IIR Approximation of FIR Filters Via Discrete-Time Vector Fitting

Ngai Wong, *Member, IEEE*, and Chi-Un Lei, *Student Member, IEEE*

***Abstract*—We present a novel technique for approximating finite-impulse-response (FIR) filters with infinite-impulse-response (IIR) structures through extending the vector fitting (VF) algorithm, used extensively for continuous-time frequency-domain rational approximation, to its discrete-time counterpart called VFz. VFz directly computes the candidate filter poles and iteratively relocates them for progressively better approximation. Each VFz iteration consists of the solutions of an overdetermined linear equation and an eigenvalue problem, with real-domain arithmetic to accommodate complex poles. Pole flipping and maximum pole radius constraint guarantee stability and robustness against finite-precision implementation. Comparison against existing algorithms confirms that VFz generally exhibits fast convergence and produces highly accurate IIR approximants.**

***Index Terms*—Approximation algorithm, finite-impulse-response (FIR) filters, infinite-impulse-response (IIR) filters, vector fitting.**

I. Introduction

Vector fitting (VF) [2], [3], since its introduction in 1999, has become a popular technique for fitting calculated or measured frequency-dependent vector/matrix data with rational function approximation. Application examples include power system and transmission line modelings, electromagnetic simulation, and lately in VLSI package and high-speed interconnect simulations [4], [5]. However, the use of VF has been limited to the Laplace domain (s-domain), and that the fitting thus obtained is primarily employed for model identification.

On the other hand, in digital filter design, recent research has been drawn to the infinite-impulse-response (IIR) approximation of finite-impulse-response (FIR) filters, e.g., [6]–[11]. This is motivated by 1) IIR design methodology through matching to a prescribed FIR filter prototype and 2) possible hardware savings due to the fewer multipliers in IIR structures. Representative IIR approximation algorithms include state-space model reduction and least-squares (LS) approximation [6]–[10], etc. However, in the approximation exercise, stringent constraints like accurate magnitude and phase matching, stability, and low algorithmic complexity have to be satisfied. Owing to the highly nonlinear nature of the problem, so far there is no optimal algorithm in terms of accuracy and computational cost [11].

The correspondence generalizes VF to its discrete-time or z-domain counterpart, called VFz. Instead of the conventional use in model identification, VFz is adapted to filter design, and in particular, the IIR approximation of FIR filters. Analogous to VF, the core of VFz is a two-step process for refining the filter poles such that the desired response may be accurately reproduced with usually low-order rational functions. VFz enjoys simple coding and is numerically well-conditioned with its use of partial fractions, instead of power series, as basis functions. Unstable poles undergo reciprocal flipping; thus, stability is always guaranteed. The maximum pole radius can also be readily constrained for finite wordlength consideration. Numerical examples then confirm the remarkable efficiency and accuracy of VFz over conventional IIR approximation algorithms.

Manuscript received August 30, 2006; revised July 30, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Timothy N. Davidson. A preliminary version of this work appeared in *Proceedings of the IEEE International Symposium on Circuits and Systems* (ISCAS), May 27–30, 2007, pp. 2343–2346.

The authors are with the Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong (e-mail: nwong@eee.hku.hk; culei@eee.hku.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2007.908935

II. Vector Fitting

VF [2] attempts to fit the rational function

$$\left(\sum_{n=1}^{N}\frac{c_n}{s-a_n}\right)+d+\mathrm{se}$$

to a set of calculated/sampled data points $f(s_k)$'s at frequencies $\{s_k\}, k=1,2,\ldots,N_s$. The poles a_n and residues c_n are either real or complex conjugate pairs, and d and e are real. Starting with a set of prescribed or approximated poles $\{\alpha_n^{(0)}\}, n=1,2,\ldots,N$, and by introducing the scaling function $\sigma(s)$, a linear problem is set up for the ith iteration, namely

$$\underbrace{\left(\sum_{n=1}^{N}\frac{c_n}{s-\alpha_n^{(i)}}\right)+d+se}_{(\sigma f)(s)}\approx\underbrace{\left(\left(\sum_{n=1}^{N}\frac{\gamma_n}{s-\alpha_n^{(i)}}\right)+1\right)}_{\sigma(s)}f(s),\qquad(1)$$

