A Weighted Least-Squares Method for the Design of Stable 1-D and 2-D IIR Digital Filters

Wu-Sheng Lu, Soo-Chang Pei, *Senior Member, IEEE*, and Chien-Cheng Tseng, *Member, IEEE*

***Abstract*— In this paper, we present a new approach to the least-squares design of stable infinite impulse response (IIR) digital filters. The design is accomplished by using an iterative scheme in which the denominator polynomial obtained from the preceding iteration is treated as a part of the weighting function, and each iteration is carried out by solving a standard quadratic programming problem that yields a stable rational function. When the iteration converges, a stable and truly least-squares solution is obtained. The method is then extended to address the least-squares design of stable IIR two-dimensional (2-D) filters. Examples are included to illustrate the proposed design techniques.**

***Index Terms*— IIR filter, quadratic programming, weighted least-squares method.**

I. Introduction

LEAST-SQUARES methods have been extensively used for dealing with various analysis and synthesis problems in science and engineering. In a digital filter context, many least-squares design methods have been proposed; see [1]–[8] and [16] among others. A literature survey shows that most of successful least-squares designs are for the finite impulse response (FIR) filters. A major problem with the existing least-squares techniques as applied to the design of infinite impulse response (IIR) filters is that the least squares cost function needs to be modified to avoid the division operation introduced by the rational transfer function of the filter so that the modified cost function can be explicitly expressed as a quadratic form with respect to the filter parameters. In doing so, however, the solution obtained by minimizing the modified cost function is no longer a truly least-squares design. Moreover, even for this sort of quasi-least-squares solution, stability of the filter designed is not guaranteed.

In this paper, we present a new approach to the least-squares design of stable IIR digital filters. The design is accomplished using an iterative scheme in which the denominator polynomial obtained from the preceding iteration is treated as a part of the weighting function, and each iteration is carried out by solving a standard quadratic programming problem that yields a stable rational function. When the iteration converges, a stable and truly least-squares solution is obtained. Moreover, by adequately adjusting the weighting function, the proposed method can also be used to design nearly equiripple IIR filters with guaranteed stability. The method is then extended to address the least-squares design of stable IIR two-dimensional (2-D) filters. Examples are included to illustrate the proposed design techniques.

Manuscript received July 29, 1995; revised June 26, 1997. This work was supported by the National Science Council, R.O.C, under Contract NSC84-2811-E002-076. The associate editor coordinating the review of this paper and approving it for publication was Dr. Victor E. DeBrunner.

W.-S. Lu is with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, B.C., Canada V8W 3P6.

S.-C. Pei and C.-C. Tseng are with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, R.O.C.

Publisher Item Identifier S 1053-587X(98)00499-1.

II. The Weighted Least-Squares Method

A. Motivation

Let $F_d(\omega)$ be the desired frequency response specified in $[0, \pi)$. We seek to find a causal stable rational function $F(z) = N(z)/D(z)$ that best approximates $F_d(\omega)$ in the weighted L_2-norm sense. For the sake of notational simplicity, we denote

$$\begin{aligned} D(z) &= 1 + \mathbf{q}_1^t(z)\mathbf{d} \\ N(z) &= \mathbf{q}_2^t(z)\mathbf{a} \end{aligned} \tag{1}$$

where $\mathbf{d} = [d_1 \cdots d_n]^t$, $a = [a_0 \ a_1 \ \cdots \ a_n]^t$, $\mathbf{q}_1(z) = [z^{-1} \cdots z^{-n}]^t$, and $\mathbf{q}_2(z) = [1 \ z^{-1} \cdots z^{-n}]^t$. With a given weighting function $W(\omega)$, the weighted L_2 cost function is defined by

$$J(\mathbf{d}, \mathbf{a}) = \tfrac{1}{2} \int_0^{\pi} W(\omega)|F_d(\omega) - F(e^{j\omega})|^2 \, d\omega. \tag{2}$$

Note that $J(\mathbf{d}, \mathbf{a})$ in (2) can be expressed as

$$J(\mathbf{d}, \mathbf{a}) = \frac{1}{2} \int_0^{\pi} \frac{W(\omega)}{|D(e^{j\omega})|^2} |F_d(\omega)D(e^{j\omega}) - N(e^{j\omega})|^2 \, d\omega. \tag{3}$$

Most of the least-squares algorithms in the literature then neglect the term $|D(e^{j\omega})|^2$ underneath $W(\omega)$ in (3), leading to the following modified cost function:

$$\hat{J}(\mathbf{d}, \mathbf{a}) = \tfrac{1}{2} \int_0^{\pi} W(\omega)|F_d(\omega)D(e^{j\omega}) - N(e^{j\omega})|^2 \, d\omega. \tag{4}$$

Evidently, the rational function $N(z)/D(z)$ that minimizes $\hat{J}(\mathbf{d}, \mathbf{a})$ does not necessarily minimize (2); hence, the solution obtained by minimizing (4) is, in general, not optimal in the least-squares sense. The approach taken in this paper treats the

term $W(\omega)/|D(e^{j\omega})|^2$ in (3) as a new and known weighting function to form a standard least-squares minimization problem. To justify this treatment, polynomial $D(z)$ must be stable so that $W(\omega)/|D(e^{j\omega})|^2$ will be a well-defined weighting, and $|D(e^{j\omega})|$ must be known. The stability problem encountered here is solved by imposing a set of linear constraints on the coefficients of $D(z)$, which ensures that all zeros of $D(z)$ are inside the open unit disk. Note that minimizing a quadratic form subject to a set of linear constraints is a typical optimization problem known as quadratic programming, to which reliable solutions can be obtained using well established techniques [9]. Furthermore, the term $|D(e^{j\omega})|^2$ underneath $W(\omega)$ in (3) is made available by adopting the following iterative scheme:

$$J_k(\mathbf{d}^{(k)}, \mathbf{a}^{(k)}) = \frac{1}{2}\int_0^{\pi} \frac{W(\omega)}{|D_{k-1}(e^{j\omega})|^2} |F_d(\omega)D_k(e^{j\omega}) - N_k(e^{j\omega})|^2 \, d\omega \tag{5}$$

where $k = 1, 2, \cdots$, $D_k(e^{j\omega}) = 1 + \mathbf{q}_1^t(\omega)\mathbf{d}^{(k)}$, $N_k(e^{j\omega}) = \mathbf{q}_2^t(\omega)\mathbf{a}^{(k)}$, $\mathbf{d}^{(k)}$, and $\mathbf{a}^{(k)}$ are the parameter vectors to be determined in the kth iteration. The initial parameter vector $\mathbf{d}^{(0)}$ can be chosen quite arbitrarily, except that $D_0(z)$ must be stable, i.e., the zeros of $D(z)$ must be inside the open unit disk in the z plane. For example, one may chose $\mathbf{d}^{(0)} = [0\ 0\ \cdots\ 0]^t$. At the kth iteration, $D_{k-1}(e^{j\omega})$ is known and stable; hence, (5) can be written as

$$J_k(\mathbf{d}^{(k)}, \mathbf{a}^{(k)}) = \tfrac{1}{2}\int_0^{\pi} W_k(\omega) |F_d(\omega)D_k(e^{j\omega}) - N_k(e^{j\omega})|^2 \, d\omega \tag{6}$$

where

$$W_k(\omega) = \frac{W(\omega)}{|D_{k-1}(e^{j\omega})|^2} \tag{7}$$

is, for a stable $D_{k-1}(e^{j\omega})$, a well-defined, nonnegative weighting function. Obviously, finding vectors $\mathbf{d}^{(k)}$ and $\mathbf{a}^{(k)}$ that minimize $J_k[\mathbf{d}^{(k)}, \mathbf{a}^{(k)}]$ in (6) has a standard procedure to follow. Details of this procedure subject to a set of stability constraints are given in parts B and C of this section. For the moment, let us assume that the sequence of parameter vectors $\{\mathbf{d}^{(k)}, \mathbf{a}^{(k)}\}$ has been generated by iteratively solving the standard least-squares problem (6) subject to a set of stability constraints and that

$$\mathbf{d}^{(k)} \to \mathbf{d} \text{ and } \mathbf{a}^{(k)} \to \mathbf{a}. \tag{8}$$

Then, it follows that

$$D_k(e^{j\omega}) \to D(e^{j\omega}) \text{ and } N_k(e^{j\omega}) \to N(e^{j\omega}) \tag{9}$$

and $\{\mathbf{d}, \mathbf{a}\}$ minimizes the limit of $J_k[\mathbf{d}^{(k)}, \mathbf{a}^{(k)}]$, namely

$$J(\mathbf{d}, \mathbf{a}) = \lim_{k\to\infty} J_k[\mathbf{d}^{(k)}, \mathbf{a}^{(k)}] \tag{10}$$

where $J(\mathbf{d}, \mathbf{a})$ is defined by (2).

