Structure Simplification of Nonrecursive 2-D Digital Filters: A New L_∞ Error Bound

W.-S. Lu

***Abstract*— This brief presents an error analysis for the SVD-based structure simplification method. Specifically, we derive a new L_∞ error bound for the simplification strategy. The error bound obtained is very tight as compared to the existing bound. The significance of having such a bound is twofold. First, if the frequency response of the original function is satisfactory and the error bound obtained is small, then there will be no spikes for the frequency response of the simplified transfer function over the entire baseband. Second, a tight bound serves as a guide to determine how many parallel sections can be neglected from the filter network without exceeding a prescribed tolerance. The new error bound therefore offers an additional justification for the use of the SVD-based simplification technique. An example is included to illustrate the algorithm for computing the error bound and demonstrate the tightness of the bound.**

I. Introduction

Simplifying the structure of a digital filter is of practical importance as it leads to a less expensive implementation as well as improved processing speed. Here we take "structure simplification" to mean an effort to obtain an approximation of the filter's transfer function with a prescribed accuracy, leading to a digital-filter network that requires less circuit elements in the case of hardware implementation or less arithmetic operations in the case of software implementation. Typically, simplification of this kind can be carried out by reducing the order of the filter. Various techniques for the order reduction of two-dimensional (2-D) systems have been developed during the last two decades [1]–[9]. As will be demonstrated in this paper, however, the simplification can also be accomplished by reducing the number of sections, of which the filter network is composed. This is particularly the case when the transfer function at hand can be expressed in a form that corresponds to a parallel-cascade structure. Several results on this type of structure simplification problems have been available for nonrecursive 2-D filters [7], [10]. The objective of this paper is to give a new L_∞ error bound for the SVD-based section reduction method. The error bound obtained here is very tight as compared to that obtained in [7]. The significance of having an L_∞ bound is twofold. First, if the frequency response of the original transfer function is satisfactory and the L_∞ bound is small, then it is guaranteed that there will be no spikes for the frequency response of the simplified transfer function over the entire baseband. Second, from the implementation point of view, a tight L_∞ bound serves as a guide to determine how many parallel sections can be neglected from the filter network without exceeding a prescribed tolerance. In the next section, the new L_∞ error bound is derived and an algorithm for its computation is presented. We also show that the evaluation of the new error bound for K varying in a set of consecutive integers (where K represents the number of sections used) can be made recursive. As an illustrative example, a circularly symmetric low-pass FIR filter will be used in Section III to demonstrate how the algorithm can be incorporated into a design procedure to obtain a simplified yet satisfactory design.

Manuscript received October 11, 1994; revised April 25, 1995. This paper was recommended by Associate Editor I. Pitas.

The author is with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, B.C. V8W 3P6 Canada.

IEEE Log Number 9413911.

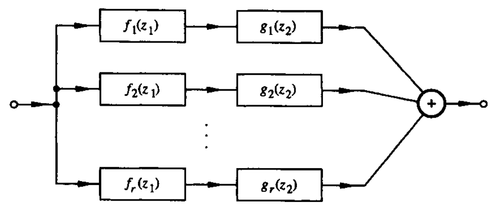

Fig. 1. The parallel structure of $H(z_1, z_2)$.

II. An L_∞ Error Bound for the SVD-Based Section Reduction Method

A. The L_∞ Bound

Consider a nonrecursive 2-D filter with a transfer function given by

$$H(z_1, z_2) = \sum_{n_1=0}^{N_1-1}\sum_{n_2=0}^{N_2-1} h_{n_1 n_2} z_1^{-n_1} z_2^{-n_2}. \tag{1}$$

Using matrix notation, the transfer function can be expressed as

$$H(z_1, z_2) = Z_1^T C Z_2 \tag{2}$$

where $Z_1 = [1 \quad z_1^{-1} \quad \cdots \quad z_1^{-(N_1-1)}]^T$ and $Z_2 = [1 z_2^{-1} \quad \cdots \quad z_2^{-(N_2-1)}]^T$, and

$$C = \begin{bmatrix} h_{00} & h_{01} & \cdots & h_{0,N_2-1} \\ \vdots & & & \vdots \\ h_{N_1-1,0} & h_{N_1-1,1} & \cdots & h_{N_1-1,N_2-1} \end{bmatrix}.$$

