Least-Squares Approximation of FIR by IIR Digital Filters

Hartmut Brandenstein, *Student Member, IEEE,* and Rolf Unbehauen, *Fellow, IEEE*

***Abstract*— In this paper, an algorithm is presented for the least-squares approximation of FIR filters by IIR filters. The algorithm is an iterative procedure where each iteration requires the solution of an overdetermined set of linear equations and some digital filtering operations. All calculations are performed with the numerator and denominator coefficients of the transfer functions. A conversion to state-space descriptions is not necessary. Examples show that the approximation error is as small as that of the IIR filters obtained with balanced model reduction. Moreover, the effects of numerical errors are negligible. Thus, our algorithm is applicable even in cases where the FIR filter length is large.**

I. Introduction

THERE EXISTS a variety of different methods for the approximation of an FIR filter by an IIR filter of reduced order (e.g., [1]–[5]). Among all approaches, those that are performed in the state space by applying model reduction techniques are the most promising. In [3] and [4], by balancing and truncating the state-space model of the FIR filter, a stable IIR filter that represents a good approximation w.r.t. the Hankel norm [6] was obtained. Since the Hankel norm is an upper bound of the l_2-norm for a stable and strictly proper transfer function [6], this solution is also a good approximation in the least-squares sense. However, if the l_2-norm error criterion is used directly as in [1] and [2], it may be possible to obtain better results. However, since, in [1], a nonlinear optimization procedure is used to solve the problem, there is no guarantee that we can find the global minimum or even a good local minimum of the error norm at all. The applicability of the algorithm, which is presented in [2], is also questionable.

Another approach that is also based on a state-space description of the filters is given in [5]. As the examples in [5] show, the resulting IIR filter is more selective than the IIR filter of the balanced model reduction approaches [3], [4]. However, with the IIR filter of [5], the approximation accuracy of the phase response is worse, and the least-squares approximation error is higher than with the IIR filter of [3] and [4]. It depends on the practical application as to which one of the two methods ([3] and [4] or [5]) is used. Here, we are interested in finding a good approximation in the least-squares sense. Thus, the approaches in [3] and [4] should be preferred. In [4], valuable simplifications were introduced by exploiting the special structure of the system matrices of the FIR filter, and it seems that among the model reduction techniques that of [4] is the most powerful in this context.

In [4], the main part of the calculations is a singular value decomposition (SVD) of an $L \times L$ Hankel matrix, where $L+1$ is the length of the FIR filter, and after that the numerator and denominator coefficients of the IIR filter's transfer function must be derived from the state-space description of the IIR filter. This can be done without a problem as long as L is small. However, for large L, the effects of finite precision arithmetic become obvious. The SVD, as well as the conversion from the state-space model to the transfer function, can produce large numerical errors that seriously deteriorate the results. The numerical errors in the state-space to transfer function conversion can be avoided to some degree by using Faddeev's method [7] to determine the characteristic polynomial of a matrix, but this considerably increases the amount of calculations. This fact, together with a possible suboptimality of the results (in the sense of least squares), is a drawback of all state-space model reduction approaches.

In this paper, we propose an algorithm for minimizing the l_2-norm of the approximation error. The starting point is a theorem that enables us to reformulate the approximation problem as a problem of designing an allpass filter with certain characteristics. This naturally leads to an algorithm that produces good results even if L is large, and considering numerous examples, we found that the l_2-norms of the approximation errors of our filters are as small as those of the filters obtained with [4], and the selectivity of the filters is similar. Moreover, due to the simplicity of our algorithm, the computational complexity of our approach is considerably smaller than the computational complexity of [4].

II. The Theorem of Walsh

Least-squares approximation of an FIR digital filter with transfer function

$$F(z) = \sum_{\lambda=0}^{L} f_\lambda z^{-\lambda} \tag{1}$$

by an IIR digital filter with transfer function

$$H(z) = \frac{P(z)}{Q(z)} \tag{2}$$

of lower order N $(N < L)$, where

$$P(z) = \sum_{\nu=0}^{N} p_\nu z^{-\nu} \tag{3}$$

Manuscript received October 23, 1995; revised June 26, 1997. The associate editor coordinating the review of this paper and approving it for publication was Dr. Victor E. DeBrunner.

The authors are with the Lehrstuhl für Allgemeine und Theoretische Elektrotechnik, Universität Erlangen-Nürnberg, Erlangen, Germany (e-mail: bra@late.e-technik.uni-erlangen.de).

Publisher Item Identifier S 1053-587X(98)00515-7.

and

$$Q(z) = \sum_{\nu=0}^{N} q_\nu z^{-\nu} \quad (q_0 = 1) \tag{4}$$

means that we have to determine $N+1$ real numerator coefficients p_ν $(\nu = 0,1,\ldots,N)$ and N real denominator coefficients q_ν $(\nu = 1,2,\ldots,N)$ such that the l_2-norm of the difference function

$$\Delta(z) = F(z) - H(z) \tag{5}$$

is minimal. The l_2-norm of $\Delta(z)$ is defined by

$$\|\Delta(z)\|_2 := \left[\frac{1}{2\pi \mathrm{j}} \oint_{|z|=1} \Delta(z)\Delta^*(z)\frac{\mathrm{d}z}{z}\right]^{\frac{1}{2}} \tag{6}$$

where the integration must be carried out in the counterclockwise sense, and "$*$" stands for conjugate complex. Since we are interested in obtaining a stable IIR filter, we have to ensure that all the poles $\alpha_1, \alpha_2, \ldots, \alpha_N$ of $H(z)$ lie in $|z| < 1$. Thus, we are concerned with rational transfer functions that are analytic in $|z| \geq 1$. This fact enables us to apply a theorem, which was stated by Walsh [8], to simplify the solution of the approximation problem by a considerable amount. Originally, it was formulated for the approximation of functions that are analytic in $|z| < 1$. However, it can easily be transferred to the case where analyticity in $|z| > 1$ is required. Using our notations, the theorem can be stated as follows.

Theorem 1: Among the set of rational functions $H(z)$, (2)–(4), with prescribed poles α_1, $\alpha_2, \ldots, \alpha_N$ that are fixed and located in $|z| < 1$, the best approximation in the sense of least squares to $F(z)$ (analytic in $|z| > 1$ and continuous in $|z| \geq 1$) is the unique function that interpolates to $F(z)$ in all the points $z = \infty$, $1/\alpha_1^*, 1/\alpha_2^*, \ldots, 1/\alpha_N^*$.

Proof: See [8] for proof. □

This theorem can be used to decouple the problem of finding the numerator coefficients p_ν $(\nu = 0,1,\ldots,N)$ and the determination of the poles α_ν $(\nu = 1,2,\ldots,N)$. It says that once we have found the poles, we get the numerator coefficients as the solution of an interpolation problem. The interpolation conditions can always be used to formulate a set of $N+1$ linear equations in the $N+1$ unknowns p_ν $(\nu = 0,1,\ldots,N)$. This set of equations is uniquely solvable. If all the α_ν $(\nu = 1,2,\ldots,N)$ are distinct from each other and distinct from zero, this fact is obvious. Defining

$$\begin{aligned} z_0 &= \infty, \\ z_\nu &= 1/\alpha_\nu^* \quad (\nu = 1,2,\ldots,N) \end{aligned} \tag{7}$$

we simply have to require

$$H(z_\nu) = F(z_\nu) \quad (\nu = 0,1,\ldots,N). \tag{8}$$

Since our assumption about the poles implies

$$z_i \neq z_j \quad (i \neq j;\ i,j \in \{0,1,\ldots,N\}) \tag{9}$$

we obtain $N+1$ linearly independent (linear) equations in the $N+1$ unknowns p_ν $(\nu = 0,1,\ldots,N)$. However, if multiple poles occur and/or if at least one pole is at $z = 0$, two or more of the equations in (8) are identical, and the solution of (8) is not unique. In this case, we have to proceed as follows. Assume that after a possible renumbering of the indices