Because of the presence of denominator $D(z)$ in (2), the objective function $J(\mathbf{d}, \mathbf{a})$ is *nonquadratic* and has, in general, multiple minimum points. Consequently, one can only expect the limiting point $\mathbf{x} = [\mathbf{d}^t\ \mathbf{a}^t]^t$ in (8) to be a local minimum point of $J(\mathbf{d}, \mathbf{a})$. To show this, note first that if

$$\mathbf{x}^{(k)} = \begin{bmatrix} \mathbf{d}^{(k)} \\ \mathbf{a}^{(k)} \end{bmatrix} \tag{11}$$

minimizes $J_k(\hat{\mathbf{x}}^{(k)})$, then

$$\left.\frac{\partial J_k[\hat{\mathbf{x}}^{(k)}]}{\partial \hat{\mathbf{x}}^{(k)}}\right|_{\hat{\mathbf{x}}^{(k)}=\mathbf{x}^{(k)}} = 0. \tag{12}$$

This, in conjunction with (8) and (10), implies that

$$\left.\frac{\partial J(\hat{\mathbf{x}})}{\partial \hat{\mathbf{x}}}\right|_{\hat{\mathbf{x}}=\mathbf{x}} = \lim_{k\to\infty} \left.\frac{\partial J_k[\hat{\mathbf{x}}^{(k)}]}{\partial \hat{\mathbf{x}}^{(k)}}\right|_{\hat{\mathbf{x}}^{(k)}=\mathbf{x}^{(k)}} = 0. \tag{13}$$

Hence, the limiting point $\mathbf{x}$ is a stationary point of the objective function $J(\mathbf{x})$. Next, we consider a neighborhood of the limiting point $\mathbf{x}$, which is denoted by B_x. By (6), it follows that for a fixed index k, J_k is *globally* convex with respect to $\mathbf{x}^{(k)}$. This is simply because in such a case, the weight $W_k(\omega)$ is known and always nonnegative for $\omega \in [0, \pi)$, and J_k is a quadratic function with a nonnegative definite Hessian matrix. If we assume that for sufficiently large k, J_k is also uniformly positive definite in B_x in the sense that for sufficiently large k there exists $\beta > 0$ such that

$$\hat{\mathbf{x}}^t\mathbf{H}_k\hat{\mathbf{x}} \geq \beta\|\hat{\mathbf{x}}\|^2 \qquad \text{for } \hat{\mathbf{x}} \in B_x \tag{14}$$

then as $k \to \infty$, we have

$$\hat{\mathbf{x}}^t\mathbf{H}(\mathbf{x})\hat{\mathbf{x}} \geq \beta\|\hat{\mathbf{x}}\|^2 \qquad \text{for } \hat{\mathbf{x}} \in B_x \tag{15}$$

where $\mathbf{H}(\mathbf{x})$ is the Hessian matrix of $J(\mathbf{d}, \mathbf{a})$ at the limiting point $\mathbf{x}$. Thus $J(\mathbf{d}, \mathbf{a})$ is strictly convex at $\mathbf{x}$, and hence, $\mathbf{x}$ is a local minimum point of J.

As a further remark, it is important to stress that different limiting points may be obtained when different initial points are chosen simply because $J(\mathbf{x})$ is not globally convex. On the other hand, our numerical experience indicates that satisfactory design can be achieved even when the design algorithm starts with the trivial initial $\mathbf{d}^{(0)} = [0 \cdots 0]^t$.

B. An Explicit Expression for $J_k[\mathbf{d}^{(k)}, \mathbf{a}^{(k)}]$ in Terms of $\mathbf{d}^{(k)}$ and $\mathbf{a}^{(k)}$

Letting $F_d(\omega)$ be the complex-valued desired frequency response, we can write

$$\begin{aligned} &|F_d(\omega)D_k(e^{j\omega}) - N_k(e^{j\omega})|^2 \\ &\quad = \mathbf{d}^{(k)^t}[|F_d(\omega)|^2\mathbf{Q}_{11}(\omega)]\mathbf{d}^{(k)} + \mathbf{a}^{(k)^t}\mathbf{Q}_{22}(\omega)\mathbf{a}^{(k)} \\ &\qquad - 2\mathbf{d}^{(k)^t}\mathbf{Q}_{12}(\omega)\mathbf{a}^{(k)} + 2\mathbf{d}^{(k)^t}[|F_d(\omega)|^2\tilde{\mathbf{q}}_1(\omega)] \\ &\qquad - 2\mathbf{a}^{(k)^t}\tilde{\mathbf{q}}_2(\omega) + |F_d(\omega)|^2 \end{aligned} \tag{16}$$

where

$$\mathbf{Q}_{11}(\omega) = \begin{bmatrix} 1 & \cos(\omega) & \cdots & \cos[(n-1)\omega] \\ \cos(\omega) & 1 & \cdots & \cos[(n-2)\omega] \\ \vdots & & \ddots & \vdots \\ \cos[(n-1)\omega] & \cos[(n-2)\omega] & \cdots & 1 \end{bmatrix}$$

$$\mathbf{Q}_{22}(\omega) = \begin{bmatrix} 1 & \cos(\omega) & \cdots & \cos(n\omega) \\ \cos(\omega) & & & \\ \vdots & & \mathbf{Q}_{11}(\omega) & \\ \cos(n\omega) & & & \end{bmatrix}$$

$$\mathbf{Q}_{12}(\omega) = \tfrac{1}{2}[F_d(\omega)\mathbf{q}_1(\omega)\overline{\mathbf{q}}_2^t(\omega) + \overline{F}_d(\omega)\overline{\mathbf{q}}_1(\omega)\mathbf{q}_2^t(\omega)]$$

$$\tilde{\mathbf{q}}_1(\omega) = [\cos(\omega) \quad \cos(2\omega) \cdots \cos(n\omega)]^t$$

$$\tilde{\mathbf{q}}_2(\omega) = \tfrac{1}{2}[F_d(\omega)\overline{\mathbf{q}}_2(\omega) + \overline{F}_d(\omega)\mathbf{q}_2(\omega)]$$

with $\overline{\mathbf{q}}_1$, $\overline{\mathbf{q}}_2$, and $\overline{F}_d$ denoting the conjugate of $\mathbf{q}_1$, $\mathbf{q}_2$, and F_d, respectively. Note that both $\mathbf{Q}_{11}(\omega)$ and $\mathbf{Q}_{22}(\omega)$ are symmetric Toeplitz matrices characterized by their first columns. Let $\Omega_L = \{\omega_i, i = 1, \cdots, L\}$ be the set of the equally spaced frequencies on Ω over which the integral in (6) is evaluated with $L \geq 4n$, as suggested in [11] for sufficient degree of accuracy; then, the cost function $J_k[\mathbf{d}^{(k)}, \mathbf{a}^{(k)}]$ can be approximated explicitly in terms of $\mathbf{d}^{(k)}$ and $\mathbf{a}^{(k)}$ in a standard positive-definite quadratic form as

$$J_k[\mathbf{d}^{(k)}, \mathbf{a}^{(k)}] \approx \tfrac{1}{2}\mathbf{d}^{(k)^t}\mathbf{K}_{11}\mathbf{d}^{(k)} + \tfrac{1}{2}\mathbf{a}^{(k)^t}\mathbf{K}_{22}\mathbf{a}^{(k)} - \mathbf{d}^{(k)^t}\mathbf{K}_{12}\mathbf{a}^k + \mathbf{d}^{(k)^t}\mathbf{b}_1 - \mathbf{a}^{(k)^t}\mathbf{b}_2 + c \quad (17)$$

where

$$\mathbf{K}_{11} = \Delta \sum_{i=1}^{L} W_k(\omega_i)|F_d(\omega_i)|^2\mathbf{Q}_{11}(\omega_i)$$

$$\mathbf{K}_{12} = \Delta \sum_{i=1}^{L} W_k(\omega_i)\mathbf{Q}_{12}(\omega_i)$$

$$\mathbf{K}_{22} = \Delta \sum_{i=1}^{L} W_k(\omega_i)\mathbf{Q}_{22}(\omega_i)$$

$$\mathbf{b}_1 = \Delta \sum_{i=1}^{L} W_k(\omega_i)|F_d(\omega_i)|^2\tilde{\mathbf{q}}_1(\omega_i)$$

$$\mathbf{b}_2 = \Delta \sum_{i=1}^{L} W_k(\omega_i)\tilde{\mathbf{q}}_2(\omega_i)$$

$$W_k(\omega_i) = \frac{W(\omega_i)}{|D_{k-1}(\omega_i)|^2},$$

where c is a constant independent of $\mathbf{d}^{(k)}$ and $\mathbf{a}^{(k)}$, and Δ is the increment for numerical approximation of J_k in (6). With

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{11} & -\mathbf{K}_{12} \\ -\mathbf{K}_{12}^t & \mathbf{K}_{22} \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ -\mathbf{b}_2 \end{bmatrix} \quad \mathbf{x}^{(k)} = \begin{bmatrix} \mathbf{d}^{(k)} \\ \mathbf{a}^{(k)} \end{bmatrix} \quad (18)$$

we can write

$$J_k[\mathbf{x}^{(k)}] \approx \tfrac{1}{2}\mathbf{x}^{(k)^t}\mathbf{K}\mathbf{x}^{(k)} + \mathbf{b}^t\mathbf{x}^{(k)} + c. \quad (19)$$

It follows from the definitions of $\mathbf{K}$ that the quadratic form $J_k[\mathbf{x}^{(k)}]$ is positive definite.