The section reduction under consideration is based on the singular value decomposition (SVD) of the coefficient matrix C, which is given by

$$C = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i \hat{u}_i \hat{v}_i^T = \sum_{i=1}^{r} u_i v_i^T \tag{3}$$

where $r = \operatorname{rank}(C)$; $\Sigma = \operatorname{diag}\{\sigma_1, \cdots, \sigma_r, 0 \cdots 0\}$ with $\sigma_1 \geq \cdots \geq \sigma_r > 0$; $\hat{u}_i$ and $\hat{v}_i$ are the ith column of U and V, respectively; and $u_i = \sigma_i^{1/2}\hat{u}_i$, $v_i = \sigma_i^{1/2}\hat{v}_i$. With (2) and (3) we write the transfer function as

$$H(z_1, z_2) = \sum_{i=1}^{r} (Z_1^T u_i)(v_i^T Z_2) \equiv \sum_{i=1}^{r} f_i(z_1)\, g_i(z_2) \tag{4}$$

that suggests a parallel structure for the realization of $H(z_1, z_2)$, where each section has two 1-D filters in cascade as shown in Fig. 1.

Note that if the 2-D filter is quadrantally symmetric, then the total number of sections in Fig. 1 does not exceed $\min((N_1-1)/2, (N_2-1)/2)$ (odd N_1 and N_2 are assumed here). As will be shown below, obtaining a simplified, more efficient network that has less sections than in Fig. 1 is often possible.

If we construct a coefficient matrix C_K to approximate C by neglecting the $r-K$ least significant singular values $\sigma_{K+1}, \cdots, \sigma_r$, i.e.,

$$C_K = \sum_{i=1}^{K} u_i v_i^T = U \Sigma_K V^T \tag{5}$$

where $\Sigma_K = \operatorname{diag}\{\sigma_1, \cdots, \sigma_K, 0, \cdots, 0\}$, then the approximate transfer function becomes

$$H_K(z_1, z_2) = Z_1^T C_K Z_2 = \sum_{i=1}^{K} f_i(z_1) g_i(z_2). \tag{6}$$

As suggested by (6), the approximate transfer function admits a realization which can be obtained directly from Fig. 1 by neglecting the last $r-K$ sections of the network. An immediate consequence one can draw from this simplification is that if $H(z_1, z_2)$ has linear phase response, so does $H_K(z_1, z_2)$. It is also obvious that the computation complexity for realizing $H_K(z_1, z_2)$ is reduced by $\frac{r-K}{r}$ over the realization scheme in Fig. 1. The rest of this section will be devoted to the problem of how much error will be introduced into an approximate realization that uses only K sections instead of r sections.

Two often used error criteria are the L_2 error and L_∞ error defined by

$$e_2^2 = \frac{1}{(2\pi j)^2} \oint_{|z_1|=1} \oint_{|z_2|=1} |H(z_1, z_2) - H_K(z_1, z_2)|^2 \times z_1^{-1} z_2^{-1} dz_1 dz_2 \tag{7}$$

and

$$e_\infty = \sup_{|z_1|=1, |z_2|=1} |H(z_1, z_2) - H_K(z_1, z_2)| \tag{8}$$

respectively. It is known that the L_2 error can be calculated precisely as [7], [10]

$$e_2 = \left[\sum_{i=K+1}^{r} \sigma_i^2 \right]^{1/2} \tag{9}$$

and a rough upper bound of e_∞ is also available from [7], that is

$$e_\infty \leq \sqrt{N_1 N_2}\, \sigma_{K+1} \equiv \gamma_K. \tag{10}$$

In what follows, our attention will be focused on obtaining a new L_∞ bound superior to (10).