$i=0,1,\ldots,N_T$, where N_T denotes the number of iterations when convergence is attained or when the upper limit is reached. The unknowns, c_n, d, e, and γ_n, are solved through an overdetermined linear equation formed by evaluating (1) at the N_s sampled frequency points. It can be observed that (1) constrains $(\sigma f)(s)$ and $\sigma(s)f(s)$ to share the same poles, which in turn implies that the original poles of $f(s)$ are canceled by the zeros of $\sigma(s)$. Solving the zeros of $\sigma(s)$ therefore produces, in the LS sense, an approximation to the poles of $f(s)$, viz. $\{\alpha_n^{(i+1)}\}$, which are then fed back to (1) as the next set of known poles. Any unstable pole is flipped about the imaginary axis to the open left half plane for stability. Upon convergence, the update in $\{\alpha_n^{(i)}\}$ diminishes and $\sigma(s)\approx 1$.

Subsequently, VF represents a two-step process: constructing $\sigma(s)$ and computing its zeros, such that the underlying (stable) system poles are successively approximated. Linear equation solves and eigenvalue solves are used exclusively in VF. Furthermore, VF is applicable to fitting vectors by replacing c_n, d, and e in (1), and hence $f(s)$, by column vectors. In that case, all entries of the fitted vector share a common set of poles $\{a_n\} := \{\alpha_n^{(N_T)}\}$(":=" denotes assignment). Recently, [12] has recognized VF to be a special case of the Sanathanan–Koerner (SK) iteration [13], but with a well-conditioned pole-based basis as compared to the conventional power-series implementations.

III. Discrete-Time Vector Fitting

Our design goal is to approximate the FIR digital filter

$$f(z)=\sum_{n=0}^{L}h_n z^{-n}\quad\text{where } h_n\in\mathbb{R}, h_L\neq 0\qquad(2)$$

with a causal and stable IIR filter

$$\hat{f}(z)=\frac{P(z)}{Q(z)}=\frac{\sum_{\mu=0}^{M}p_\mu z^{-\mu}}{\sum_{\upsilon=0}^{N}q_\upsilon z^{-\upsilon}}\quad\text{where } p_\mu, q_\upsilon\in\mathbb{R}, q_0=1.\qquad(3)$$

Therefore, all poles of $\hat{f}(z)$ (zeros of $Q(z)$) must lie in $|z|<1$. Using $\bar{(\circ)}$ to denote complex conjugate operation, obviously, $\overline{f}(e^{j\Omega})=f(e^{-j\Omega})$ and $\overline{\hat{f}}(e^{j\Omega})=\hat{f}(e^{-j\Omega}), \forall\Omega\in[-\pi,\pi)$. To exclude the trivial case of $\hat{f}(z)\equiv f(z)$, we assume $M<L$. In the following, we formulate VFz as the discrete-time counterpart of VF and adopt it to IIR filter design.

A. Pole Relocation and Stabilization

Similar to VF, we use partial fraction basis to seek a rational approximation of the FIR filter $f(z)$ in (2). This is done by equating (approximating) it to the IIR filter $\hat{f}(z)$ in (3), namely

$$\hat{f}(z) = \left(\sum_{n=1}^{N} \frac{c_n}{z^{-1} - a_n} \right) + d \approx f(z) = \sum_{n=0}^{L} h_n z^{-n}, \quad (4)$$

over the (digital) frequency band(s) of interest. Similarly, c_n and a_n are either real or complex conjugate pairs. We note that in (1) the "se" term is included for a generic continuous-time *passive* transfer function which is not needed in the digital filter regime. To ensure stability, the set of poles $\{1/a_n\}$ in (4) must be within the unit circle and therefore $|a_n| > 1$.As in (1), supposing an initial set of pole reciprocals $\{\alpha_n^{(0)}\}, |\alpha_n^{(0)}| > 1$, is specified, we build

$$\underbrace{\left(\sum_{n=1}^{N} \frac{c_n}{z^{-1} - \alpha_n^{(i)}} \right) + d}_{(\sigma f)(z)} \approx \underbrace{\left(\left(\sum_{n=1}^{N} \frac{\gamma_n}{z^{-1} - \alpha_n^{(i)}} \right) + 1 \right)}_{\sigma(z)} f(z), \quad (5)$$