C. The Stability Issue

The rational function generated from the kth iteration must be stable. It is well known [10], [11] that $D_k(z)$ is stable if

$$\mathrm{Re}[D_k(e^{j\omega})] > 0 \qquad \text{for } \omega \in [0, \pi] \quad (20)$$

where $\mathrm{Re}[D_k(e^{j\omega})]$ denotes the real part of $D_k(e^{j\omega})$. It is important to note that (20) offers only a sufficient condition for the stability; hence, the set of polynomials D_k satisfying (20) is a nontrivial subset of the set of all stable polynomials of order n. Consequently, the solution obtained by minimizing J_k in (19) subject to constraint (20) can only be claimed as suboptimal as it is possible that the optimal solution may have excluded by constraint (20) in the minimization process. On the other hand, however, (20) is less conservative than other known linear constraints that ensure the stability of D_k [11], [12]. Refer to [11, Appendixes I and II] for a detailed analysis on this matter. In practice, (20) is replaced by

$$\mathrm{Re}[D_k(e^{j\omega})] \geq \delta \qquad \omega \in [0, \pi] \quad (21)$$

where δ is a small and positive number and is then implemented on a dense grid of points over $[0, \pi]$. Let $\Theta = \{\Omega_i, i = 1, \cdots, M\}$ be the set of grid points on $[0, \pi]$. Using matrix notation, condition (21) on set Θ becomes

$$\mathbf{B}\mathbf{x}^{(k)} \leq (1 - \delta)\mathbf{e}_{2n+1} \quad (22)$$

where

$$\mathbf{B} = -\begin{bmatrix} \mathbf{q}_1^t(e^{j\Omega_1}) & \\ \vdots & 0 \\ \mathbf{q}_1^t(e^{j\Omega_M}) & \end{bmatrix}_{M\times(2n+1)} \qquad \mathbf{e}_{2n+1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{M\times 1}. \quad (23)$$

Thus, the task at the kth iteration is to minimize $J_k[\mathbf{x}^{(k)}]$ in (19) subject to constraints (22), which is a typical quadratic programming problem. With a positive definite $\mathbf{K}$, the solution of (19) and (22) can be computed efficiently. For example, one can convert the problem at hand first to a so-called least distance programming problem, which can be further converted to a non-negative least squares (NNLS) problem, and the NNLS problem can then be solved using an "active set" method. Refer to [9, ch. 23] for the details of this method.

D. The Constrained Least-Squares Design

As was mentioned in Section II-A, the iteration begins by choosing an initial $\mathbf{d}^{(0)} = [0 \cdots 0]^t$. Matrices $\mathbf{K}$, $\mathbf{b}$, and $\mathbf{B}$ are then evaluated, and $\mathbf{x}^{(1)}$ is computed as the solution of the quadratic programming problem (19), (22). In the next iteration, $\mathbf{d}^{(1)}$ is utilized to update matrices $\mathbf{K}$ and $\mathbf{b}$, and then, $\mathbf{x}^{(2)}$ is computed as the solution of the same quadratic programming problem. The iteration continues until $||\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}||$ is less than a prescribed tolerance ε. At that

time, the convergence is claimed, and $\mathbf{x}^{(k)}$ is deemed as the solution of the constrained least-squares problem.

There are two issues that need to be addressed here: i) whether the design algorithm described above always converges and ii) if the algorithm does converge, whether the limiting point is a solution to the constrained optimization problem

$$\begin{aligned} &\underset{\mathbf{x}}{\text{Minimize}}\ J(\mathbf{x}) \\ &\text{Subject to } \mathbf{B}\mathbf{x} \le (1-\delta)\mathbf{e}_{2n+1}. \end{aligned} \tag{24}$$

Let us consider the second issue first. Assume $\mathbf{x}^{(k)} \to \mathbf{x}$ as $k \to \infty$, where $\mathbf{x}^{(k)}$ minimizes J_k in (19) subject to constraint (22). It follows that $\mathbf{x}^{(k)}$ satisfies the Kuhn–Tucker (KT) conditions [17, ch. 10], which are a set of necessary conditions for $\mathbf{x}^{(k)}$ to be the solution of problem (19), (22). As $k \to \infty$, the assumption that $\mathbf{x}^{(k)} \to \mathbf{x}$ in conjunction with the smoothness of the objective function J_k with respect to $\mathbf{x}^{(k)}$ now implies that the KT conditions for $\mathbf{x}^{(k)}$ converge to the KT conditions for the limiting point $\mathbf{x}$. In addition, as $k \to \infty$, the constraints in (22) become a set of linear inequality constraints that guarantee the stability of the IIR filter produced by the limiting point $\mathbf{x}$. Furthermore, by using an argument similar to that made in Section II-A, one can show that the objective function $J(\mathbf{x})$ is strictly convex at the limiting point $\mathbf{x}$, provided that the Hessian matrices of J_k's are uniformly positive definite for sufficiently large k. Under these circumstances, one concludes that $\mathbf{x}$ is a local solution of the constrained optimization problem (24) [17]. It is important to stress that the term "weighted least-squares design" used in the rest of the paper is referred to a local solution of (24). The local nature of the solution is primarily due to the high degree of nonlinearity of $J(\mathbf{x})$, and the complexity of the stability requirement has forced us to consider a set of linear sufficient constraints in (24). Consequently, a solution to (24) can only be deemed suboptimal.

Let us now address the convergence issue with two remarks. The algorithm proposed above converges with a wide range of initial points. However, we have also identified a very limited number of occasions where the algorithm converges rather slowly or does not converge, at least within a reasonable number of iterations, say, 50. A simple modification of the algorithm described below is found to be effective in improving the robustness of the algorithm. Let Φ be the operator that maps an initial point to the solution of the quadratic programming problem in (19) and (22). The iterative algorithm can then be described by

$$\mathbf{x}^{(k+1)} = \Phi(\mathbf{x}^{(k)}). \tag{25}$$

We now modify this to

$$\mathbf{u}^{(k)} = \Phi[\mathbf{x}^{(k)}])$$

and

$$\mathbf{x}^{(k+1)} = \alpha\mathbf{u}^{(k)} + (1-\alpha)\mathbf{x}^{(k)} \tag{26}$$

where $0 < \alpha < 1$ is a relaxation constant. In other words, the next point $\mathbf{x}^{(k+1)}$ is obtained by combining the solution of (19) and (22) with the initial point used. A large α means that point $\mathbf{x}^{(k+1)}$ relies more on solution $\mathbf{u}^{(k)}$ so that the algorithm would converge faster at the risk of numerical instability, whereas a small α tends to stabilize the algorithm by using more information from the preceding iteration result at the expense of reduced convergence rate. An α in the range [0.3, 0.5] is often found appropriate, and the modified algorithm with a $\alpha \in [0.3, 0.5]$ was successfully used to design a variety of digital filters with different (but stable) initial points. It should be pointed out that this algorithm modification was motivated by several recent design methods for multirate systems [18], [19], where a similar relaxation technique was used to improve convergence of the algorithms, although the design scenario there differs considerably from ours.

As the second remark, we present a sufficient condition for the convergence of the modified algorithm. Define the ratio

$$\eta_k = \frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|}{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}. \tag{27}$$

It can be shown that if η_k has a less-than-unity upper bound, i.e.,

$$\eta_k \le \gamma < 1 \tag{28}$$

for $k \ge L$ where L is a positive integer, then sequence $\{\mathbf{x}^{(k)}\}$ converges. As a matter of fact, this condition implies that

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \le \gamma\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|. \tag{29}$$

Therfore, for sufficiently large m and n with $m > n \ge L$, we have

$$\|\mathbf{x}^{(m)} - \mathbf{x}^{(n)}\| \le \frac{\gamma^{n-L+1} - \gamma^{m-L+1}}{1-\gamma}\,\|\mathbf{x}^{(L)} - \mathbf{x}^{(L-1)}\| \tag{30}$$

which approaches zero when $m, n \to \infty$, and hence, $\{\mathbf{x}^{(k)}\}$ is a Cauchy sequence in a finite-dimensional Euclidean space. Further, notice that the above sufficient condition is equivalent to

$$\|\Phi[\mathbf{x}^{(k)}] - \Phi[\mathbf{x}^{(k-1)}]\| \le \beta\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \tag{31}$$

for a $\beta \in (0, 1)$. In other words, $\{\mathbf{x}^{(k)}\}$ is convergent if Φ is a *contraction mapping* when it applies to the sequence produced by $\mathbf{x}^{(k+1)} = \Phi[\mathbf{x}^{(k)}]$. Although a rigorous proof is not available to date, with $\alpha \in [0.3, 0.5]$, the modified algorithm was always successful in producing $\{\mathbf{x}^{(k)}\}$ with ratio $\eta_k \le \gamma < 1$ in our extensive simulation study.

E. A Quasi-Equiripple Design

Like the approach used in [6], the weighting function $W(\omega)$ can be updated properly to achieve a quasi-equiripple design. At each iteration, one can use a second iteration loop as applied to the weighting function $W(\omega)$ so that a nearly equiripple design can be achieved. In this second iteration loop, $W(\omega)$ is updated on each frequency band contained in Ω, namely, $W_{l+1}(\omega) = W_l(\omega)\nu_l(\omega)$, where $\nu_l(\omega) > 0$ is determined by the extreme values of the design error at the lth iteration in conjunction with an interpolation technique that ensures $\nu_l(\omega) \ne 0$. The second iteration loop is terminated when the extreme values of the design error are nearly equal. See [6] for the details of this design technique.