$$z_i \neq z_j \quad (i \neq j;\ i,j \in \{0,1,\ldots,r-1\}) \tag{10}$$

and

$$z_i \in \{z_0, z_1, \ldots, z_{r-1}\} \quad (i \in \{r, r+1, \ldots, N\}) \tag{11}$$

are valid, where r $(1 \leq r \leq N)$ denotes the number of different points in the set $\{z_\nu\}$ $(\nu = 0,1,\ldots,N)$. Then, each point z_k $(k \in \{0,1,\ldots,r-1\})$ occurs with multiplicity s_k in the set z_ν $(\nu = 0,1,\ldots,N)$, and

$$\sum_{k=0}^{r-1} s_k = N+1 \tag{12}$$

is valid. Now, interpolation of $H(z)$ to $F(z)$ in the points z_k means the requirements

$$\begin{aligned} \frac{\mathrm{d}^{\sigma_k}}{(\mathrm{d}z^{-1})^{\sigma_k}} H(z)\bigg|_{z=z_k} &= \frac{\mathrm{d}^{\sigma_k}}{(\mathrm{d}z^{-1})^{\sigma_k}} F(z)\bigg|_{z=z_k} \\ (\sigma_k = 0,1,\ldots,s_k-1;\ k = 0,1,\ldots,r-1). \end{aligned} \tag{13}$$

These $N+1$ equations are linearly independent and linear in the unknowns p_ν $(\nu = 0,1,\ldots,N)$. Thus, the determination of $P(z)$ is straightforward.

Note that (13) is valid whether or not $F(z)$ is an FIR transfer function. The only requirement on $F(z)$ is that $F(z)$ is analytic in $|z| > 1$ and continuous in $|z| \geq 1$. In our case, however, $F(z)$ is an FIR transfer function, and this fact can be employed to simplify the procedure by making use of (13) implicitly. This is done in the following section.

III. Determination of the Numerator

Let $F(z)$ and $Q(z)$ be given. We want to determine $P(z)$ such that (13) is satisfied. Equation (13) implies that $N+1$ zeros of the difference function $\Delta(z)$ are located at $z = z_\nu$ $(\nu = 0,1,\ldots,N)$. Thus, $\Delta(z)$ can be written in the form

$$\Delta(z) = \frac{z^{-(N+1)}Q(z^{-1})}{Q(z)} R(z) \tag{14}$$

where

$$R(z) = \sum_{\lambda=0}^{L-1} r_\lambda z^{-\lambda} \tag{15}$$

is an unknown FIR transfer function with real coefficients. Substituting (2) and (14) into (5) yields

$$F(z)Q(z) - P(z) = z^{-(N+1)}Q(z^{-1})R(z) \tag{16}$$

which is equivalent to

$$\begin{aligned} &[z^{-L}F(z^{-1})][z^{-N}Q(z^{-1})] - z^{-L}[z^{-N}P(z^{-1})] \\ &\quad = Q(z)[z^{-(L-1)}R(z^{-1})] \end{aligned} \tag{17}$$

and with the definitions

$$\begin{aligned} X_1(z) &:= z^{-L}F(z^{-1}) \\ X_2(z) &:= z^{-N}P(z^{-1}) \\ Y(z) &:= z^{-(L-1)}R(z^{-1}) \end{aligned} \tag{18}$$

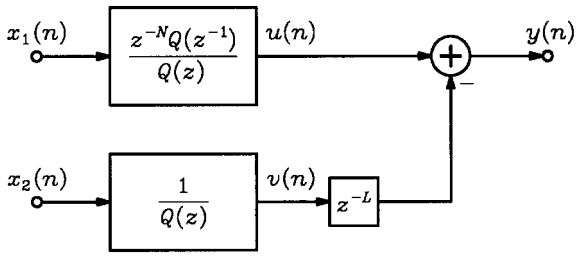

Fig. 1. Digital filter described by (19).

we obtain from (17)

$$Y(z)=\frac{z^{-N}Q(z^{-1})}{Q(z)}X_1(z)-\frac{z^{-L}}{Q(z)}X_2(z). \tag{19}$$

The functions $X_1(z)$, $X_2(z)$, and $Y(z)$ can be regarded as z-transforms of the signals with finite support

$$x_1(n)=\begin{cases}0, & n<0\\ f_{L-n}, & 0\le n\le L\\ 0, & n>L\end{cases} \tag{20}$$

$$x_2(n)=\begin{cases}0, & n<0\\ p_{N-n}, & 0\le n\le N\\ 0, & n>N\end{cases} \tag{21}$$

and

$$y(n)=\begin{cases}0, & n<0\\ r_{L-1-n}, & 0\le n\le L-1\\ 0, & n>L-1\end{cases} \tag{22}$$

respectively. Now, (19) can be interpreted as the input–output description of the digital filter shown in Fig. 1. The relationship between the output $y(n)$ and the two signals $u(n)$ and $v(n)$, which were both introduced in Fig. 1, is given by

$$y(n)=u(n)-v(n-L). \tag{23}$$

Especially in the interval $0\le n\le L-1$, we have

$$y(n)=u(n)\quad(0\le n\le L-1). \tag{24}$$

Substituting (22) into (24) yields

$$r_{L-1-n}=u(n)\quad(n=0,1,\ldots,L-1). \tag{25}$$

Now, a simple method to calculate $R(z)$ from $F(z)$ and $Q(z)$ has been found. By allpass filtering of the signal $x_1(n)$ that is constructed by reversing the order of the samples of the impulse response of $F(z)$, we obtain a signal $u(n)$, whose samples in the interval $0\le n\le L-1$ are the coefficients of $R(z)$, according to (25). The allpass filter is described by its transfer function

$$A(z)=\frac{z^{-N}Q(z^{-1})}{Q(z)}. \tag{26}$$

Having determined $R(z)$, we finally get $P(z)$ [see (3)], which is a polynomial in z^{-1} of degree N, from (16) as

$$P(z)=F(z)Q(z)-z^{-(N+1)}Q(z^{-1})R(z). \tag{27}$$

Note that we do not need to know the roots of $Q(z)$ explicitly in order to obtain $P(z)$ such that (13) is satisfied. As a consequence, our calculations are not influenced by the numerical inaccuracies of a root-finding procedure that would be necessary if (13) were used directly.

As we saw, $P(z)$ can easily be determined from $F(z)$ and $Q(z)$. Thus, we only consider $Q(z)$ in the following and try to find a procedure to adjust its coefficients such that $\|\Delta(z)\|_2$ [see (6)] is minimized.

IV. Determination of the Denominator

Denoting the approximation error $\|\Delta(z)\|_2$ as E, we get from (6), (14), (15), and (25)

$$E^2:=\|\Delta(z)\|_2^2=\|R(z)\|_2^2=\sum_{\lambda=0}^{L-1}r_\lambda^2=\sum_{n=0}^{L-1}u^2(n). \tag{28}$$

The signal $u(n)$ is the output of the allpass filter with transfer function $A(z)$ if the input signal is $x(n):=x_1(n)$ see (20). Now, $Q(z)$ shall be chosen such that E^2 is minimal. Since, for an allpass filter, the energy of the input signal equals the energy of the output signal, we have

$$\sum_{n=0}^{\infty}x^2(n)=\sum_{n=0}^{\infty}u^2(n). \tag{29}$$

The approximation problem can now be formulated as follows: Determine $Q(z)$ such that for the allpass filter $A(z)$ with input signal $x(n)=x_1(n)$, the energy of the output signal $u(n)$ in $0\le n\le L-1$ is as small as possible. In view of (29), this implies that the energy of $u(n)$ should be concentrated in $n\ge L$. A similar requirement of distributing the output energy of an allpass filter in time has already been given by Friedman [1], although the approach in [1] is different from ours. In [1], a network structure for the realization of the IIR filter is derived, and the multiplier values of this filter structure are obtained by nonlinear optimization with a steepest descent algorithm. However, there is no guarantee to find a good solution. Here, to avoid this difficulty, we do not use optimization techniques to solve the problem. Instead, we propose the following iterative procedure that produces results at low expense.