Note that

$$\begin{aligned} H(z_1, z_2) - H_K(z_1, z_2) &= Z_1^T \left(\sum_{i=K+1}^{r} u_i v_i^T \right) Z_2 \\ &= (Z_1^T U_K)(V_K^T Z_2) \end{aligned}$$

where

$$U_K = [u_{K+1} \quad \cdots \quad u_r], \quad V_K = [v_{K+1} \quad \cdots \quad v_r]. \tag{11}$$

It follows from the Schwarz inequality that

$$e_\infty \leq \beta_K \tag{12}$$

where

$$\beta_K = \mu_K \nu_K \tag{13}$$

$$\mu_K = \max_{|z_1|=1} \|Z_1^T U_K\|_2 \tag{14}$$

$$\nu_K = \max_{|z_2|=1} \|V_K^T Z_2\|_2 \tag{15}$$

and $\|\cdot\|_2$ denotes the Euclidean norm of the vector involved.

B. Computation of μ_K and ν_K

To calculate μ_K, we denote

$$U_K = \begin{bmatrix} p_0 \\ \vdots \\ p_{N_1-1} \end{bmatrix} \tag{16}$$

where each p_i is a row vector of dimension $r-K$, and write

$$\|Z_1^T U_K\|_2^2 = \left\| \sum_{l=0}^{N_1-1} p_l z^{-l} \right\|_2^2 = \sum_{k=-(N_1-1)}^{(N_1-1)} \phi_k z^k \tag{17}$$

where ϕ_k is determined by

$$\phi_k = \sum_{i=0}^{N_1-1} p_{i-k} p_i^T \tag{18}$$

with the convention that $p_i = 0$ for $i < 0$. By (18) it is obvious that $\phi_k = \phi_{-k}$ for $k = 1, \cdots, N_1 - 1$. Hence on the unit circle $z_1 = e^{j\omega}$ we have

$$\sum_{k=-(N_1-1)}^{N_1-1} \phi_k z^k = \sum_{k=0}^{N_1-1} \hat{\phi}_k \cos k\omega \tag{19}$$

with

$$\hat{\phi}_k = \begin{cases} \phi_0 & \text{for } k = 0 \\ 2\phi_k & \text{for } 0 < k \leq N_1 - 1. \end{cases} \tag{20}$$

From (17) and (19) we now obtain

$$\mu_K = \max_{0 \leq \omega \leq \pi} \left(\sum_{k=0}^{N_1-1} \hat{\phi}_k \cos k\omega \right)^{1/2} \tag{21}$$

where the interval of ω has been reduced by half since the function involved is an even function. Feasible numerical methods are available [11], [12] for obtaining an accurate maximum value of the sum in (21). Since the function involved has multiple local maximums, a set of coarse grid point needs to be placed on [0, π] and a certain one-dimensional optimization algorithm such as golden section search method can be applied to each subinterval to find a local maximum. The global maximum is then obtained by examining the finite number of local maximums. An alternative approach to finding μ_K is to evaluate the sum in (21) over a set of fine grid points on [0, π] and identify the largest value obtained.

Likewise, if we denote

$$V_K^T = [q_0 \quad \cdots \quad q_{N_2-1}] \tag{22}$$

and define

$$\psi_k = \sum_{i=0}^{N_2-1} q_i^T q_{i-k} \tag{23}$$

$$q_i = 0 \quad \text{for } i < 0 \tag{24}$$

$$\hat{\psi}_k = \begin{cases} \psi_0 & \text{for } k = 0 \\ 2\psi_k & \text{for } 0 < k \leq N_2 - 1 \end{cases} \tag{25}$$

then

$$\nu_K = \max_{0 \leq \omega \leq \pi} \left(\sum_{k=0}^{N_2-1} \hat{\psi}_k \cos k\omega \right)^{1/2}. \tag{26}$$

The above derivation of μ_K and ν_K leads to the following algorithm for the evaluation of β_K.

Algorithm 1 (for computing the L_∞ error bound β_K):

Input: Coefficient matrix C, and K—the number of sections to be used.

Step 1. Perform the SVD of C.

Step 2. Form U_K and V_K using (11).

Step 3. Calculate $\{\hat{\phi}_k,\ 0 \leq k \leq N_1 - 1\}$ and $\{\hat{\psi}_k,\ 0 \leq k \leq N_2 - 1\}$ by using (16), (18), (20), and (22), (23), (25), respectively.

Step 4. Find μ_K and ν_K defined in (21) and (26) by using numerical optimization or the fine-grid approach.

Step 5. Obtain the error bound β_K using (13).