III. Extension to the Two-Dimensional Case

In the 2-D case, we seek to find a stable 2-D rational function of order (n_1, n_2)

$$F(z_1, z_2) = \frac{N(z_1, z_2)}{D(z_1, z_2)} \tag{32}$$

that best approximates the desired frequency response $F_d(\omega_1, \omega_2)$ given on $\Omega_2 = [-\pi, \pi) \times [-\pi, \pi)$ in weighted $L_2(\Omega)$ norm.

Using vector notation, the denominator and numerator of $F(z_1, z_2)$ can be expressed as

$$\begin{aligned} D(z_1, z_2) &= 1 + \mathbf{p}_1^t(z_1, z_2)\mathbf{r} \\ N(z_1, z_2) &= \mathbf{p}_2^t(z_1, z_2)\mathbf{s} \end{aligned}$$

where

$$\begin{aligned} \mathbf{p}_1(z_1, z_2) &= [z_1^{-1} \cdots z_1^{-n_1}\ z_2^{-1}\ z_1^{-1}z_2^{-1} \cdots z_1^{-n_1}z_2^{-1} \\ &\qquad \cdots z_1^{-n_1}z_2^{-n_2}]^t \\ \mathbf{p}_2(z_1, z_2) &= [1\ \mathbf{p}_1^t(z_1, z_2)]^t \\ \mathbf{r} &= [r_1 \cdots r_m]^t \\ \mathbf{s} &= [s_1 \cdots s_{m+1}]^t \\ m &= (n_1 + 1)(n_2 + 1) - 1. \end{aligned} \tag{33}$$

Define the weighted L_2 cost function by

$$\begin{aligned} J_2(\mathbf{r}, \mathbf{s}) = \tfrac{1}{2} \int\int_{\Omega_2} & W_2(\omega_1, \omega_2)|F_d(\omega_1, \omega_2) \\ & - F(\omega_1, \omega_2)|^2\, d\omega_1\, d\omega_2 \end{aligned} \tag{34}$$

where $W_2(\omega_1, \omega_2) \geq 0$ is the weighting function given on Ω_2, and $F(\omega_1, \omega_2)$ is characterized by (32). Like the 1-D case, (34) can be expressed as

$$\begin{aligned} J_2(\mathbf{r}, \mathbf{s}) = \frac{1}{2} \int\int_{\Omega_2} & \frac{W_2(\omega_1, \omega_2)}{|D(\omega_1, \omega_2)|^2}\, |F_d(\omega_1, \omega_2)D(\omega_1, \omega_2) \\ & - N(\omega_1, \omega_2)|^2\, d\omega_1\, d\omega_2 \end{aligned} \tag{35}$$

which suggests the following iterative least-squares scheme for the design problem at hand:

$$\begin{aligned} J_{2k}(\mathbf{r}^{(k)}, \mathbf{s}^{(k)}) = \tfrac{1}{2} \int\int_{\Omega_2} & W_{2k}(\omega_1, \omega_2)|F_d(\omega_1, \omega_2)D_k(\omega_1, \omega_2) \\ & - N_k(\omega_1, \omega_2)|^2\, d\omega_1\, d\omega_2 \end{aligned} \tag{36}$$

with $k = 1, 2, \cdots$, $\mathbf{r}^{(0)} = [0 \cdots 0]^t$

$$\begin{aligned} D_k(\omega_1, \omega_2) &= 1 + \mathbf{p}_1^t(\omega_1, \omega_2)\mathbf{r}^{(k)} \\ N_k(\omega_1, \omega_2) &= \mathbf{p}_2^t(\omega_1, \omega_2)\mathbf{s}^{(k)} \\ W_{2k}(\omega_1, \omega_2) &= \frac{W_2(\omega_1, \omega_2)}{|D_{k-1}(\omega_1, \omega_2)|^2}. \end{aligned} \tag{37}$$

At the kth iteration, $W_{2k}(\omega_1, \omega_2)$ given by (37) is known and non-negative, provided that $D_{k-1}(z_1, z_2)$ is stable in the 2-D sense. Thus, minimizing (36) with respect to parameters $\mathbf{r}^{(k)}$ and $\mathbf{s}^{(k)}$ is a least-squares problem. As will be shown in Section III-B, the requirement that $D_k(z_1, z_2)$ is stable can be satisfied by imposing a set of linear constraints on $\mathbf{r}^{(k)}$. In what follows, we first derive an explicit expression for J_{2k} in terms of $\mathbf{r}^{(k)}$ and $\mathbf{s}^{(k)}$.

A. An Explicit Expression for $J_{2k}[\mathbf{r}^{(k)}, \mathbf{s}^{(k)}]$

With the above matrix notation, we compute

$$\begin{aligned} & |F_d(\omega_1, \omega_2)D_k(\omega_1, \omega_2) - N_k(\omega_1, \omega_2)|^2 \\ & \quad = |F_d(\omega_1, \omega_2)|^2|D_k(\omega_1, \omega_2)|^2 + |N_k(\omega_1, \omega_2)|^2 \\ & \qquad - 2\mathrm{Re}[F_d(\omega_1, \omega_2)D_k(\omega_1, \omega_2)\overline{N}_k(\omega_1, \omega_2)] \end{aligned}$$

where

$$\begin{aligned} |D_k(\omega_1, \omega_2)|^2 &= 1 + 2\tilde{\mathbf{p}}_1^t(\omega_1, \omega_2)\mathbf{r}^{(k)} + \mathbf{r}^{(k)^t}\mathbf{P}_{11}(\omega_1, \omega_2)\mathbf{r}^{(k)} \\ |N_k(\omega_1, \omega_2)|^2 &= \mathbf{s}^{(k)^t}\mathbf{P}_{22}(\omega_1, \omega_2)\mathbf{s}^{(k)} \end{aligned} \tag{38}$$

with

$$\begin{aligned} \tilde{\mathbf{p}}_1(\omega_1, \omega_2) = [&\cos(\omega_1) \cdots \cos(n\omega_1) \quad \cos(\omega_2) \\ &\cos(\omega_1 + \omega_2) \cdots \cos(n_1\omega_1 + \omega_2) \\ &\cdots \cos(n_1\omega_1 + n_2\omega_2)]^t \\ \mathbf{P}_{11}(\omega_1, \omega_2) = &\text{the symmetric Toeplitz matrix} \\ &\text{determined by its first column} \\ &[1 \quad \cos(\omega_1) \cdots \cos[(n_1 - 1)\omega_1] \\ &\cos(\omega_2 - \omega_1) \quad \cos(\omega_2) \\ &\cdots \cos[\omega_2 + (n_1 - 1)\omega_1] \\ &\cdots \cos[n_2\omega_2 + (n_1 - 1)\omega_1]]^t \\ \mathbf{P}_{22}(\omega_1, \omega_2) = &\text{the symmetric Toeplitz matrix} \\ &\text{determined by its first column} \\ &[1 \quad \cos(\omega_1) \cdots \cos(n_1\omega_1) \quad \cos(\omega_2) \\ &\cos(\omega_1 + \omega_2) \cdots \cos(n_1\omega_1 + \omega_2) \\ &\cdots \cos(n_1\omega_1 + n_2\omega_2)]^t \end{aligned}$$

and

$$\begin{aligned} & \mathrm{Re}[F_d(\omega_1, \omega_2)D_k(\omega_1, \omega_2)\overline{N}_k(\omega_1, \omega_2)] = \\ & \quad \tilde{\mathbf{p}}^t(\omega_1, \omega_2)\mathbf{s}^{(k)} + \mathbf{r}^{(k)^t}\mathbf{P}_{12}(\omega_1, \omega_2)\mathbf{s}^{(k)} \end{aligned} \tag{39}$$

with

$$\begin{aligned} \tilde{\mathbf{p}}_2(\omega_1, \omega_2) &= \tfrac{1}{2}[F_d(\omega_1, \omega_2)\overline{\mathbf{p}}_2(\omega_1, \omega_2) \\ &\qquad + \overline{F}_d(\omega_1, \omega_2)\mathbf{p}_2(\omega_1, \omega_2)] \\ \mathbf{P}_{12}(\omega_1, \omega_2) &= \tfrac{1}{2}[F_d(\omega_1, \omega_2)\mathbf{p}_1(\omega_1, \omega_2)\overline{\mathbf{p}}_2^t(\omega_1, \omega_2) \\ &\qquad + \overline{F}_d(\omega_1, \omega_2)\overline{\mathbf{p}}_1(\omega_1, \omega_2)\mathbf{p}_2^t(\omega_1, \omega_2)]. \end{aligned}$$