Starting with $Q^{(0)}(z)=1$, we recursively calculate polynomials

$$\begin{aligned}Q^{(k)}(z)&=\sum_{\nu=0}^{N}q_\nu^{(k)}z^{-\nu}\quad\left(q_0^{(k)}=1\right)\\ &=1+z^{-1}Q_1^{(k)}(z)\quad(k=1,2,\ldots)\end{aligned} \tag{30}$$

where

$$Q_1^{(k)}(z)=\sum_{\nu=0}^{N-1}q_{\nu+1}^{(k)}z^{-\nu}\quad(k=1,2,\ldots) \tag{31}$$

such that for a sufficiently large K, the allpass $A(z)$ with denominator

$$Q(z)=Q^{(K)}(z) \tag{32}$$

has the desired characteristics. We first consider a digital filter with transfer function

$$A^{(k)}(z)=\frac{z^{-N}Q^{(k)}(z^{-1})}{Q^{(k-1)}(z)}. \tag{33}$$

It approaches an allpass if $||Q^{(k)}(z) - Q^{(k-1)}(z)||_2 \to 0$ for $k \to \infty$. If the input signal of $A^{(k)}(z)$ is $x(n) = x_1(n)$ [see (20)] with its z-transform

$$X(z) = z^{-L}F(z^{-1}) \tag{34}$$

then the z-transform of the output signal is given by

$$U^{(k)}(z) = \frac{z^{-N}Q^{(k)}(z^{-1})}{Q^{(k-1)}(z)}X(z) = \sum_{n=0}^{\infty} u^{(k)}(n)z^{-n}. \tag{35}$$

Defining

$$X^{(k)}(z) := \frac{X(z)}{Q^{(k-1)}(z)} = \sum_{n=0}^{\infty} x^{(k)}(n)z^{-n} \tag{36}$$

we can write

$$U^{(k)}(z) = z^{-N}Q^{(k)}(z^{-1})X^{(k)}(z). \tag{37}$$

Now, we first want to determine a $Q^{(k)}(z)$ that minimizes

$$E_1^{(k)} = \left[\sum_{n=0}^{L-1}[u^{(k)}(n)]^2\right]^{\frac{1}{2}}. \tag{38}$$

Substituting (30) into (37) yields

$$X^{(k)}(z)\left[z^{-(N-1)}Q_1^{(k)}(z^{-1})\right] = U^{(k)}(z) - z^{-N}X^{(k)}(z). \tag{39}$$

Equating the coefficients of $z^0, z^{-1}, \ldots, z^{-(L-1)}$ on both sides of (39), we get with the vectors

$$\begin{aligned}\mathbf{q}^{(k)} &:= \left[q_N^{(k)}, q_{N-1}^{(k)}, \ldots, q_1^{(k)}\right]^{\mathrm{T}}\\ \mathbf{u}^{(k)} &:= [u^{(k)}(0), u^{(k)}(1), \ldots, u^{(k)}(L-1)]^{\mathrm{T}}\\ \mathbf{b}^{(k)} &:= -[0, \ldots, 0, x^{(k)}(0), \ldots, x^{(k)}(L-N-1)]^{\mathrm{T}}\end{aligned} \tag{40}$$

and the $L \times N$ matrix

$$\mathbf{A}^{(k)} := \begin{bmatrix} x^{(k)}(0) & 0 & \cdots & 0 \\ x^{(k)}(1) & x^{(k)}(0) & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ x^{(k)}(N-1) & \cdots & & x^{(k)}(0) \\ \vdots & & & \vdots \\ x^{(k)}(L-1) & \cdots & & x^{(k)}(L-N) \end{bmatrix} \tag{41}$$

the set of equations

$$\mathbf{A}^{(k)}\mathbf{q}^{(k)} = \mathbf{u}^{(k)} + \mathbf{b}^{(k)} \tag{42}$$

where the vectors $\mathbf{q}^{(k)}$ and $\mathbf{u}^{(k)}$ are unknown. Solving

$$\mathbf{A}^{(k)}\mathbf{q}^{(k)} = \mathbf{b}^{(k)} \tag{43}$$

in the least-squares sense, we obtain a vector $\mathbf{q}^{(k)}$ that causes the vector

$$\mathbf{u}^{(k)} = \mathbf{A}^{(k)}\mathbf{q}^{(k)} - \mathbf{b}^{(k)} \tag{44}$$

to be of minimum Euclidean norm, and as a consequence, we have found $Q^{(k)}(z)$ such that $E_1^{(k)}$ is minimum.

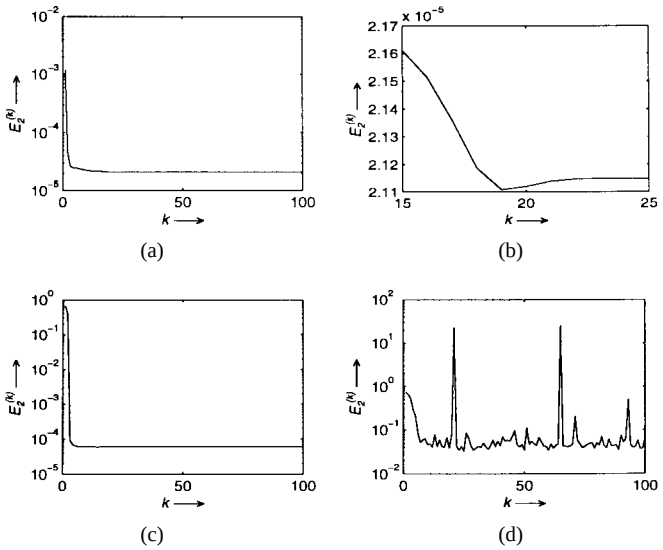

Fig. 2. Error curves of Example 1 ($L = 99$). (a) Error curve for the linear phase case ($N = 49$). (b) $E_2^{(k)}$ for the linear phase case in the interval $15 \le k \le 25$. (c) Error curve for the maximum phase case with $N = 85$. (d) Error curve for the maximum phase case with $N = 75$.

Repeating this procedure for $k = 1, 2, \ldots$, where the elements of $\mathbf{A}^{(k)}$ and $\mathbf{b}^{(k)}$ must be updated by allpole filtering of $x(n) = x_1 = (n)$ [see (34)], according to (36), we can construct a sequence of polynomials $Q^{(k)}(z)$; see (30). For each $Q^{(k)}(z)$, we can find a polynomial

$$R^{(k)}(z) = \sum_{\lambda=0}^{L-1} r_\lambda^{(k)} z^{-\lambda} \tag{45}$$

such that $Q^{(k)}(z)$ and $R^{(k)}(z)$ satisfy a relation analoguous to (16) (by allpass filtering of $x(n) = x_1(n)$ with the allpass $z^{-N}Q^{(k)}(z^{-1})/Q^{(k)}(z)$ as described in Section III). This polynomial $R^{(k)}(z)$ can be used to calculate

$$E_2^{(k)} = ||R^{(k)}(z)||_2 \tag{46}$$

which is the l_2-norm of the approximation error if $F(z)$ is approximated by an IIR transfer function with denominator $Q^{(k)}(z)$. From the sequence of polynomials $Q^{(k)}(z)$, we can now choose that one with minimum error norm $E_2^{(k)}$ as $Q(z)$. The resulting IIR filter is always stable. This is a consequence of the following theorem.

Theorem 2: The polynomial $Q^{(k)}(z)$ [see (30)] that minimizes $E_1^{(k)}$ [see (38) and cf. (35) and (37)] for arbitrarily given $X^{(k)}(z)$ has no zeros in $|z| \ge 1$.

Proof: See the Appendix. □

Fig. 2(a) shows an example (see Example 1) of how $E_2^{(k)}$ decreases in 100 iterations. This error curve is a typical one. We obtained similar curves in all cases, where the FIR filter was linear or minimum phase as well as in most maximum-phase cases. However, for the latter, we observed effects that are described at the end of this section.