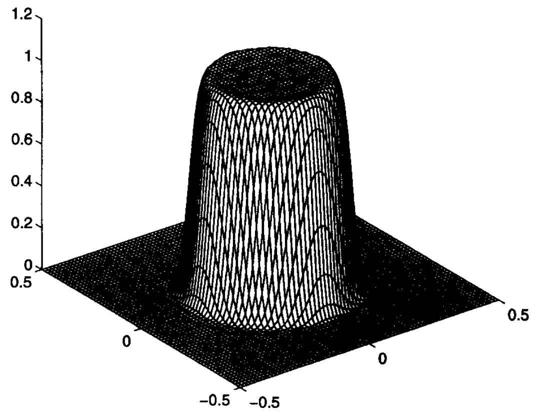

Fig. 2. Amplitude response of a circularly symmetric low-pass FIR filter of order 35 × 35.

Algorithm 1 indicates that as long as the coefficient matrix of a 2-D FIR transfer function (which very likely comes from a preliminary design) is available, the designer can ran Algorithm 1 with the section number K going through a selected set of integers in order to reach a conclusion as to how many sections of subfilters can be neglected from the given transfer function such that the simplified filter remains satisfactory. This is particularly the case when the preliminary design happens to be an overqualified one in terms of the required design specifications. The designer in such a case does not have to perform a re-design but simplify the preliminary design by keeping an appropriate portion of it with less (sometimes substantially less) sections of subfilters. We shall revisit this issue in the next section using an illustrative example.

C. Remarks on Algorithm 1

1) It is noted that for those filters with $N_1 = N_2$ and symmetric coefficient matrix C, the weighted singular vectors obey $u_i = \pm v_i$ for $1 \le i \le r$. In this case it can readily be verified that $\mu_K = \nu_K$ for any $K \le r$, and consequently $\beta_K = \mu_K^2$, i.e.,

$$\beta_K = \max_{0\le\omega\le\pi} \sum_{k=0}^{N_1-1} \hat{\phi}_k \cos k\omega$$

making the algorithm more efficient.

2) As mentioned earlier, in order to include Algorithm 1 in a design procedure, it becomes necessary to evaluate β_K with K varying in a certain set of integers, say from $K_0, K_0+1 \equiv K_1, \cdots$, to $K_0 + m \equiv K_m$. For each K_j one needs to calculate coefficients $\phi_k^{(j)}$ and $\psi_k^{(j)}$. This remark claims that the calculations can be done recursively and therefore efficiently. To see this, denote

$$U_{K_j} = \begin{bmatrix} p_0^{(j)} \\ \vdots \\ p_{N-1}^{(j)} \end{bmatrix} \tag{27}$$

and

$$\phi_k^{(j)} = \sum_{i=0}^{N_1-1} p_{i-k}^{(j)} p_i^{(j)T} \tag{28}$$

for $j = 0, 1, \cdots, m$, and note that U_{K_j} is an $N_1 \times [r-(K_0+j)]$ matrix. It is obvious that given $U_{K_{j-1}}$, matrix U_{K_j} can be obtained by augmenting U_{K_j} by one column on the left. Hence one can write

$$U_{K_{j-1}} = \begin{bmatrix} \alpha_0^{(j)} & p_0^{(j)} \\ \vdots & \vdots \\ \alpha_{N_1-1}^{(j)} & p_{N_1-1}^{(j)} \end{bmatrix} \tag{29}$$

where $\alpha_0^{(j)}, \cdots, \alpha_{N_1-1}^{(j)}$ are *scalars*. From (27)–(29) it follows that

$$\begin{aligned}\phi_k^{(j-1)} &= \sum_{i=0}^{N_1-1} p_{i-k}^{(j-1)} p_i^{(j-1)T} \\ &= \sum_{i=0}^{N_1-1} p_{i-k}^{(j)} p_i^{(j)T} + \sum_{i=0}^{N_1-1} \alpha_{i-k}^{(j)} \alpha_i^{(j)} \\ &= \phi_k^{(j)} + \delta_k^{(j)}\end{aligned} \tag{30}$$

where

$$\delta_k^{(j)} = \sum_{i=0}^{N_1-1} \alpha_{i-k}^{(j)} \alpha_i^{(j)}. \tag{31}$$

The recursive relation (30) in conjunction with (31) indicates that once $\phi_k^{(m)}$ is computed, $\phi_k^{(j)}$ for $j = m-1, \cdots, 0$ can be computed recursively using (30) where each time only the term $\delta_k^{(j)}$ needs to be evaluated. A relation similar to (30) holds for $\psi_k^{(j)}$ and $\psi_k^{(j-1)}$.