A discretization of the integral in (36) then gives

$$J_{2k}[\mathbf{r}^{(k)}, \mathbf{s}^{(k)}] \approx \tfrac{1}{2}\mathbf{x}^{(k)^t}\mathbf{G}\mathbf{x}^{(k)} + \mathbf{b}^t\mathbf{x}^{(k)} + c \tag{40}$$

where

$$\begin{aligned}
\mathbf{G} &= \begin{bmatrix} \mathbf{G}_{11} & -\mathbf{G}_{12} \\ -\mathbf{G}_{12}^t & \mathbf{G}_{22} \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ -\mathbf{b}_2 \end{bmatrix} \\
\mathbf{G}_{11} &= \Delta^2 \sum_i \sum_j W_{2k}(\omega_{1i}, \omega_{2j}) \\
&\quad \cdot |F_d(\omega_{1i}, \omega_{2j})|^2 \mathbf{P}_{11}(\omega_{1i}, \omega_{2j}) \\
\mathbf{G}_{12} &= \Delta^2 \sum_i \sum_j W_{2k}(\omega_{1i}, \omega_{2j}) \mathbf{P}_{12}(\omega_{1i}, \omega_{2j}) \\
\mathbf{G}_{22} &= \Delta^2 \sum_i \sum_j W_{2k}(\omega_{1i}, \omega_{2j}) \mathbf{P}_{22}(\omega_{1i}, \omega_{2j}) \\
\mathbf{b}_1 &= \Delta^2 \sum_i \sum_j W_{2k}(\omega_{1i}, \omega_{2j}) |F_d(\omega_{1i}, \omega_{2j})|^2 \\
&\quad \cdot \tilde{\mathbf{p}}_1(\omega_{1i}, \omega_{2j}) \\
\mathbf{b}_2 &= \Delta^2 \sum_i \sum_j W_{2k}(\omega_{1i}, \omega_{2j}) \tilde{\mathbf{p}}_2(\omega_{1i}, \omega_{2j}). \qquad (41)
\end{aligned}$$

c is a constant independent of $\mathbf{r}^{(k)}$ and $\mathbf{s}^{(k)}$, and $\{\omega_{1i}, \omega_{2j}\}$ is a dense grid points over the region Ω_2 with $1 \le i \le L_1$, $1 \le j \le L_2$, $L_1 \ge 4n_1$, and $L_2 \ge 4n_2$.

B. The Stability Issue

As was mentioned earlier, the stability constraint on $D_k(z_1, z_2)$ has to be imposed. It is well known [13] that $D_k(z_1, z_2)$ is a stable polynomial if and only if i) polynomial $D_k(z_1, z_2)$ with $z_2^{-1} \equiv 0$ is stable, and ii) for each fixed z_1^{-1} on the unit circle, $D(z_1, z_2)$ is stable. This stability condition, in conjunction with the 1-D sufficient stability condition that has been used in Section II-C, yields the following sufficient condition for the stability of $D_k(z_1, z_2)$: Polynomial $D_k(z_1, z_2)$ is 2-D stable if

$$\mathrm{Re}[\hat{D}_k(e^{j\omega_1})] > 0 \qquad \text{for } \omega_1 \in [0, \pi] \qquad (42a)$$

$$\mathrm{Re}[D_k(e^{j\omega_1}, e^{j\omega_2})] > 0 \qquad \text{for } \omega_1, \omega_2 \in [0, \pi] \qquad (42b)$$

where $\hat{D}_k(z_1)$ is the 1-D polynomial defined by

$$\hat{D}_k(z_1) = D_k(z_1, z_2)|_{z_2^{-1}=0}. \qquad (43)$$

From (33) and (42), it follows that $D_k(z_1, z_2)$ is stable if

$$-\hat{\mathbf{p}}_1^t(\omega_1)\mathbf{r}_1^{(k)} \le 1 - \delta \qquad \omega_1 \in [0, \pi] \qquad (44a)$$

$$-\tilde{\mathbf{p}}_1^t(\omega_1, \omega_2)\mathbf{r}^{(k)} \le 1 - \delta \qquad \omega_1, \omega_2 \in [0, \pi] \qquad (44b)$$

where $\mathbf{r}_1^{(k)}$ is the vector consisting of the first n_1 components of $\mathbf{r}^{(k)}$, $\delta > 0$ is a small constant, and $\hat{\mathbf{p}}_1(\omega_1) = [\cos(\omega_1) \cdots \cos(n_1\omega_1)]^t$, $\tilde{\mathbf{p}}_1(\omega_1)$ was defined in Section III-A. Implementing constraints (44) on a set of dense grid points over Ω_2, say, $\Theta_2 = \{(\omega_{1i}, \omega_{2j}), 1 \le i \le I, 1 \le j \le J\}$, we obtain the following linear constraints for the stability of $D_k(z_1, z_2)$:

$$\mathbf{R}\mathbf{x}^{(k)} \le (1 - \delta)\mathbf{e}_K \qquad (45)$$

where

$$\mathbf{R} = -\begin{bmatrix} -\hat{\mathbf{p}}_1^t(\omega_{1,1}) & \\ \vdots & \\ -\hat{\mathbf{p}}_1^t(\omega_{1,I}) & \\ -\tilde{\mathbf{p}}_1(\omega_{1,1}, \omega_{2,1}) & 0 \\ \vdots & \\ -\tilde{\mathbf{p}}_1(\omega_{1,I}, \omega_{2,J}) & \end{bmatrix} \quad \mathbf{e}_K = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{K \times 1} \qquad (46)$$

with $K = I + I \times J$.

C. The Constrained Least-Squares Design

The design begins by choosing a stable initial $\mathbf{r}^{(0)}$ such as $\mathbf{r}^{(0)} = [0 \cdots 0]^t$. Matrices $\mathbf{G}$ and $\mathbf{b}$ in (40) and $\mathbf{R}$ in (45) are then evaluated, and the quadratic programming problem in (40) and (45) is solved to obtain $\mathbf{x}^{(1)}$. In the next iteration, matrices $\mathbf{G}$ and $\mathbf{b}$ in (40) are updated using $\mathbf{x}^{(1)}$, and then, the same quadratic programming problem in (40) and (45) is solved to obtain $\mathbf{x}^{(2)}$. The iteration continues until $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| < \varepsilon$, which is a prescribed tolerance.

D. The Design of Quadrantally Symmetric 2-D Filters

It is noted that even for a moderate density of grid points on Ω_2, the number of stability constraints in (45) can easily exceed 2000, this in conjunction with a large size matrix $\mathbf{G}$ in (40) for a filter of moderate order, leads to a large size quadratic programming problem. However, the computational burden can be considerably reduced by restricting the filter being designed to the class of quadrantally symmetric filters. It is well known that all circularly symmetric filters, various regularization filters [14], fan, and diamond-shaped filters possess a quadrantally symmetric frequency response. It is also known [15] that the transfer function of a quadrantally symmetric 2-D filter has a separable denominator, i.e.,

$$F(z_1, z_2) = \frac{N(z_1, z_2)}{g(z_1)h(z_2)}. \qquad (47)$$

Obviously, the design idea addressed in the preceding subsections is applicable to $F(z_1, z_2)$ in (47) with reduced computational complexity.

Let

$$\begin{aligned}
g(z_1) &= 1 + \mathbf{q}_1^t(z_1)\mathbf{g} \\
h(z_2) &= 1 + \mathbf{q}_2^t(z_2)\mathbf{h} \\
\mathbf{g} &= [g_1 \cdots g_{n_1}]^t \\
\mathbf{h} &= [h_1 \cdots h_{n_2}]^t \\
\mathbf{q}_1(z_1) &= [z_1^{-1} \cdots z_1^{-n_1}]^t \\
\mathbf{q}_2(z_2) &= [z_2^{-1} \cdots z_2^{-n_2}]^t.
\end{aligned}$$

The cost function J_{2k} in (36) in this case becomes

$$\begin{aligned}
J_{2k}[\mathbf{g}^{(k)}, \mathbf{h}^{(k)}, \mathbf{s}^{(k)}] = \tfrac{1}{2} \int\int_{\Omega_2} & W_{2k}(\omega_1, \omega_2) \\
& \cdot |F_d(\omega_1, \omega_2) g_k(\omega_1) h_k(\omega_2) \\
& - N_k(\omega_1, \omega_2)|^2 \, d\omega_1 \, d\omega_2 \qquad (48)
\end{aligned}$$

with $k = 1, 2, \cdots$, $\mathbf{g}^{(0)} = [0 \cdots 0]^t$, $\mathbf{h}^{(0)} = [0 \cdots 0]^t$

$$\begin{aligned} g_k(\omega_1) &= 1 + \mathbf{q}_1(\omega_1)\mathbf{g}^{(k)} \\ h_k(\omega_2) &= 1 + \mathbf{q}_2(\omega_2)\mathbf{h}^{(k)} \\ N_k(\omega_1, \omega_2) &= \mathbf{p}_2^t(\omega_1, \omega_2)\mathbf{s}^{(k)} \\ W_{2k}(\omega_1, \omega_2) &= \frac{W_2(\omega_1, \omega_2)}{|g_{k-1}(\omega_1)h_{k-1}(\omega_2)|^2}. \end{aligned}$$