From Fig. 2(a), it is obvious that $E_2^{(k)}$ converges, but note that the minimum of $E_2^{(k)}$ and the limit point of $E_2^{(k)}$ are not necessarily identical. This can be seen from Fig. 2(b),

where $E_2^{(k)}$ for $k = 15 \ldots 25$ is given. The minimum occurs at $k = 19$. In this case, we would choose $K = 19$ in (32).

It can also be realized from Fig. 2(a) that $E_2^{(k)}$ decreases rapidly in the first iterations. Thus, only few iterations are necessary to obtain a good approximation. In the examples given in the next section, as well as in numerous other examples, we found that 20 iterations are sufficient in most cases, and increasing the number of iterations yields only small improvements. One could try to further decrease the approximation error by applying a gradient procedure with E [cf. (28)] as objective function and with the coefficients of $Q(z)$ [see (32)] as initial values. However, we found that this would not significantly improve the results, which indicates that with our algorithm, we reach a minimum of the l_2-norm of $\Delta(z)$.

If the FIR filter is maximum phase, the error curves are, in most cases, similar to that of Fig. 2(a) and (b) [see Fig. 2(c)], but for $L > 50$, it is possible that the progression of $E_2^{(k)}$ is like that given in Fig. 2(d). Obviously, $E_2^{(k)}$ does not converge, and there are large peaks in the approximation error. The peaks occur when the solution of (43) leads to a polynomial $Q^{(k)}(z)$ with zeros in $|z| \geq 1$. This originates from numerical influences in cases when the matrix $\mathbf{A}^{(k)}$ is ill conditioned. Nevertheless, the polynomial $Q^{(k)}(z)$ that belongs to the minimum value $E_2^{(k)}$ has all its zeros in $|z| < 1$. Thus, our approach is applicable also in the maximum phase case.

V. Examples

The proposed algorithm is illustrated by the following examples, where we first (Example 1) consider its convergence properties and then (Examples 2–5) compare the IIR filters obtained with our approach (20 iterations) with the IIR filters obtained with balanced model reduction [4]. In Example 6, we finally show the applicability of our approach when the FIR filter length is large.

Example 1: In this example, we want to complete the study of the convergence behavior of our algorithm and give the details of Fig. 2. The FIR filter is a lowpass filter with $L = 99$, passband edge frequency $\Omega_p = 0.6\pi$, and stopband edge frequency $\Omega_s = 0.7\pi$.

First, we chose $F(z)$ to be linear phase, and with an IIR filter order $N = 49$, we obtained $E_2^{(k)}$, as shown in Fig. 2(a). Here, we have $\lim_{k\to\infty} E_2^{(k)} = 2.1143 \cdot 10^{-5}$. However, the minimum occurs at $k = 19$ [see Fig. 2(b)]. Its value is $E_2^{(19)} = 2.1109 \cdot 10^{-5}$.

Now, we consider the maximum-phase case. With $N = 85$, we obtained $E_2^{(k)}$ as given in Fig. 2(c). Here, $E_2^{(k)}$ converges to $\lim_{k\to\infty} E_2^{(k)} = 5.9724 \cdot 10^{-5}$. The minimum occurs at $k = 15$. Its value is $E_2^{(15)} = 5.9151 \cdot 10^{-5}$. With $N = 75$, we obtained $E_2^{(k)}$, as shown in Fig. 2(d). Obviously, $E_2^{(k)}$ does not converge. The minimum occurs at $k = 25$. Its value is $E_2^{(25)} = 0.0324$, and the IIR filter with denominator $Q^{(25)}(z)$ is stable. Note that in this case, the IIR filter we got with [4] was unstable.

In the following, we consider only linear phase FIR filters. They were computed with the Remez exchange algorithm [9].

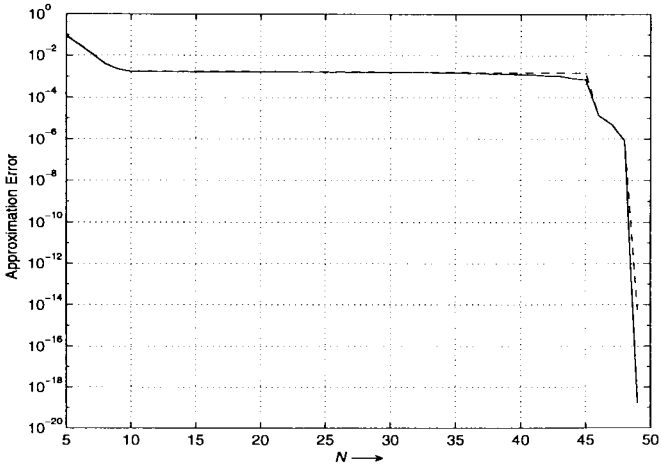

Fig. 3. Approximation of a linear-phase FIR lowpass filter ($L = 50$) by an IIR filter of order N: Approximation errors for $N = 5 \cdots 49$ (—— our algorithm; – – – the method of [4]).

For a given FIR filter of length $L + 1$, we calculated IIR filters of various orders N in two ways: with our algorithm and with the method of [4]. Thus, we are able to compare the two approaches w.r.t. the approximation quality in the least-squares sense. The approximation errors of our approach were computed according to (28), and the approximation errors of [4] were obtained by evaluating (6).

Subsequently, we consider the IIR filters we obtained with the two approaches for a fixed N, where N was chosen such that the approximation errors were sufficiently small. We compare the frequency responses of both IIR filters with the frequency reponse of the FIR filter, regarding the minimum stopband attenuation (MSA), the maximum passband attenuation (MPA), and the deviation of the group delay from the constant value $L/2$. The MPA is found to be sufficiently small as soon as the MSA is large enough.

Example 2: In this example, the FIR filter is a narrowband lowpass filter with $L = 50$, passband edge frequency $\Omega_p = 0.1\pi$, and stopband edge frequency $\Omega_s = 0.2\pi$. Fig. 3 shows the approximation errors of the IIR filters obtained with our approach (IIR1) and the approximation errors of the IIR filters obtained with [4] (IIR2) for $N = 5 \cdots 49$. Obviously, the difference between the two error curves is small. The same can be observed in Examples 3–5.

As can be seen from Fig. 3, a small approximation error can already be achieved with $N = 10$, which equals $L/5$, and it would not be reasonable to choose $N > 10$ since there is no significant reduction of the approximation error until $N > 45$. In Fig. 4, the frequency responses of the two IIR filters with $N = 10$ are shown, and it can be confirmed that both IIR filters satisfy the magnitude specifications of the FIR filter (FIR: MSA 48.78 dB; IIR1: MSA 48.77 dB; IIR2: MSA 49.26 dB) and that the deviations of the IIR filters' group delays from the constant $L/2$ are small. The latter also applies for the following examples.

Example 3: In this example, the FIR filter is a wideband lowpass filter with $L = 71$, passband edge frequency $\Omega_p = 0.8\pi$, and stopband edge frequency $\Omega_s = 0.9\pi$. Fig. 5 shows the approximation errors of IIR1 and IIR2 for $N = 10 \cdots 70$.