$i = 0, 1, \ldots, N_T$. Ambiguity in the solution for $\sigma(z)$ is removed by matching it to unity as z approaches the origin. Now (5) is linear in its unknowns c_n, d, and γ_n. Writing (5) for the N_s frequency points $z_k = e^{j\Omega_k}, \Omega_k \in [0, \pi), k = 1, 2, \ldots, N_s, N_s > 2N + 1$, gives an overdetermined linear problem. Specifically, rewriting (5) at $z = z_k$

$$\left(\sum_{n=1}^{N} \frac{c_n}{z_k^{-1} - \alpha_n^{(i)}} \right) + d - \left(\sum_{n=1}^{N} \frac{\gamma_n f(z_k)}{z_k^{-1} - \alpha_n^{(i)}} \right) \approx f(z_k) \quad (6)$$

which can be put into

$$A_k x = b_k \quad (7)$$

where

$$\begin{aligned} A_k &= \left[\frac{1}{z_k^{-1} - \alpha_1^{(i)}} \quad \cdots \quad \frac{1}{z_k^{-1} - \alpha_N^{(i)}} \quad 1 \quad \frac{-f(z_k)}{z_k^{-1} - \alpha_1^{(i)}} \quad \cdots \quad \frac{-f(z_k)}{z_k^{-1} - \alpha_N^{(i)}} \right] \\ x &= [\, c_1 \quad \cdots \quad c_N \quad d \quad \gamma_1 \quad \cdots \quad \gamma_N \,]^T, \quad b_k = f(z_k). \end{aligned} \quad (8)$$

Here, A_k and x are row and column vectors, respectively, and b_k is a scalar. Repeating (7) at the N_s frequency points and stacking the A_k's and b_k's into a (tall) column matrix and a vector, respectively, gives the overdetermined linear equation for each i, namely

$$Ax = b. \quad (9)$$

Real arithmetic is preferred in actual computation. When all poles in (6) are real, and therefore real $\alpha_n^{(i)}$'s, x is a real vector while A and b are complex. Accordingly, (9) is solved in the real domain by

$$\begin{bmatrix} \Re A \\ \Im A \end{bmatrix} x = \begin{bmatrix} \Re b \\ \Im b \end{bmatrix} \quad (10)$$

where $\Re$ and $\Im$ denote the real and imaginary parts, respectively. Using the last N elements of the LS solve of x, i.e., γ_1 to γ_N, $\sigma(z)$ in (5) can be reconstructed whose zeros, $\{1/\alpha_n^{(i+1)}\}$, then form the new set of starting poles in the next VFz iteration. Similar to the VF analysis [2], it can be shown that the reciprocals of zeros of $\sigma(z)$, $\{\alpha_n^{(i+1)}\}$, are conveniently obtained as the eigenvalues of

$$\Psi = \begin{bmatrix} \alpha_1^{(i)} & & & \\ & \alpha_2^{(i)} & & \\ & & \ddots & \\ & & & \alpha_N^{(i)} \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} [\, \gamma_1 \quad \gamma_2 \quad \cdots \quad \gamma_N \,]. \quad (11)$$

Ψ is real when only real poles are considered. To ensure stability, it is required that every $|\alpha_n^{(i+1)}| > 1$. If it is not so, its reciprocal is taken, viz. $\alpha_n^{(i+1)} := 1/\alpha_n^{(i+1)}$, such that the pole is flipped back inside the unit circle. This has the physical meaning of multiplying the all-pass filter

$$\frac{z^{-1} - \alpha_n^{(i+1)}}{1 - \bar{\alpha}_n^{(i+1)} z^{-1}}$$

to both sides of (5) (now with $i := i + 1$), thus changing the phase without altering the magnitude response. Here, a real $\alpha_n^{(i+1)}$ is assumed but flipping of conjugate poles follows exactly by multiplying two all-pass filters, with conjugate poles, at a time. Such stability enforcement parallels the pole flipping strategy in the s-domain VF. So far, only real poles are considered. Special care must be paid to complex conjugate poles.