3) Applying the Schwarz inequality to (14) and (15), we have

$$\mu_k \le \max_{|z_1|=1} \|Z_1\|_2 \, \|U_K\|_2 = \sqrt{N_1 \sigma_{K+1}}$$
$$\nu_k \le \max_{|z_2|=1} \|Z_2\|_2 \, \|V_K\|_2 = \sqrt{N_2 \sigma_{K+1}}$$

which imply that

$$\beta_K \le \gamma_K \tag{32}$$

with γ_K given by (10), which indicates that β_K is a better upper bound to use. As a matter of fact, our simulation study shows that for a variety of FIR filters such as circularly symmetic low-pass, boundpass, bandstop, and high-pass filter, fan-shaped and diamond-shaped filter, β_K is much tighter than γ_k with a typical ratio $\gamma_K/\beta_K \approx 6$.

TABLE I
ERROR BOUNDS β_K

K	β_K
10	7.69870155×10^{-8}
9	7.68157185×10^{-6}
8	1.29893761×10^{-4}
7	6.43246554×10^{-4}
6	0.00301410
5	0.01093786
4	0.03197229

III. An Example

As an illustrative example, let us consider a linear-phase, non-recursive, circularly symmetric low-pass 2-D filter with normalized passband edge = 0.2125 and stopband edge = 0.2875. With order $N_1 = N_2 = 35$, the application of the SVD design method [10] produces a linear-phase transfer function $H(z_1, z_2)$ whose amplitude response is depicted in Fig. 2. The maximum ripple of the filter in the passband and stopband are 0.0427 and 0.0407, respectively. The coefficient matrix of $H(z_1, z_2)$ has $\text{rank}(C) = 11$ meaning that the filter needs 11 sections for the implementation.

Using Algorithm 1 in the recursive manner as discussed in Remark 2 of the preceding section for $K = 10, \cdots, 4$, β_K are evaluated and the results are listed in Table I. It follows that if a 5% ripple is allowed in both passband and stopband, then it is sufficient to keep only the first *six* sections that correspond to the six most

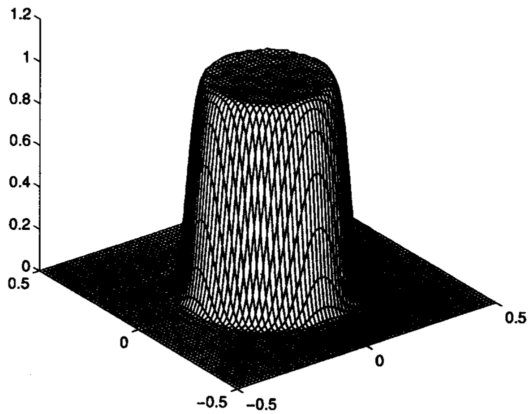

Fig. 3. Amplitude response of $H_6(z_1, z_2)$.

significant singular vectors of matrix C. The amplitude response of the simplified transfer function $H_6(z_1, z_2)$ is depicted in Fig. 3. The actual maximum ripples of $H_6(z_1, z_2)$ in the passband and stopband are 0.0431 and 0.0414, respectively.

Note from Table I that β_6 =0.00301410. To see the tightness of the error bound, we use a very fine grid over the entire baseband and find the numerical value of the L_∞ error

$$\sup_{|z_1|=1, |z_2|=1} |H(z_1, z_2) - H_6(z_1, z_2)| \approx 0.00266762.$$

For comparison, (10) gives an upper bound $\gamma_6 = \sqrt{N_1 N_2}\,\sigma_7 = 0.01906959$ which is 6.3 times larger than β_6.

IV. Concluding Remarks

A new L_∞ error bound for the SVD-based section reduction method has been derived, and the illustrative example described in Section III has demonstrated the feasibility of the section reduction method as well as the tightness of the new error bound. At this point of research one would seek to establish a similar L_∞ (or L_2) error bound for the recursive case as a natural extension of what has been achieved here. This is indeed an on-going research project and the results will be reported elsewhere.