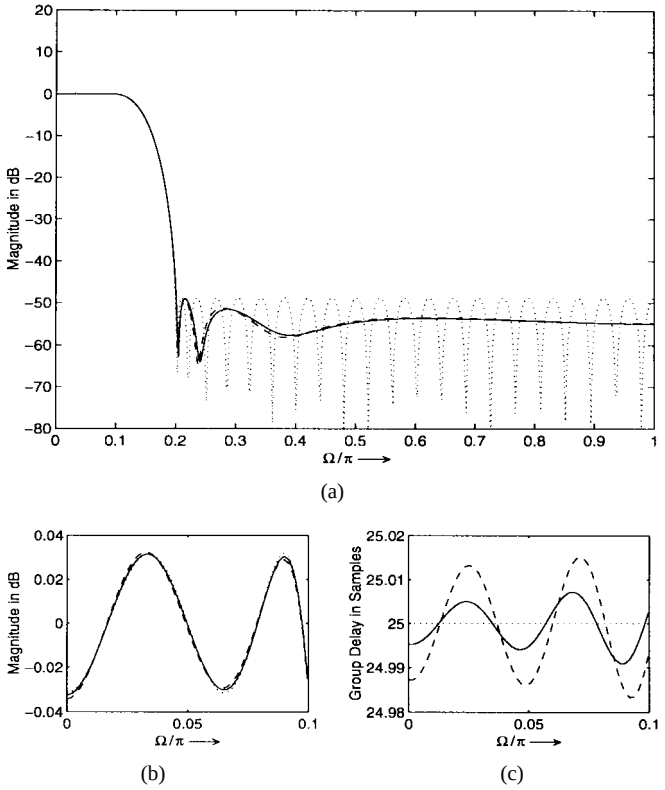

Fig. 4. Frequency responses of two IIR filters with $N = 10$ that approximate a linear-phase FIR lowpass filter ($L = 50$). (a) Magnitude frequency responses of the filters. (b) Magnitude frequency responses of the filters in the passband. (c) Group delays of the filters in the passband. ($\cdots\cdots$ FIR filter; —— IIR filter obtained with our algorithm; – – – IIR filter obtained with the method of [4]).

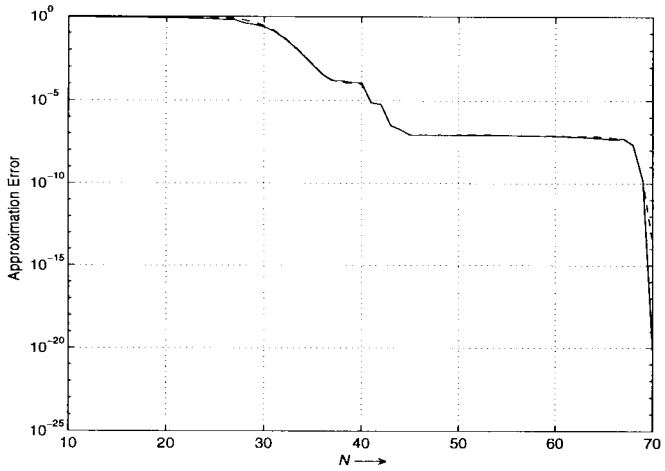

Fig. 5. Approximation of a linear-phase FIR lowpass filter ($L = 71$) by an IIR filter of order N. Approximation errors for $N = 10 \cdots 70$ (—— our algorithm; – – – the method of [4]).

It can be seen that a good approximation can be achieved with $N = 40$. In Fig. 6, the frequency responses of the two IIR filters with $N = 40$ are shown. The FIR filter has a MSA of 64.36 dB. The MSA of IIR1 is 63.16 dB, and the MSA of IIR2 is 64.26 dB.

Example 4: In this example, the FIR filter is a bandstop filter with $L = 100$. Its specifications are listed in Table I, where the edge frequencies are normalized w.r.t. half the

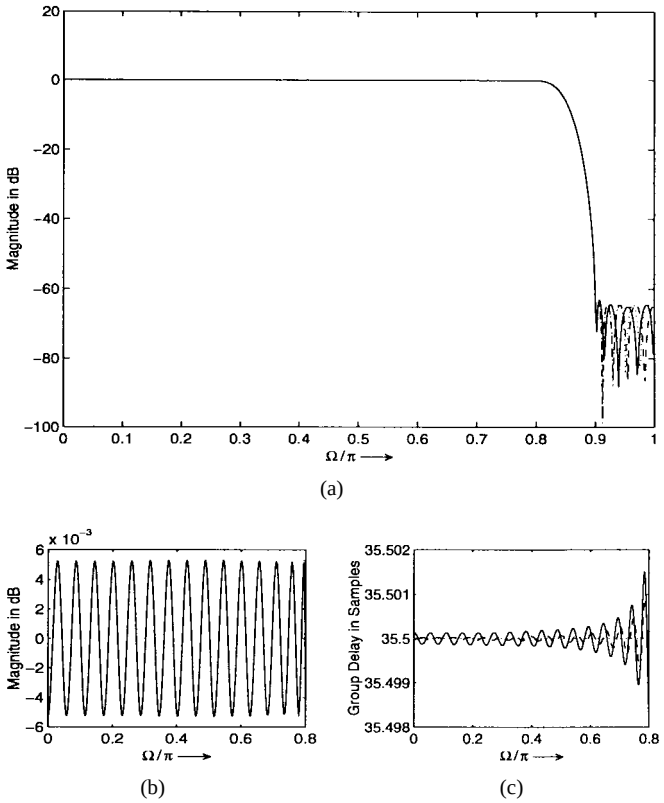

Fig. 6. Frequency responses of two IIR filters with $N = 40$ that approximate a linear phase FIR lowpass filter ($L = 71$). (a) Magnitude frequency responses of the filters. (b) Magnitude frequency responses of the filters in the passband. (c) Group delays of the filters in the passband. ($\cdots\cdots$ FIR filter; —— IIR filter obtained with our algorithm; – – – IIR filter obtained with the method of [4]).

sampling rate. In Fig. 7, the approximation errors of IIR1 and IIR2 for $N = 10 \cdots 99$ are given. To achieve a good approximation, we have to choose $N > L/2$. Fig. 8 shows the frequency responses of the two IIR filters with $N = 54$. In Fig. 8(d) and (e), for a better graphical representation, the constant $L/2 = 50$ is subtracted from the group delays of the filters. The MSA of the FIR filter is 80.31 dB. The MSA of IIR1 is 81.40 dB and the MSA of IIR2 is 81.41 dB, and there is no visible difference between the frequency responses of IIR1 and IIR2.

Example 5: In this example, the FIR filter is a bandpass filter with $L = 120$. Its specifications are listed in Table II. In Fig. 9, the approximation errors of IIR1 and IIR2 for $N = 12 \cdots 119$ are given. Obviously, we have a small approximation error for $N = 60$. Fig. 10 shows the frequency responses of the two IIR filters of order $N = 60$. In Fig. 10(c), for a better graphical representation, the constant $L/2 = 60$ is subtracted from the group delays of the filters. The FIR filter has a MSA of 98.90 dB in both stopbands. The MSA of IIR1 is 96.93 dB in the left stopband and 96.87 dB in the right stopband, and the MSA of IIR2 is 98.09 dB in the left and 97.98 dB in the right stopband.

As can be seen from Examples 2–5, the approximation error curve of our algorithm and the approximation error curve of [4] are always nearly identical, and the frequency responses of the filters are similar. With our algorithm, however, the necessary

TABLE I
DESIGN SPECIFICATIONS FOR THE LINEAR-PHASE FIR BANDSTOP FILTER OF EXAMPLE 4

	Band 1	Band 2	Band 3
Lower edge	0.00	0.49	0.60
Upper edge	0.40	0.51	1.00
Desired value	1	0	1

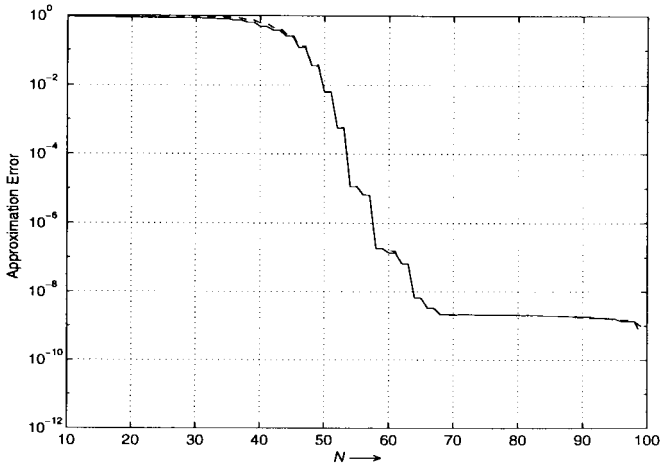

Fig. 7. Approximation of a linear-phase FIR bandstop filter ($L = 100$) by an IIR filter of order N. Approximation errors for $N = 10 \cdots 99$ (—— our algorithm; – – – the method of [4]).

amount of calculations is considerably smaller than with the method of [4]. Furthermore, our algorithm works well, even in cases where L is very large. To prove this, we consider the following example.