B. Complex Poles

The poles in (5) or (6) must be either real or complex conjugates such that $\hat{f}(z)$ stays as a real-coefficient approximant to the original $f(z)$ whose tap coefficients are real. In case of complex conjugate pole pairs in (6), i.e., $\alpha_{n+1}^{(i)} = \bar{\alpha}_n^{(i)}$, the c_n's and γ_n's in x must also be in conjugate pairs. To maintain a real x, the corresponding entries in A_k and x need to be modified. Without loss of generality, suppose the first two poles in A_k are conjugate pair, i.e., $\alpha_2^{(i)} = \bar{\alpha}_1^{(i)}$. With reference to (7) and (8), we have

$$\begin{aligned} A_k &= \left[\frac{1}{z_k^{-1} - \alpha_1^{(i)}} \quad \frac{1}{z_k^{-1} - \bar{\alpha}_1^{(i)}} \quad \cdots \right] \\ x &= [\, c_1 \quad \bar{c}_1 \quad \cdots \,]^T. \end{aligned} \quad (12)$$

To maintain a real x, (12) is rewritten as

$$\begin{aligned} A_k &= \left[\left(\frac{1}{z_k^{-1} - \alpha_1^{(i)}} + \frac{1}{z_k^{-1} - \bar{\alpha}_1^{(i)}} \right) \quad j \left(\frac{1}{z_k^{-1} - \alpha_1^{(i)}} - \frac{1}{z_k^{-1} - \bar{\alpha}_1^{(i)}} \right) \quad \cdots \right] \\ x &= [\, \Re c_1 \quad \Im c_1 \quad \cdots \,]^T. \end{aligned} \quad (13)$$

Modification for the last N entries in A_k and x (i.e., γ_1 to γ_N) for conjugate pole pairs follows similarly [see (14), shown at the bottom of the page]. Subsequently, (10) is formulated and solved with these new expressions of A_k's and x.

To compute the zeros of $\sigma(z)$, which now contains complex poles, we apply similarity transform to (11) to bring it back to a real matrix.

$$\begin{aligned} A_k &= \left[\cdots \quad -\left(\frac{1}{z_k^{-1} - \alpha_1^{(i)}} + \frac{1}{z_k^{-1} - \bar{\alpha}_1^{(i)}} \right) f(z_k) \quad -j \left(\frac{1}{z_k^{-1} - \alpha_1^{(i)}} - \frac{1}{z_k^{-1} - \bar{\alpha}_1^{(i)}} \right) f(z_k) \quad \cdots \right] \\ x &= [\, \cdots \quad \Re\gamma_1 \quad \Im\gamma_1 \quad \cdots \,]^T. \end{aligned} \quad (14)$$

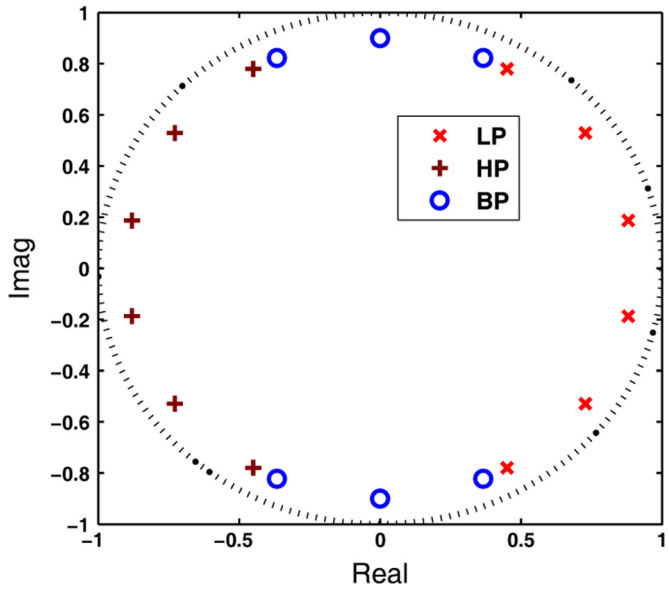

Fig. 1. Typical initial pole placements in VFz for the six-pole IIR approximations of general lowpass (LP), bandpass (BP), and highpass (HP) filters. All poles are sitting on a circle with radius$= 0.9$.

Each pair of conjugate poles now manifest as a 2×2 diagonal block in Ψ. Specifically, using the notions in (14), Ψ is transformed into

$$\Psi = \begin{bmatrix} \Re\alpha_1^{(i)} & \Im\alpha_1^{(i)} & & \\ -\Im\alpha_1^{(i)} & \Re\alpha_1^{(i)} & & \\ & & \ddots & \end{bmatrix} - \begin{bmatrix} 2 \\ 0 \\ \vdots \end{bmatrix} [\,\Re\gamma_1 \quad \Im\gamma_1 \quad \cdots\,]. \tag{15}$$

Apart from accelerating convergence, allowing complex poles in the (real) VFz arithmetics is critical, if not necessary, as all practical digital filters have poles distributed in certain sectors in the complex plane. For example, Fig. 1 shows the typical initial pole placements for general lowpass (LP), bandpass (BP), and highpass (HP) filters.