At the kth iteration, one seeks to find parameter vectors $\mathbf{g}^{(k)}$, $\mathbf{h}^{(k)}$, and $\mathbf{s}^{(k)}$ that minimize (48). At this point, we must stress that unlike the general case (36), minimizing (48) is no longer a least-squares problem since the parameters in $\mathbf{g}^{(k)}$ and $\mathbf{h}^{(k)}$ are multiplied by each other before the square. There are several approaches to modify J_{2k} so that it can be reformulated as a least-squares problem. For example, the modification of J_{2k} given by

$$\begin{aligned} \hat{J}_{2k} &= \tfrac{1}{2} \int\int_{\Omega_2} W_{2k}(\omega_1, \omega_2)|F_d(\omega_1, \omega_2)\hat{D}_k(\omega_1, \omega_2) \\ &\quad - N_k(\omega_1, \omega_2)|^2 \, d\omega_1 \, d\omega_2 \\ \hat{D}_k(\omega_1, \omega_2) &= \begin{cases} g_k(\omega_1)h_{k-1}(\omega_2) & \text{for even } k \\ g_{k-1}(\omega_1)h_k(\omega_2) & \text{for odd } k \end{cases} \end{aligned} \tag{49}$$

leads to a least-squares problem. For an even k, parameters $\mathbf{g}^{(k)}$ and $\mathbf{s}^{(k)}$ are sought to minimize $\hat{J}_{2k}$, whereas for an odd k, parameters $\mathbf{h}^{(k)}$ and $\mathbf{s}^{(k)}$ are sought to minimize $\hat{J}_{2k}$. From (49), it follows that if $\mathbf{g}^{(k)} \to \mathbf{g}$, $\mathbf{h}^{(k)} \to \mathbf{h}$, and $\mathbf{s}^{(k)} \to \mathbf{s}$ under certain stability constraints (which will be detailed shortly) as $k \to \infty$, then the 2-D transfer function associated with the limiting vectors $\mathbf{g}$, $\mathbf{h}$, and $\mathbf{s}$ minimizes

$$\begin{aligned} \hat{J}_2 &= \tfrac{1}{2} \int\int_{\Omega_2} W_2(\omega_1, \omega_2) \left| F_d(\omega_1, \omega_2) \right. \\ &\quad \left. - \frac{N(\omega_1, \omega_2)}{g(\omega_1)h(\omega_2)} \right|^2 d\omega_1 \, d\omega_2 \end{aligned} \tag{50}$$

and, therefore, offers a weighted least-squares solution to the design problem.

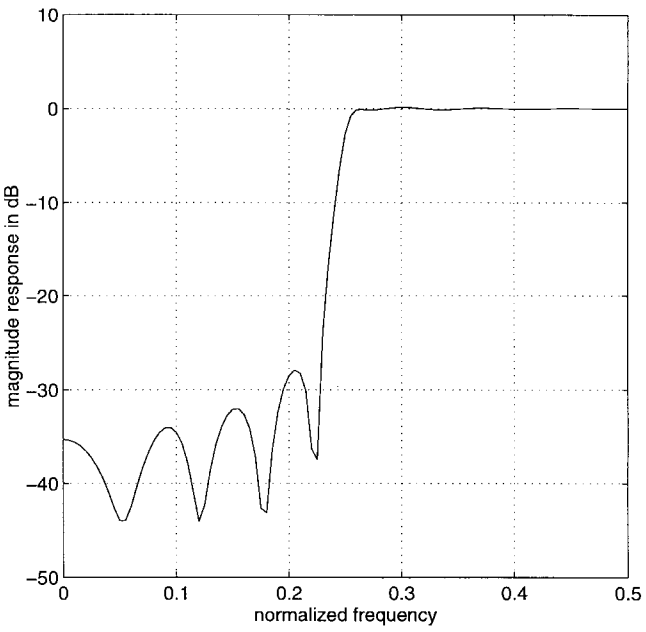
Given $\hat{D}_k(\omega_1, \omega_2)$ by (49), explicit expressions of $\hat{J}_{2k}$ for even and odd k's can be derived in a manner similar to that in Section III-A. These expressions are given below without deriving the details.

For even k

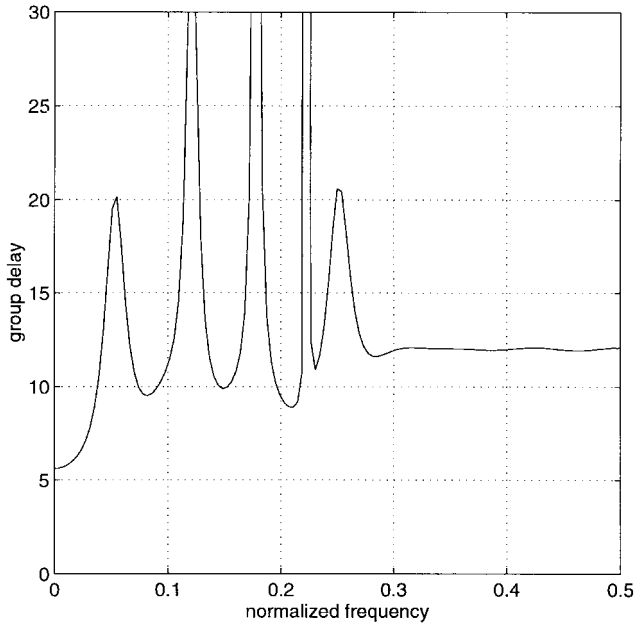
$$\hat{J}_{2k}[\mathbf{g}^{(k)}, \mathbf{s}^{(k)}] \approx \tfrac{1}{2}\mathbf{x}_e^{(k)^t}\mathbf{G}_e\mathbf{x}_e^{(k)} + \mathbf{b}_e^t\mathbf{x}_e^{(k)} + c_e \tag{51}$$

where

$$\begin{aligned} \mathbf{G}_e &= \begin{bmatrix} \mathbf{G}_{11e} & -\mathbf{G}_{12e} \\ -\mathbf{G}_{12e}^t & \mathbf{G}_{22e} \end{bmatrix} \quad \mathbf{b}_e = \begin{bmatrix} \mathbf{b}_{1e} \\ -\mathbf{b}_{2e} \end{bmatrix} \\ \mathbf{x}_e^{(k)} &= \begin{bmatrix} \mathbf{g}^{(k)} \\ \mathbf{s}^{(k)} \end{bmatrix} \end{aligned} \tag{52}$$



(a)



(b)

Fig. 1. (a) Magnitude response of the filter in Example 1 and (b) group delay of the filter in Example 1.

and

$$\begin{aligned} \mathbf{G}_{11e} &= \Delta^2 \sum_i \sum_j W_{2k}(\omega_{1i}, \omega_{2j})|F_d(\omega_{1i}, \omega_{2j})|^2 \\ &\quad \cdot |h_{k-1}(\omega_{2j})|^2 \mathbf{P}_{11e}(\omega_{1i}) \\ \mathbf{G}_{12e} &= \Delta^2 \sum_i \sum_j W_{2k}(\omega_{1i}, \omega_{2j})\mathbf{P}_{12e}(\omega_{1i}, \omega_{2j}) \\ \mathbf{G}_{22e} &= \mathbf{G}_{22} \text{ in } (41) \\ \mathbf{b}_{1e} &= \Delta^2 \sum_i \sum_j W_{2k}(\omega_{1i}, \omega_{2j})|F_d(\omega_{1i}, \omega_{2j})|^2 \\ &\quad \cdot |h_{k-1}(\omega_{2j})|^2 \hat{\mathbf{p}}_1(\omega_{1i}) \\ \mathbf{b}_{2e} &= \Delta^2 \sum_i \sum_j W_{2k}(\omega_{1i}, \omega_{2j})\tilde{\mathbf{p}}_{2e}(\omega_{1i}, \omega_{2j}) \end{aligned} \tag{53}$$

TABLE I
COEFFICIENTS OF THE TRANSFER FUNCTION IN EXAMPLE 1

Denominator	Numerator
1.0	0.00216452
0.85183505	0.01422263
1.40277206	0.01070862
1.15151797	-0.00173896
0.95948743	-0.00056474
0.81499579	0.01648599
0.68685173	0.01450372
0.52355787	-0.01424925
0.35584783	-0.01101010
0.23561367	0.04991283
0.15595331	0.04637753
0.08260605	-0.19765241
0.01749834	0.33139538
-0.01347975	-0.25569216
-0.01109415	0.13691517

with

$$\begin{aligned}
\hat{\mathbf{p}}_1(\omega_1) &= [\cos(\omega_1) \cdots \cos(n_1\omega_1)]^t \\
\tilde{\mathbf{p}}_{2e}(\omega_1, \omega_2) &= \tfrac{1}{2}[F_d(\omega_1, \omega_2)h_{k-1}(\omega_2)\overline{\mathbf{p}}_2(\omega_1, \omega_2) \\
&\quad + \overline{F}_d(\omega_1, \omega_2)\overline{h}_{k-1}(\omega_2)\mathbf{p}_2(\omega_1, \omega_2)] \\
\mathbf{P}_{12e}(\omega_1, \omega_2) &= \tfrac{1}{2}[F_d(\omega_1, \omega_2)h_{k-1}(\omega_2)\mathbf{p}_{1e}(\omega_1)\overline{\mathbf{p}}_2^t(\omega_1, \omega_2) \\
&\quad + \overline{F}_d(\omega_1, \omega_2)\overline{h}_{k-1}(\omega_2)\overline{\mathbf{p}}_{1e}(\omega_1)\mathbf{p}_2^t(\omega_1, \omega_2)] \\
\mathbf{p}_{1e}(\omega_1) &= [e^{-j\omega_1} \cdots e^{-jn_1\omega_1}]^t. \qquad (54)
\end{aligned}$$

where $\mathbf{P}_{11e}(\omega_1)$ is the symmetric Toeplitz matrix determined by its first column $[1 \quad \cos(\omega_1) \cdots \cos[(n_1-1)\omega_1]]^t$. For odd k