Example 6: In this example, the FIR filter is a lowpass filter with $L = 1000$, passband edge frequency $\Omega_p = 0.5\pi$, and stopband edge frequency $\Omega_s = 0.51\pi$. Its magnitude frequency response is given in Fig. 11(a). Fig. 11(b) shows the approximation errors of the IIR filters, which were obtained with our algorithm for $N = 100, 150, \ldots, 950$. Obviously, a good approximation can be achieved with an IIR filter of order $N = 500$. The magnitude frequency response of this IIR filter is given in Fig. 11(c); its group delay in the passband is shown in Fig. 11(d). The IIR filter is stable, and the approximation error is $E = 2.0989 \cdot 10^{-5}$.

It is desirable to choose the lowest IIR filter order N that leads to a good approximation of $F(z)$ in the least-squares sense, in advance, rather than applying a trial-and-error method. In [4], the Hankel singular values plot (HSVP) was introduced as an efficient means. An HSVP is a graphical representation of the singular values of the $L \times L$ Hankel matrix

$$\mathbf{H} := \begin{bmatrix} f_1 & f_2 & \cdots & f_L \\ f_2 & f_3 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ f_L & 0 & \cdots & 0 \end{bmatrix}$$

that consists of the coefficients $f_1, f_2, \ldots, f_L$ of $F(z)$. In an HSVP, the singular values $\sigma_1, \sigma_2, \ldots, \sigma_L$ of $\mathbf{H}$ that are arranged in descending order

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_L$$

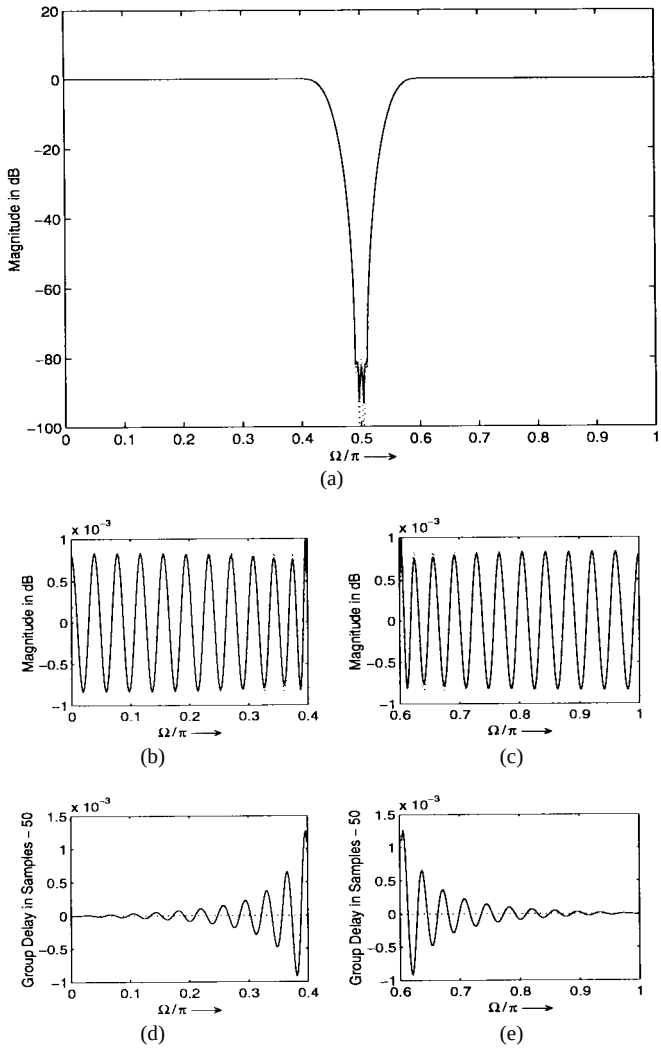

Fig. 8. Frequency responses of two IIR filters with $N = 54$ that approximate a linear-phase FIR bandstop filter ($L = 100$). (a) Magnitude frequency responses of the filters. (b) Magnitude frequency responses of the filters in the left passband. (c) Magnitude frequency responses of the filters in the right passband. (d) Group delays of the filters in the left passband. (e) Group delays of the filters in the right passband. ($\cdots\cdots$ FIR filter; —— IIR filter obtained with our algorithm; – – – IIR filter obtained with the method of [4]).

TABLE II
DESIGN SPECIFICATIONS FOR THE LINEAR-PHASE FIR BANDPASS FILTER OF EXAMPLE 5

	Band 1	Band 2	Band 3
Lower edge	0.00	0.25	0.85
Upper edge	0.15	0.75	1.00
Desired value	0	1	0

are plotted versus their index. Fig. 12 shows the HSVP's for the FIR filters of Examples 2–5. In Fig. 12, the singular values are represented in a logarithmic scale. Comparing Fig. 12(a) to Fig. 3, Fig. 12(b) to Fig. 5, Fig. 12(c) to Fig. 7, and Fig. 12(d) to Fig. 9, we can realize the similarity between the progression of the singular values' magnitude and the progression of the approximation error. Thus, the HSVP can be used to determine a proper N. We can decide from the HSVP how large N must be at least to achieve a good approximation and how the approximation error is reduced (qualitatively) when N is further increased.

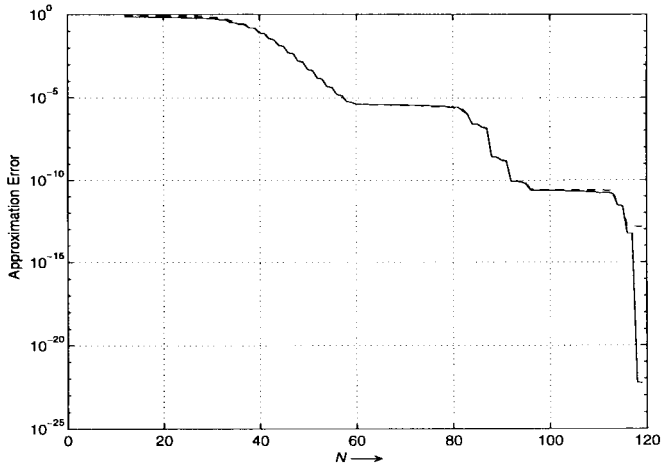

Fig. 9. Approximation of a linear-phase FIR bandpass filter ($L = 120$) by an IIR filter of order N. Approximation errors for $N = 12 \cdots 119$ (—— our algorithm; – – – the method of [4]).

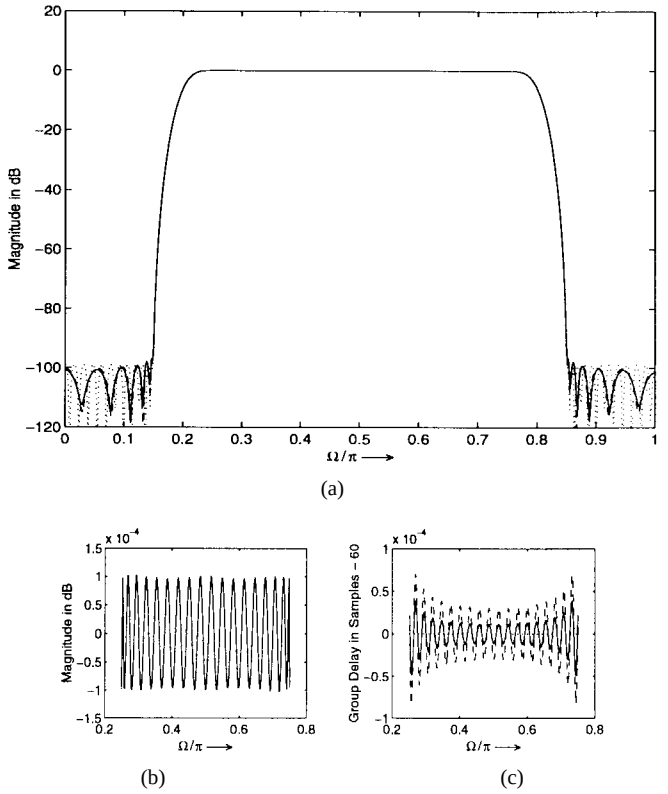

Fig. 10. Frequency responses of two IIR filters with $N = 60$ that approximate a linear-phase FIR bandpass filter ($L = 120$). (a) Magnitude frequency responses of the filters. (b) Magnitude frequency responses of the filters in the passband. (c) Group delays of the filters in the passband. ($\cdots\cdots$ FIR filter; —— IIR filter obtained with our algorithm; – – – IIR filter obtained with the method of [4]).