References

[1] P. N. Paraskevopoulos, "Padé type order reduction of two-dimensional systems," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 413–416, May 1980.

[2] B. Lashgari, L. M. Silverman, and J. F. Abramatic, "Approximation of 2-D separable in denominator filters," *IEEE Trans. Circuits Syst.*, vol. CAS-30, pp. 107–121, Feb. 1983.

[3] W.-S. Lu, E. B. Lee, and Q.-T. Zhang, "Model reduction for 2-D systems," in *Proc. 1986 IEEE ISCAS*, May 1986, pp. 79–82.

[4] E. I. Jury and K. Premaratne, "Model reduction of two-dimensional discrete systems," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 558–562, May 1986.

[5] A. Kumar, F. W. Fairman, and J. R. Svensson, "Separately balanced realization and model reduction of 2-D separable denominator transfer functions from input output data," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 233–239, Mar. 1987.

[6] K. Premaratne and E. I. Jury, "Model reduction of two-dimensional discrete systems via balanced realization," in *Proc. 21st Asilomar Conf. Signals, Syst. and Comput.*, Nov. 1987.

[7] W.-S. Lu and E. B. Lee, "Structure simplification of 2-D digital filters: An overview," in *Lecture Notes in Control and Information Sciences*, M. Thoma and A. Wyner, Eds. New York: Springer-Verlag, 1989, vol. 129, pp. 163–177.

[8] G. Gu, P. P. Khargonekar, and E. B. Lee, "Approximation of infinite-dimensional sytems," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 610–618, June 1989.

[9] G. Gu and B. A. Shenoi, "A novel approach to the synthesis of recursive digital filters with linear phase," *IEEE Trans. Circuits Syst.*, vol. 38, pp. 602–612, June 1991.

[10] W.-S. Lu and A. Antoniou, *Two-Dimensional Digital Filters*. New York: Marcel Dekker, 1992.

[11] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed. Reading, MA: Addison-Wesley, 1984.

[12] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. New York: Wiley, 1987.

A Separation Theorem for Finite Precision Digital Filters

Kelly K. Johnson and Irwin W. Sandberg

***Abstract*—For models of zero-input direct-form digital filters implemented in fixed-point digital hardware, it is known that if saturation arithmetic is used, and stability holds when the quantization is ignored, then the amplitude of all limit cycles can be made arbitrarily small by making the bound on the magnitude of the quantization sufficiently small. In that sense, the effects of quantization and overflow can be considered separately. Here we extend this proposition to the important case in which the input need not be zero and a certain more general type of arithmetic is used.**

I. Introduction

A direct-form digital filter implemented in fixed-point digital hardware is governed by the relation

$$w_n = f\left(\sum_{k=1}^{m} b_k w_{n-k} + \sum_{k=0}^{m} a_k u_{n-k}\right) + q_n, \quad n \geq 0 \qquad (1)$$

in which the b_k and a_k are real coefficients, $\{u_n\}_0^\infty$ and $\{w_n\}_0^\infty$ are the input and output sequences, respectively, and $w_{-m}, \cdots, w_{-1}$ and $u_{-m}, \cdots, u_{-1}$ are initial values. The sequence $\{q_n\}_0^\infty$ takes into account quantization, and the function f accounts for overflow. The initial values and the elements of the input, output, and quantization sequences are real numbers typically assumed to belong to $[-1, 1]$. In what follows, it is assumed that the infinite precision counterpart of the digital filter is stable (i.e., that $1 - \sum_{k=1}^{m} b_k z^{-k} \neq 0$ for $|z| \geq 1$). It is well-known that (1) may exhibit zero-input limit cycle behavior due to the effects of the finite-precision elements.

Consider the case in which the input to the digital filter is zero and quantization is ignored. That is, consider

$$w_n = f\left(\sum_{k=1}^{m} b_k w_{n-k}\right), \quad n \geq 0 \qquad (2)$$

Manuscript received November 3, 1994; revised April 26, 1995. This paper was recommended by Associate Editor I. Pitas.

The authors are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA.

IEEE Log Number 9413920.