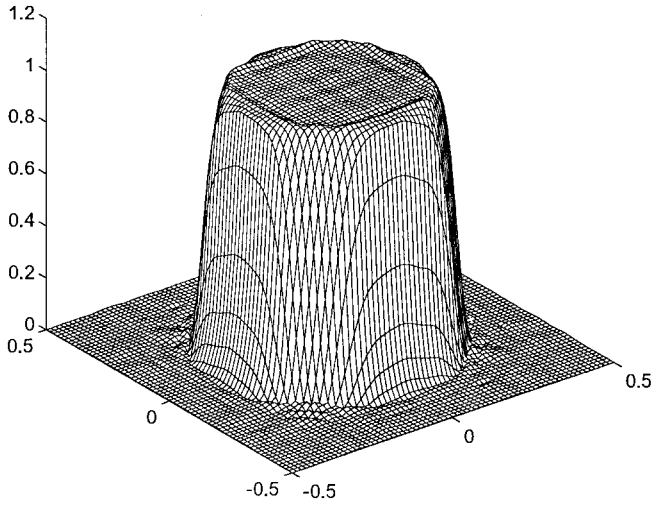
$$\hat{J}_{2k}[\mathbf{h}^{(k)}, \mathbf{s}^{(k)}] \approx \tfrac{1}{2}\mathbf{x}_o^{(k)^t}\mathbf{G}_o\mathbf{x}_o^{(k)} + \mathbf{b}_o^t\mathbf{x}_o^{(k)} + c_o \qquad (55)$$

where

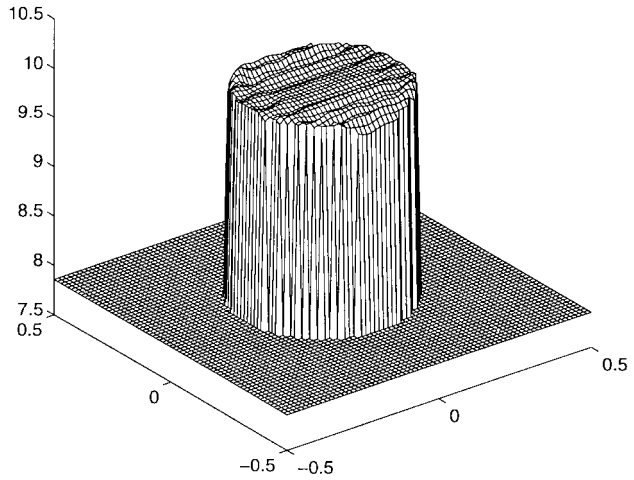
$$\begin{aligned}
\mathbf{G}_o &= \begin{bmatrix} \mathbf{G}_{11o} & -\mathbf{G}_{12o} \\ -\mathbf{G}_{12o}^t & \mathbf{G}_{22o} \end{bmatrix} \quad \mathbf{b}_o = \begin{bmatrix} \mathbf{b}_{1o} \\ -\mathbf{b}_{2o} \end{bmatrix} \\
\mathbf{x}_o^{(k)} &= \begin{bmatrix} \mathbf{h}^{(k)} \\ \mathbf{s}^{(k)} \end{bmatrix} \qquad (56)
\end{aligned}$$

and

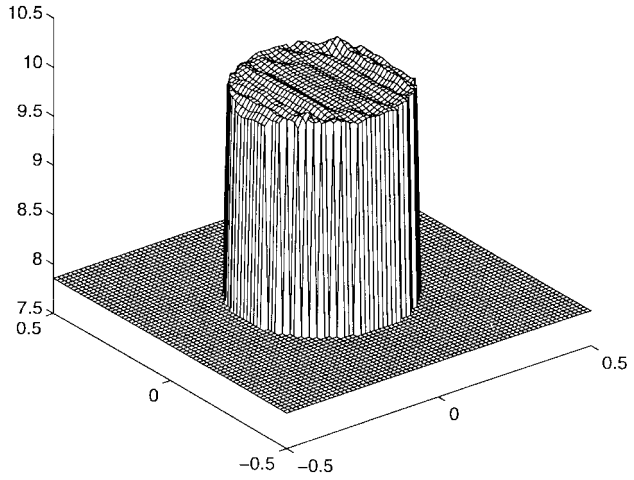
$$\begin{aligned}
\mathbf{G}_{11o} &= \Delta^2 \sum_i \sum_j W_{2k}(\omega_{1i}, \omega_{2j})|F_d(\omega_{1i}, \omega_{2j})|^2 \\
&\quad \cdot |g_{k-1}(\omega_{2i})|^2\mathbf{P}_{11o}(\omega_{2j}) \\
\mathbf{G}_{12o} &= \Delta^2 \sum_i \sum_j W_{2k}(\omega_{1i}, \omega_{2j})\mathbf{P}_{12o}(\omega_{1i}, \omega_{2j}) \\
\mathbf{G}_{22o} &= \mathbf{G}_{22} \text{ in } (41) \\
\mathbf{b}_{1o} &= \Delta^2 \sum_i \sum_j W_{2k}(\omega_{1i}, \omega_{2j})|F_d(\omega_{1i}, \omega_{2j})|^2 \\
&\quad \cdot |g_{k-1}(\omega_{1i})|^2\hat{\mathbf{p}}_2(\omega_{2j}) \\
\mathbf{b}_{2o} &= \Delta^2 \sum_i \sum_j W_{2k}(\omega_{1i}, \omega_{2j})\tilde{\mathbf{p}}_{2o}(\omega_{1i}, \omega_{2j}) \qquad (57)
\end{aligned}$$



(a)



(b)



(c)

Fig. 2. (a) Magnitude response of the filter in Example 2. (b) Group delay (along $\omega 1$) in the passband of the filter in Example 2. (c) Group delay (along $\omega 2$) in the passband of the filter in Example 2.

with

$$\begin{aligned}\hat{\mathbf{p}}_2(\omega_2) &= [\cos(\omega_2) \cdots \cos(n_2\omega_2)]^t \\ \tilde{\mathbf{p}}_{2o}(\omega_1, \omega_2) &= \tfrac{1}{2}[F_d(\omega_1, \omega_2)g_{k-1}(\omega_1)\overline{\mathbf{p}}_2(\omega_1, \omega_2) \\ &\quad + \overline{F}_d(\omega_1, \omega_2)\overline{g}_{k-1}(\omega_1)\mathbf{p}_2(\omega_1, \omega_2)] \\ \mathbf{P}_{12o}(\omega_1, \omega_2) &= \tfrac{1}{2}[F_d(\omega_1, \omega_2)g_{k-1}(\omega_1)\mathbf{p}_{2o}(\omega_2)\overline{\mathbf{p}}_2^t(\omega_1, \omega_2) \\ &\quad + \overline{F}_d(\omega_1, \omega_2)\overline{g}_{k-1}(\omega_1)\overline{\mathbf{p}}_{2o}(\omega_2)\mathbf{p}_2^t(\omega_1, \omega_2)] \\ \mathbf{p}_{2o}(\omega_2) &= [e^{-j\omega_2} \cdots e^{-jn_2\omega_2}]^t. \end{aligned} \tag{58}$$

where $\mathbf{P}_{11o}(\omega_1)$ is the symmetric Toeplitz matrix determined by its first column $[1 \quad \cos(\omega_2) \cdots \cos[(n_2-1)\omega_2]]^t$. Note that in either case, the number of parameters involved in $\hat{J}_{2k}$ has been reduced from the general case of $2(n_1+1)(n_2+1)-1$ to $(n_1+1)(n_2+1)+n_1$ (for even k) or $(n_1+1)(n_2+1)+n_2$ (for odd k). Since the filter is quadrantally symmetric, the region Ω_2 can be reduced to $\Omega_2 = [0, \pi) \times [0, \pi)$. Moreover, the stability constraints (45) are in this case replaced by

$$\mathbf{R}_e\mathbf{x}^{(k)} \leq (1-\delta)\mathbf{e}_{K1} \tag{59}$$

for even k, where $\mathbf{x}^{(k)} = [\mathbf{p}^{(k)^t} \ \mathbf{s}^{(k)^t}]^t$, and

$$\mathbf{R}_e = \begin{bmatrix} -\hat{\mathbf{p}}_1^t(\omega_{1,1}) & \\ \vdots & 0 \\ -\hat{\mathbf{p}}_1^t(\omega_{1,I}) & \end{bmatrix} \quad \mathbf{e}_{K1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{I\times 1} \tag{60}$$

and by

$$\mathbf{R}_o\mathbf{x}^{(k)} \leq (1-\delta)\mathbf{e}_{K2} \tag{61}$$

for odd k, where $\mathbf{x}^{(k)} = [\mathbf{h}^{(k)^t} \ \mathbf{s}^{(k)^t}]^t$, and

$$\mathbf{R}_o = \begin{bmatrix} -\hat{\mathbf{p}}_2^t(\omega_{2,1}) & \\ \vdots & 0 \\ -\hat{\mathbf{p}}_2^t(\omega_{2,J}) & \end{bmatrix} \quad \mathbf{e}_{K2} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{J\times 1}. \tag{62}$$

It follows that the number of constraints is reduced from $I + I \times J$ to I for even k and to J for odd k.

IV. Design Examples

The design method proposed in this paper is fairly general in the sense that

1) it can be used to find a 1-D or 2-D transfer function that approximates arbitrary amplitude and phase responses;
2) the design obtained is optimal in the least-squares sense;
3) the stability of the filter obtained is guaranteed;
4) by using appropriate weighting, a quasi-equiripple design can be achieved.

In this section, two examples are presented to illustrate this design methodology.