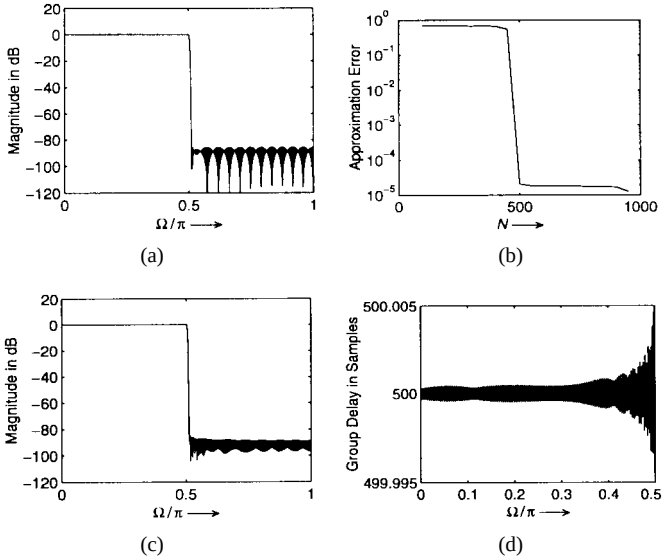

Fig. 11. Approximation of a linear-phase FIR lowpass filter ($L = 1000$) by an IIR filter of order N. (a) Magnitude frequency response of the FIR filter. (b) Approximation errors for $N = 100, 150, \ldots, 950$. (c) Magnitude frequency response of the approximating IIR filter with $N = 500$. (d) Group delay in the passband of the IIR filter with $N = 500$.

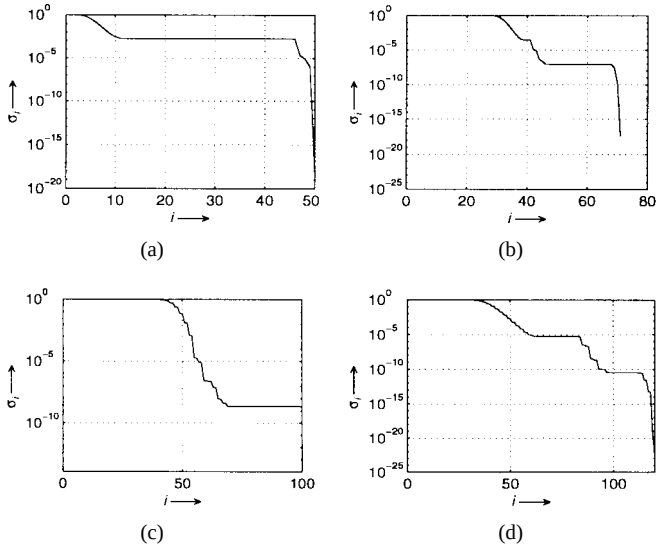

Fig. 12. Hankel singular values plots for the FIR filters of Examples 2–5. (a) Example 2. (b) Example 3. (c) Example 4. (d) Example 5.

VI. Conclusion

We have presented an algorithm for the approximation of an FIR filter by an IIR filter in the least-squares sense. A theorem was used to reformulate the approximation problem as a problem of finding an allpass filter with certain characteristics. A closed-form expression for the approximation error was given, and it was shown how the allpass filter coefficients can be obtained such that the approximation error is minimized. The minimization was achieved by an iterative procedure where each step required the solution of a set of linear equations and digital filtering operations. From the allpass filter, the interesting IIR filter could easily be constructed. All calculations were performed on the numerator and denominator coefficients of the transfer functions. No eigenvalue calculations and no conversions to and from state-space descriptions were necessary. As a consequence, the effects of finite-precision arithmetics on the results are negligible. This makes our algorithm applicable even in cases were the filter orders are large.

Our algorithm can also be employed for the approximation of an IIR filter $H_1(z)$ of order M by an IIR filter $H(z)$ of order

N ($N < M$). The following two-step procedure is proposed: First, an FIR filter $F(z)$ is constructed whose impulse response $f(n)$ satisfies $f(n) = h_1(n)$ ($n = 0, 1, \ldots, L$), where $h_1(n)$ is the impulse response of $H_1(z)$. Then, this FIR filter is approximated by an IIR filter $H(z)$ of order N. To ensure that $F(z)$ is close to $H_1(z)$, we have to choose L sufficiently large. However, approximating an FIR filter with large L is no problem with our algorithm, especially in this case, where $N \ll L$ is valid. Thus, it can be expected to obtain a good approximation.

APPENDIX

Proof of Theorem 2: We assume that for a given $X^{(k)}(z)$, the polynomial $Q^{(k)}(z)$ [see (30)] that minimizes $E_1^{(k)}$ [see (38)] has been obtained from (37) by solving (43). It is possible that the order of $Q^{(k)}(z)$ is lower than N, which means that $q_l^{(k)} \neq 0$ and $q_\nu^{(k)} = 0$ ($\nu = l+1, l+2, \ldots, N$) is valid, where $1 \leq l \leq N$. Note that the trivial case $Q^{(k)}(z) \equiv 1$, for which the theorem is obviously true, is excluded.

In the following, we omit the superscript "(k)" for convenience. Thus, we are concerned with the expressions $Q(z)$, $X(z)$, $U(z)$, and E_1 and suppose that (30), (35)–(38) are changed accordingly. The proof is accomplished by adopting the arguments in [10].

We consider the mirror-image polynomial of $Q(z)$

$$\tilde{Q}(z) := z^{-N} Q(z^{-1}) = \sum_{\nu=0}^{N} \tilde{q}_\nu z^{-\nu} \quad (\tilde{q}_N = 1). \tag{47}$$

To take into account the case, when the order of $Q(z)$ [see (30)] is lower than N, we assume

$$\tilde{q}_\nu = 0 \quad (\nu < m) \quad \text{and} \quad \tilde{q}_m \neq 0 \tag{48}$$

where m is restricted by

$$0 \leq m \leq N - 1. \tag{49}$$

Let $\zeta = r\mathrm{e}^{\mathrm{j}\varphi}$ ($r > 0$, $0 \leq \varphi < 2\pi$) be a zero of $\tilde{Q}(z)$. Then, $\tilde{Q}(z)$ can be written as

$$\tilde{Q}(z) = z^{-m}(z^{-1} - r^{-1}\mathrm{e}^{-\mathrm{j}\varphi})\bar{Q}(z). \tag{50}$$

With (47) and (50), we obtain from (37)

$$U(z) = z^{-m}(z^{-1} - r^{-1}\mathrm{e}^{-\mathrm{j}\varphi})S(z) \tag{51}$$

where

$$S(z) = \sum_{\mu=0}^{\infty} s_\mu z^{-\mu} \quad (s_0 \neq 0) \tag{52}$$

is a function with complex coefficients s_μ. It can be seen from (51) and (52) that the coefficients $u(n)$ of $U(z)$ [see (35)] vanish for $n < m$. As a consequence, the error norm E_1 [see (38)] can be calculated by

$$E_1 = \left[\sum_{n=m}^{L-1} |u(n)|^2 \right]^{\frac{1}{2}}. \tag{53}$$

We want to express E_1 in terms of r, φ, and s_μ. Substituting $S(z)$ in (51) by (52), the values $u(n)$ ($n = m, m+1, \ldots, L-1$) can be derived as

$$\begin{aligned} u(m) &= -r^{-1}\mathrm{e}^{-\mathrm{j}\varphi} s_0 \\ u(m+\mu) &= s_{\mu-1} - r^{-1}\mathrm{e}^{-\mathrm{j}\varphi} s_\mu \\ &\qquad (\mu = 1, 2, \ldots, L-1-m). \end{aligned} \tag{54}$$