C. Building the IIR Filter

Suppose a converged set of filter poles (or their reciprocals $\{\alpha_n^{(N_T)}\}$) are obtained, the final step is to reconstruct the IIR filter $\hat{f}(z)$. With reference to (5) and (6), we should now have $\sigma(z) \approx 1$ and the following relationship holds:

$$\hat{f}(z_k) = \left(\sum_{n=1}^{N} \frac{c_n}{z_k^{-1} - \alpha_n^{(N_T)}} \right) + d \approx f(z_k), \tag{16}$$

$k = 1, 2, \ldots, N_s$. The residues c_n of $\hat{f}(z)$ are computed in exactly the same manner as in the previous two sections, except that the last N elements in both A_k and x are now discarded. This partial fraction decomposition of $\hat{f}(z)$ may then be summed up to a rational function commonly used in IIR filter representation.

The computation of VFz lies in its two major steps: the overdetermined equation solve in (10) requires $O(N^2 N_s)$ operations, and the eigenvalue solve in (11) or (15) requires $O(N^3)$ operations. As will be discussed later, we usually choose $N_s \approx L \approx 4N$ and N_T is consistently within 10–15, therefore overall VFz constitutes an $O(N^3)$ algorithm.

D. Filter Stability and Finite Wordlength Consideration

Pole stability is not necessarily guaranteed in some IIR approximation algorithms (e.g., see [11]). Explicit pole computation in VFz, however, allows simple reciprocal flipping of unstable poles whose relationship to all-pass filter multiplication has been described in Section III-A. The multiplication of all-pass filters, while preserving the magnitude, always alters the phase response. The effect of this phase change, however, is then offset/suppressed through the solution of (10) which seeks the residue and dc coefficient update, under the new set of (stable) poles, that provides the LS fit to the objective FIR frequency samples $\{f(z_k)\}$ in terms of both magnitude and phase. Because the residues are related to the numerator and therefore zeros of the rational function approximant, this step is also known as the numerator update or zero placement.

In hardware implementation, not only the stability of the ideal filter but also the relative stability due to coefficient quantization has to be considered. Specifically, finite wordlength (FWL) dictates that only a finite and usually nonuniform constellation of poles on the unit circle can be realized [14]. To impose a FWL stability margin [15], we adopt the common, if ad hoc, scheme by restricting the maximum pole radius to be inside a bound smaller than unity. Again, with the explicit poles obtained from (11) or (15), this is simply done by a scaling of any violating pole.

In our experiments in Section V, pole flipping always occurs during the first few (~ 5) iterations before the poles eventually settle to a relatively fixed region in the unit circle. The effects of maximum pole radius and FWL are also studied in Example 5. In any case, convergence is maintained despite the fact that a few more iterations may be incurred.[1] Such excellent convergence is attributed to the connection of VFz to the Steiglitz–McBride (SM) Iteration, as discussed below.

IV. VFZ as Steiglitz–McBride Iteration

Analogous to the equivalence between VF and SK iteration [13], [12], VFz can be regarded as a reformulation of the rational function fitting procedure called SM iteration [17]. Using the notations from Section III, given a transfer function or response $f(z)$, SM iteration replaces the nonlinear LS approximation objective $\widehat{G}_{L_2} = \Sigma_{k=1}^{N_s} |f(z_k) - (P(z_k))/(Q(z_k))|^2$ with a linearized $\widehat{G}_{\mathrm{SM}}$ where

$$\widehat{G}_{\mathrm{SM}} = \sum_{k=1}^{N_s} \frac{1}{|Q^{(i-1)}(z_k)|^2} \left| Q^{(i)}(z_k) f(z_k) - P^{(i)}(z_k) \right|^2. \tag{17}$$