Example 1: A halfband highpass IIR filter of order 14 with linear phase response over its passband is designed using the iterative constrained least-squares method proposed in Section II. The desired frequency is given by

$$F_d(\omega) = \begin{cases} 1e^{-j12\omega} & \omega \geq 0.525\pi \\ 0 & \omega \leq 0.475\pi \end{cases}. \tag{63}$$

As was observed in [11], good designs are usually achieved when the group delay is set to be between $n-1$ and $n/2$, where n denotes the order of the filter. The desired group delay in (63) is obviously in agreement with this observation. The weighting function $W(\omega)$ is chosen as 1 in the passband and stopband. With a trivial initial $\mathbf{d}^{(0)} = [0 \cdots 0]^t$ and tolerance $\varepsilon = 10^{-4}$, the algorithm converges after six iterations. The amplitude response and group delay are shown in Fig. 1, and the coefficients of the filter designed are given in Table I. The maximum modulus of the filter poles is 0.9276; hence, the filter is stable. The peak error in the passband is 0.1406 dB, and the peak error in the stopband is -27.8974 dB.

Example 2: In this example, a 2-D circularly symmetric lowpass IIR filter of order $(n_1, n_2) = (14, 14)$ with linear phase over its passband is designed using the least-squares method developed in Section III-D. The desired response is given by

$$F_d(\omega_1, \omega_2) = \begin{cases} 1e^{-j10(\omega_1+\omega_2)} & \sqrt{\omega_1^2+\omega_2^2} \leq 0.5\pi \\ 0 & \sqrt{\omega_1^2+\omega_2^2} \geq 0.7\pi \end{cases}. \tag{64}$$

Again, the desired group delay in (64) is in the range $[n_i/2, n_i - 1]$ $(i = 1, 2)$ as was suggested in [11]. The weighting function $W(\omega_1, \omega_2)$ is chosen as 5 in the passband and 1 in the stopband. With trivial initial $\mathbf{g}^{(0)} = [0 \cdots 0]^t$, $\mathbf{h}^{(0)} = [0 \cdots 0]^t$ and tolerance $\varepsilon = 5 \times 10^{-3}$, the design algorithm converges after 12 iterations. The maximum pole radius of the designed filter is 0.9236; hence, the filter is stable. The resultant magnitude response is shown in Fig. 2(a). The maximum ripple in the passband is 0.0118, and the maximum ripple in the stopband is 0.0268. Moreover, let the phase response of the designed filter be $p(\omega_1, \omega_2)$. The group delays of the filter corresponding to ω_1- and ω_2-axii are defined by

$$\begin{aligned}\tau_1(\omega_1, \omega_2) &= -\frac{\partial}{\partial\omega_1}p(\omega_1, \omega_2) \\ \tau_2(\omega_1, \omega_2) &= -\frac{\partial}{\partial\omega_2}p(\omega_1, \omega_2).\end{aligned}$$

Fig. 2(b) and (c) shows the 3-D plots of group delay $\tau_1(\omega_1, \omega_2)$ and $\tau_2(\omega_1, \omega_2)$ in the passband of the filter designed. It is observed that the peak-to-peak deviation in the passband for both τ_1 and τ_2 is 0.2564, representing a 2.564% group delay distortion.

V. Conclusion

We have proposed a new approach to the weighted least-squares design of stable IIR 1-D and 2-D digital filters. The difficulty of taking the denominator of the transfer function being designed into account in a least-squares setting is overcome by means of an iterative strategy that treats the denominator as a part of the weighting function, and the

stability of the filter designed is guaranteed by a set of linear constraints on the denominator coefficients. The usefulness of the proposed method has been demonstrated by two design examples.

ACKNOWLEDGMENT

The authors are grateful to the reviewers for their constructive comments, which have led to this improved version of the paper.

REFERENCES

[1] D. C. Farden and L. L. Scharf, "Statistical design of nonrecursive digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 188–196, June 1974.
[2] V. R. Algazi and M. Suk, "On the frequency weighted least squares design of finite duration filter," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 943–953, Dec. 1975.
[3] V. R. Algazi, M. Suk, and C. S. Rim,"Design of almost minimax FIR filters in one and two dimensions by WLS technique," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 590–596, June 1986.
[4] M. O. Ahmad and J. D. Wang, "An analytical least squares solution to the design problem of two-dimensional FIR filters with quadrantally symmetric or antisymmetric frequency response," *IEEE Trans. Circuits Syst.*, vol. 36, pp. 968–979, July 1989.
[5] T. Kobayashi and S. Imai, "Design of IIR digital filters with arbitrary log magnitude function by WLS technique," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 247–252, Feb. 1990.
[6] Y. C. Lim, J. H. Lee, C. K. Chen, and R. H. Yang, "A weighted least squares algorithm for quasi-equiripple FIR and IIR digital filter design," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 40, pp. 551–558, Mar. 1992.
[7] S. C. Pei and J. J. Shyu, "Design of arbitrary FIR log filters by weighted least squares technique," *IEEE Trans. Signal Processing*, vol. 42, pp. 2495–2499, Sept. 1994.
[8] S. Sunder and V. Ramachandran, "Design of recursive differentiators with constant group delay characteristics," *Signal Process.*, vol. 39, pp. 79–88, Sept. 1994.
[9] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems.* Englewood Cliffs, NJ: Prentice-Hall, 1974.
[10] E. A. Robinson, *Statistical Communication and Detection.* New York: Hafner, 1967.
[11] A. T. Chottera and G. A. Jullien, "A linear programming approach to recursive digital filter design with linear phase," *IEEE Trans. Circuits Syst.*, vol. CAS-29, pp. 139–149, Mar. 1982.
[12] E. I. Jury, *Theory and Applications of the z-Transform Method.* New York: Wiley, 1964.
[13] W.-S. Lu and A. Antoniou, *Two-Dimensional Digital Filters.* New York: Marcel Dekker, 1992.
[14] C. M. Leung and W.-S. Lu, "Image restoration by 2-D recursive regularization filters," in *Proc. Canadian Conf. Elect. Comput. Eng.*, Toronto, Ont., Canada, Sept. 1992, pp. WM3.24.1–WM3.24.4.
[15] P. K. Rajan and M. N. S. Swamy, "Quadrantal symmetry associated with two-dimensional digital transfer functions," *IEEE Trans. Circuits Syst.*, vol. CAS-29, pp. 340–343, June 1983.
[16] I. W. Selesnick, M. Lang, and C. S. Burrus, "Constrained least squares design of FIR filters without specified transition bands," in *Proc. ICASSP*, 1995, pp. 1260–1263.
[17] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed. Reading, MA: Addison-Wesley, 1984.
[18] C. K. Chen and J. H. Lee, "Design of quadrature mirror filters with linear phase in the frequency domain," *IEEE Trans. Circuits Syst.*, vol. 39, pp. 593–605, Sept. 1992.
[19] H. Xu, W.-S. Lu, and A. Antoniou, "Improved iterative methods for the design of quadrature mirror-image filter bands," *IEEE Trans. Circuits Syst. II*, vol. 43, pp. 363–371, May 1996.

Wu-Sheng Lu received the B.S. degree in mathematics from Fudan University, Fudan, China, in 1964 and the M.S. degree in electrical engineering and the Ph.D. degree in control science from the University of Minnesota, Minneapolis, in 1983 and 1984, respectively.

He was a Postdoctoral Fellow at the University of Victoria, Victoria, B.C., Canada, in 1985, and held a Visiting Assistant Professorship at the University of Minnesota in 1986. Since 1987, he has been with the University of Victoria, where he is currently a Professor. He is the co-author, with A. Antoniou, of *Two-Dimensional Digital Filters* (New York: Marcel Dekker, 1992).

Dr. Lu served as Associate Editor of the *Canadian Journal of Electrical and Computer Engineering* in 1989, Editor of the same journal from 1990 to 1992, and Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II from 1993 to 1995. He is an Associate Editor of *Multidimensional Systems and Signal Processing*.

Soo-Chang Pei (SM'89) was born in Soo-Auo, Taiwan, in 1949. He received the B.S.E.E. degree from the National Taiwan University, Taipei, Taiwan, R.O.C., in 1970 and the M.S.E.E. and Ph.D. degrees from the University of California, Santa Barbara (UCSB), in 1972 and 1975, respectively.

He was an Engineering Officer with the Chinese Navy Shipyard from 1970 to 1971. From 1971 to 1975, he was a Research Assistant at UCSB. He was Professor and Chairman with the Department of Electrical Engineering, Tatung Institute of Technology, Taiwan, from 1981 to 1983. Presently, he is a Professor and Chairman of the Department of Electrical Engineering, National Taiwan University. His research interests include digital signal processing, image processing, optical information processing, and laser holography.

Dr. Pei is a member of Eta Keppa Nu and the Optical Society of Ameria.

Chien-Cheng Tseng (S'90–M'96) was born in Taipei, Taiwan, R.O.C., on August 25, 1965. He received the B.S. degree, with honors, from Tatung Institute of Technology, Taipei, in 1988, and the M.S. and Ph.D. degrees from the National Taiwan University, Taipei, in 1990 and 1995, respectively, all in electrical engineering.

He is currently an Associate Research Engineer at Telecommunication Laboratories, Chunghwa Telecom Co., Ltd., Taoyuan, Taiwan. His research interests include digital signal processing, pattern recognition, and electrical commerce.