Note that the $u(m+\mu)$ ($\mu = 0, 1, \ldots, L-1-m$) are real. With the abbreviation

$$p := L - 1 - m \tag{55}$$

we get from (53) and (54)

$$E_1^2 = r^{-2}|s_0|^2 + \sum_{\mu=1}^{p} \left| s_{\mu-1} - r^{-1}\mathrm{e}^{-\mathrm{j}\varphi} s_\mu \right|^2. \tag{56}$$

Now, r^{-1} in (56) is replaced by the parameter σ, which is allowed to vary in the following. Thus, we obtain a real function

$$E_1^2(\sigma) = \sigma^2|s_0|^2 + \sum_{\mu=1}^{p} \left| s_{\mu-1} - \sigma\mathrm{e}^{-\mathrm{j}\varphi} s_\mu \right|^2 \tag{57}$$

that depends on the real parameter σ. With the same substitution ($r^{-1} \to \sigma$) in (50), we can construct a one parameter family of polynomials $\tilde{Q}_\sigma(z)$. These polynomials can have complex coefficients. $\tilde{Q}(z)$ is a special case of these polynomials. It has real coefficients, and it is the one with the minimum value of E_1; see (53). As a consequence, $E_1^2(\sigma)$ attains its minimum value if and only if

$$\sigma = r^{-1}. \tag{58}$$

We can derive an alternative expression for the extremal point by using (57), which can be written as

$$E_1^2(\sigma) = a_2\sigma^2 - 2a_1\sigma + a_0 \tag{59}$$

where

$$\begin{aligned} a_0 &= \sum_{\mu=0}^{p-1} |s_\mu|^2 \\ a_1 &= \frac{1}{2} \sum_{\mu=1}^{p} \left(\mathrm{e}^{-\mathrm{j}\varphi} s_\mu s_{\mu-1}^* + \mathrm{e}^{\mathrm{j}\varphi} s_\mu^* s_{\mu-1} \right) \\ a_2 &= \sum_{\mu=0}^{p} |s_\mu|^2. \end{aligned} \tag{60}$$

The minimum is characterized by

$$\frac{\mathrm{d}E_1^2(\sigma)}{\mathrm{d}\sigma} = 2(a_2\sigma - a_1) = 0 \tag{61}$$

and we obtain

$$\frac{1}{r} = \sigma = \frac{a_1}{a_2}. \tag{62}$$

Note that a_1 must be positive since $r > 0$ and $a_2 > 0$, according to (60). We can give an estimation of a_1 by using

$$a_1 = \mathrm{Re} \sum_{\mu=1}^{p} \mathrm{e}^{-\mathrm{j}\varphi} s_\mu s_{\mu-1}^* \tag{63}$$

as

$$a_1 \le \left| \sum_{\mu=1}^{p} \mathrm{e}^{-\mathrm{j}\varphi} s_\mu s_{\mu-1}^* \right| \le \sum_{\mu=1}^{p} \left| \mathrm{e}^{-\mathrm{j}\varphi} s_\mu s_{\mu-1}^* \right| \le \sqrt{\sum_{\mu=1}^{p} |s_\mu|^2} \sqrt{\sum_{\mu=0}^{p-1} |s_\mu|^2} < \sum_{\mu=0}^{p} |s_\mu|^2 = a_2. \quad (64)$$

In summary

$$a_1 < a_2 \quad (65)$$

where the Cauchy–Schwarz inequality was used in the third step, and $s_0 \neq 0$ in the fourth step. With (62), we get

$$r > 1 \quad (66)$$

which means that $|\zeta| > 1$. This indicates that $\tilde{Q}(z)$ must have all its zeros in $|z| > 1$. As a consequence, the zeros of $Q(z)$ all lie in $|z| < 1$. □

REFERENCES

[1] D. H. Friedman, "On approximating an FIR filter using discrete orthonormal exponentials," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 923–926, Aug. 1981.

[2] A. I. Saleh and M. F. Fahmy, "Design of recursive digital filters through the use of discrete exponentials," in *Proc. Euro. Conf. Circuit Theory Design*, Sept. 1983, pp. 352–354.

[3] V. Sreeram and P. Agathoklis, "Design of linear-phase IIR filters via impulse-response gramians," *IEEE Trans. Signal Processing*, vol. 40, pp. 389–394, Feb. 1992.

[4] B. Beliczynski, J. Kale, and G. D. Cain, "Approximation of FIR by IIR digital filters: An algorithm based on balanced model reduction," *IEEE Trans. Signal Processing*, vol. 40, pp. 532–542, Mar. 1992.

[5] M. F. Fahmy, Y. M. Yassin, G. Abdel-Raheem, and N. El-Gayed, "Design of linear-phase IIR filters from FIR specifications," *IEEE Trans. Signal Processing*, vol. 42, pp. 437–440, Feb. 1994.

[6] S.-Y. Kung and D. W. Lin, "Optimal Hankel-norm model reductions: Multivariable systems," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 832–852, Aug. 1981.

[7] F. R. Gantmacher, *Matrizenrechnung I*. Berlin, Germany: VEB Deutscher Verlag der Wissenschaften, 1965.

[8] J. L. Walsh, *Interpolation and Approximation by Rational Functions in the Complex Domain*. Providence, RI: Amer. Math. Soc., 1965.

[9] J. H. McClellan, T. W. Parks, and L. R. Rabiner, "A computer program for designing optimum FIR linear phase digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 506–526, Dec. 1973.

[10] G. Schmeisser, D. Raghuramireddy, and R. Unbehauen, "An alternative proof for the stability of least squares inverse polynomials," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 822–824, Aug. 1986.

Hartmut Brandenstein (S'93) was born in Wonfurt, Germany, in 1965. He received the Dipl.-Ing. degree in electrical engineering from the University of Erlangen–Nürnberg, Erlangen, Germany, in 1992.

Since then, he has been a scientific assistant at the Lehrstuhl für Allgemeine und Theoretische Elektrotechnik, University of Erlangen-Nürnberg. His research interests are multidimensional digital signal processing, filter design, and multirate systems.

Rolf Unbehauen (F'91) received the diploma in mathematics, the Ph.D. degree in electrical engineering, and the qualification for inauguration as academic lecturer (Habilitation) from Stuttgart University, Stuttgart, Germany, in 1954, 1957, and 1964, respectively.

From 1955 to 1966, he was a member of the Institute of Mathematics, the Computer Center, and the Institute of Electrical Engineering, Stuttgart University, where he was appointed Associate Professor in 1965. Since 1966, he has been Full Professor of Electrical Engineering, University of Erlangen-Nürnberg, Erlangen, Germany. He has published a great number of papers on his research results. He has authored four books in German and is co-author of *MOS Switched-Capacitor and Continuous-Time Integrated Circuits and Systems* (Berlin, Germany: Springer, 1989) and *Neural Networks for Optimization and Signal Processing* (Berlin, Germany/New York: Teubner Verlag/Wiley, 1993). His teaching and research interests are in the areas of network theory and its applications, system theory, and electromagnetics.

Dr. Unbehauen was an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS from 1990 to 1991. Currently, he is Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS and of *Multidimensional Systems and Signal Processing*. In 1959, he received the NTG Best Paper Award, in 1991 the IEEE Fellow Award, in 1994 a Best Book Award and a honorary doctorate degree. He is a Member of the Informationstechnische Gesellschaft of Germany and of URSI, Commission C: Signals and Systems.