Here, $P^{(i)}$ and $Q^{(i)}$ are, respectively, the numerator and denominator determined during the ith SM iteration (thus $Q^{(i-1)}$ is assumed predetermined). Although $\widehat{G}_{\mathrm{SM}}$ is not equivalent to $\widehat{G}_{L_2}$, by using the triangle inequality, if we approximate $f(z)$ by an Nth-order system, $\|\widehat{G}_{L_2} - \widehat{G}_{\mathrm{SM}}\|_2 \leq 2\sigma_{N+1}$, where σ_i denotes the ith singular value of a Hankel-form matrix constructed by the coefficients h_n's of $f(z)$, whose order $L \gg N$ [18] and σ_i measures the significance of the ith approximant order. In general, SM iteration converges to a near-global-optimal approximant as in the LS sense for noise-free data, with an *a priori* error bound for an Nth-order approximant

$$\min_{\deg(\frac{P}{Q}=N)} \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \left| f(e^{j\omega}) - \frac{P^{(i)}(e^{j\omega})}{Q^{(i)}(e^{j\omega})} \right|^2 d\omega \right)^{1/2} \leq \sigma_{N+1}. \tag{18}$$

Such error bound is important as it provides a certificate for the approximant accuracy and can be used to select the approximant order. For example, from our extensive experience with VFz and the Hankel-form matrices thus analyzed, a rule of thumb is to set $N \approx L/4$ for the number of IIR poles to obtain a good (viz. $\sigma_{N+1} \approx 0$) approximation. SM iteration has been used in digital filter design [19], [20], but the

[1]We note that, like VF, convergence of VFz can be destroyed by noisy frequency samples so robust schemes should be used to restore it [16]. In our context, the samples $\{f(z_k)\}$ arising from the predetermined FIR prototype are always noise free so we refrain from further elaboration.

TABLE I
FILTER SPECIFICATIONS AND RESULTS

Filter	FIR Taps	Filter Type [Passband/π]	IIR Poles	CPU Time (sec) LS [8]	VFz	BMR [7]
Ex. 1	80	LP [0, 0.3]	20	0.2392	0.4707	0.9357
Ex. 2	120	BP [0.4, 0.6]	30	0.2293	0.3906	1.7946

corresponding formulations require complicated linearization tradeoff and additional optimization constraints. This is in contrast to the simple codings of VFz as well as its use of numerically well-conditioned partial fraction basis that eliminates high powers of $z = e^{j\omega}$ in a direct SM implementation.

V. NUMERICAL EXAMPLES

The balanced model reduction (BMR) [7], [9] method is known to produce good IIR approximants, but is restricted by its fast growth in algorithmic complexity with respect to the FIR filter order. In [11], various non-model-reduction, essentially iterative, algorithms are examined, namely, the (weighted) least-squares (LS), least-squares inverse (LSI), FIR fitting, and mixed-domain fitting. Among these, it has been found that the LS scheme often produces the most accurate IIR approximants. Subsequently, we contrast VFz against BMR [7] and LS [8] schemes for its accuracy and complexity. All experiments are done in the Matlab 7.2 environment using a 512 MB-RAM 1.4-GHz laptop. All algorithms (VFz, LS, and BMR) are coded in standard m-script files. The FIR prototypes are designed with the Matlab routine `fir1` using the window method.

Example 1: An 80-tap FIR linear-phase LP filter is reduced to a 20-pole IIR filter (recall our $N \approx L/4$ heuristics from Section IV). The numerator and denominator in each IIR filter are of the same order. The filter specification and CPU times of respective algorithms are listed in Table I. Specifically, VFz uses 100 linearly spaced sampling points in the passband and transition band, and 30 in the stopband, and is run for 15 iterations, which is more than sufficient. The initial poles follow the typical LP distribution as in Fig. 1 by uniformly distributing 20 poles with a radius of 0.9 in the angle range $\pm 0.8\pi/2$. For fairness, the number of iterations in LS equals that in VFz. From Table I, the iterative schemes, LS and VFz (both of $O(N^3)$ complexity), exhibit much higher efficiency than BMR (of $O(L^3)$ complexity). The time of VFz is slightly higher than that of LS mainly because of the eigenvalue solve, which is not needed in the latter. Fig. 2 and Table II show the magnitude and phase (represented by the group delay) responses, as well as the approximation errors. Obviously, all schemes produce satisfactory results but among them, VFz always renders the best approximants, seen graphically and verified numerically (cf. Table II where passband accuracy has a higher importance in filter design). In other words, VFz achieves similar computational efficiency to LS, and comparable or even better accuracy than BMR.

