cost is heavily dependent on the number of operations per second required by the filters, so that short-length FIR filters are a must. For image processing applications, we can extended the algorithm to the design of two-dimensional filters on arbitrary periodic sampling lattices. The general multidimensional versions of our proposed algorithm can be found in [3], [4].

References

[1] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1983.

[2] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975, ch. 5.

[3] H. S. Malvar, "Optimal pre- and post-filters in noisy sampled-data systems," Ph.D. dissertation, Dep. Elec. Eng., Massachusetts Inst. Technol., Cambridge, Aug. 1986. (Also as Tech. Rep. 519, Res. Lab. Electron., Massachusetts Inst. Technol., Aug. 1986.)

[4] H. S. Malvar and D. H. Staelin, "Statistical design of optimal multidimensional FIR filters for decimation and interpolation of random signals," in *Proc. IEEE Conf. Antennas Commun.*, Montreal, P.Q., Canada, Oct. 1986, pp. 330–333.

[5] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[6] W. A. Gardner and L. E. Franks, "Characterization of cyclostationary random signal processes," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 4–14, Jan. 1975.

[7] G. Oetken, T. W. Parks, and H. Schüssler, "New results in the design of digital interpolators," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 301–309, June 1975.

[8] A. D. Polydoros and E. N. Protonotarios, "Digital interpolation of stochastic signals," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 916–922, Nov. 1979.

[9] R. G. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 1153–1160, Dec. 1981.

[10] S. Kay, "Some new results in linear interpolation theory," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 746–749, June 1983.

[11] D. Radbel and R. J. Marks, II, "An FIR estimation filter based on the sampling theorem," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 455–460, Apr. 1985.

[12] R. E. Blahut, *Fast Algorithms for Digital Signal Processing*. Reading, MA: Addison-Wesley, 1985, ch. 11.
[13] F. G. Gustavson and D. Y. Y. Yun, "Fast algorithms for rational Hermite approximation and solution of Toeplitz systems," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 750–755, Sept. 1979.
[14] R. Kumar, "A fast algorithm for solving a Toeplitz system of equations," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 254–267, Feb. 1985.
[15] C. Manolakis, N. Kalouptsidis, and G. Carayannis, "Efficient determination of FIR Wiener filters with linear phase," *Electron. Lett.*, vol. 18, pp. 429–431, May 13, 1982.
[16] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*. Los Angeles, CA: Univ. California Press, 1958, ch. 10.
[17] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. New York: Academic, 1982, ch. 2.
[18] G. Dahlquist and A. Björk, *Numerical Methods*. Englewood Cliffs, NJ: Prentice-Hall, 1974, ch. 6.
[19] D. G. Luenberger, *Linear and Non-Linear Programming*. Reading, MA: Addison-Wesley, 1984, ch. 7.
[20] T. Abatzoglou and B. O'Donnel, "Minimization by coordinate descent," *J. Optimiz. Theory Appl.*, vol. 36, pp. 163–174, Feb. 1982.
[21] P. R. Chevillat and G. Ungerboeck, "Optimum FIR transmitter and receiver filters for data transmission over band-limited channels," *IEEE Trans. Commun.*, vol. COM-30, pp. 1909–1915, Aug. 1982.
[22] R. Hummel, "Sampling for spline reconstruction," *SIAM J. Appl. Math.*, pp. 278–288, Apr. 1983.
[23] J. N. Ratzel, "The discrete representation of spatially continuous images," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, 1980.
[24] P. Faubert, "Optimisation conjointe du pré-filtre et du post-filtre pour la decimation et l'interpolation des signaux numeriques multidimensionnels," M.Sc. thesis, Univ. Québec, Québec, Canada, 1985.
[25] J. H. McClellan, T. W. Parks, and L. R. Rabiner, "FIR linear phase filter design program," in *Programs for Digital Signal Processing*. New York: IEEE Press, 1979, sect. 5.1.
[26] G. Oetken, "A computer program for digital interpolator design," in *Programs for Digital Signal Processing*. New York: IEEE Press, 1979, sect. 8.1.
[27] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978, ch. 8.

Henrique S. Malvar (M'79–S'82–M'86) was born in Rio de Janeiro, Brazil, in 1957. He received the B.S. degree in 1977 from the Universidade de Brasilia, Brazil, the M.S. degree in 1979 from the Universidade Federal do Rio de Janeiro, Brazil, and the Ph.D. degree in 1986 from the Massachusetts Institute of Technology, Cambridge, all in electrical engineering.

Since 1979 he has been on the faculty of the Universidade de Brasilia, Brazil, where he is now an Associate Professor of Electrical Engineering. From 1982 to 1987 he was on leave with the Massachusetts Institute of Technology, where he had been a doctoral student from 1982 to 1986, and a Visiting Assistant Professor of Electrical Engineering for the 1986–1987 academic year. From 1985 to 1987 he had also been a consultant with PictureTel Corporation, Peabody, MA, where he was part of the design team for the C-2000 video codec. His teaching, research, and consulting interests include analog and digital filter synthesis, digital signal processing, and low-bandwidth coding of video and speech. He has several publications in these areas, and pending patent applications in the U.S.A, Canada, Europe, and Japan.

Dr. Malvar is a member of Sigma Xi. He was the recipient of the Young Scientist Award from the Marconi International Fellowship in 1981.

★

David H. Staelin (S'59–M'65–SM'75–F'79) was born in Toledo, OH, in 1938. He received the S.B., S.M., and Sc.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge.

He has been on the faculty at M.I.T. since 1965 and is now a Professor of Electrical Engineering. His current research involves video image processing, passive microwave remote sensing, optical stellar interferometry, and radio astronomy. His teaching is in the areas of signal processing and electromagnetic radiation, and his consulting is in the areas of satellite and video communications, and remote sensing. He is a founder of PictureTel Corporation, Peabody, MA, and served as Chairman from 1984 to 1987.