To demonstrate the robustness and insensitivity of VFz against initial pole placement, the starting poles are now assigned in the opposite HP region as in Fig. 3. The "self-correcting" pole relocation can be seen which reaches convergence after seven iterations, eventually giving the same results as with LP starting poles (for which the convergence is attained after only four iterations).

Example 2: A 120-tap FIR linear-phase BP filter is reduced to a 30-pole IIR filter. Again, the filter specification and CPU times are in Table I. This time, we reduce the number of VFz sampling points and iterations. Specifically, VFz uses 80 linearly spaced sampling points in the passband and transition band, and 20 in the stopband, and is run for

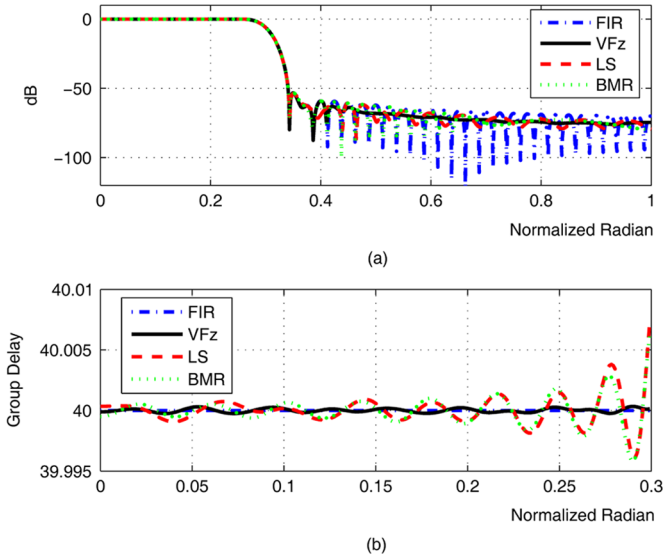

Fig. 2. Frequency responses of the LP filters in Example 1: (a) magnitude responses and (b) passband group delays.

TABLE II
ABSOLUTE APPROXIMATION ERRORS IN EXAMPLES 1 AND 2. I: PASSBAND MAGNITUDE, II: PASSBAND GROUP DELAY, III: STOPBAND MAGNITUDE. THE "WINNER" IN EACH CATEGORY IS UNDERLINED

	L_2 error			L_∞ error		
Filter	VFz	LS	BMR	VFz	LS	BMR
Ex. 1 (I)	6.8e-5	2.5e-4	2.5e-4	1.0e-5	4.1e-5	3.7e-5
(II)	0.0027	0.0207	0.0199	3.4e-4	7.3e-3	6.4e-3
(III)	0.0043	0.0048	0.0026	5.0e-4	7.4e-4	3.1e-4
Ex. 2 (I)	0.0012	0.0019	0.0017	9.7e-5	1.6e-4	1.7e-4
(II)	0.1031	0.1717	0.1633	0.0208	0.0391	0.0413
(III)	0.0075	0.0082	0.0079	6.5e-4	6.8e-4	6.4e-4

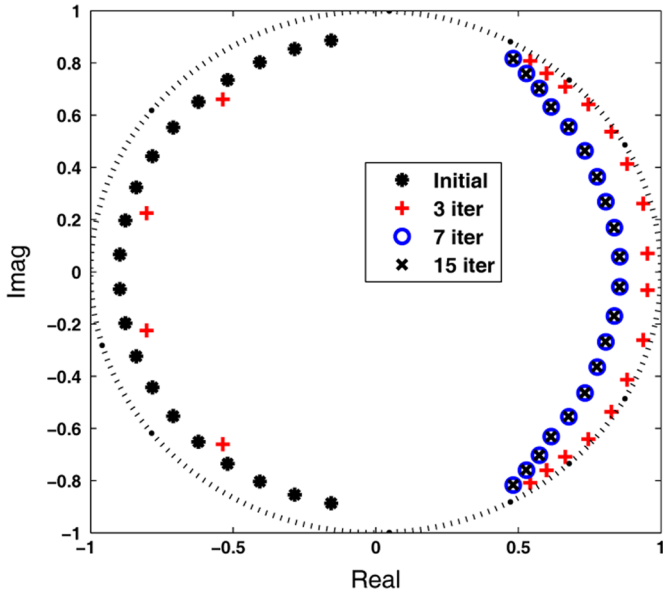

Fig. 3. VFz pole relocation of the LP filter in Example 1 using HP initial poles: at the 3rd, 7th and 15th iterations.

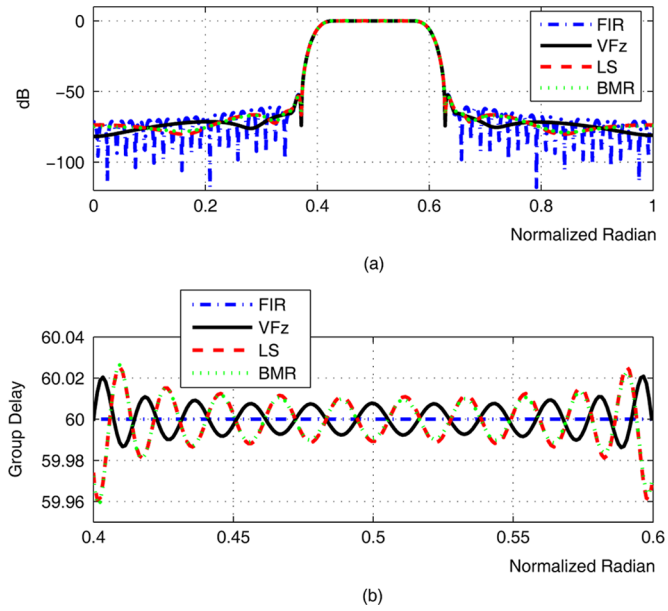

Fig. 4. Frequency responses of the BP filters in Example 2: (a) magnitude responses and (b) passband group delays.

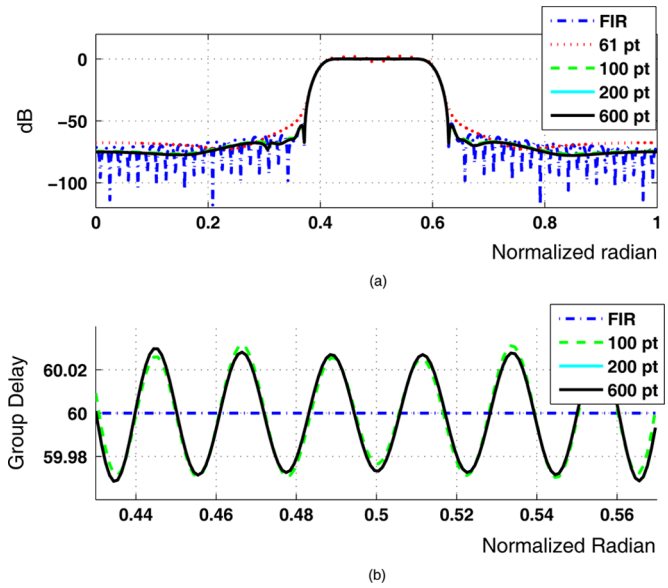

Fig. 5. Frequency responses of IIR approximants by VFz for different numbers of sampling points (iterations fixed at 8): (a) magnitude response and (b) passband group delays.

eight iterations. Responses and errors are shown in Fig. 4 and Table II. Similar observations as in Example 1 are obtained.

Example 3: The influence of the number of frequency sampling points on VFz is investigated. The filter in Example 2 is used. This time, all the sampling points are linearly spaced over $[0, \pi)$, and VFz is run for eight iterations. Fig. 5 and Table III show the results. The group delay curve of the 61-point case is omitted since it does not give an accurate result. Obviously, to enable solution of (9) [therefore (10)], at least $2N + 1$ (N: IIR poles) points are needed. It can be observed in Fig. 5 that in general more sampling points give rise to more accurate approximants, but the gain quickly tapers off beyond a certain number (100 in this test). Excessive sampling points do not improve accuracy much but increase the CPU time. From our experience in various tests,

TABLE III
CPU TIMES OF VFZ UNDER DIFFERENT NUMBERS OF SAMPLING POINTS (ITERATIONS FIXED AT 8)

No. of pt	61	100	200	600
Time (sec)	0.1875	0.2031	0.3906	1.9688

TABLE IV
CPU TIMES OF VFZ UNDER DIFFERENT NUMBERS OF ITERATIONS (SAMPLING POINTS FIXED AT 100)

No. of Iter	1	5	8	16
Time (sec)	0.0469	0.1406	0.2031	0.4375

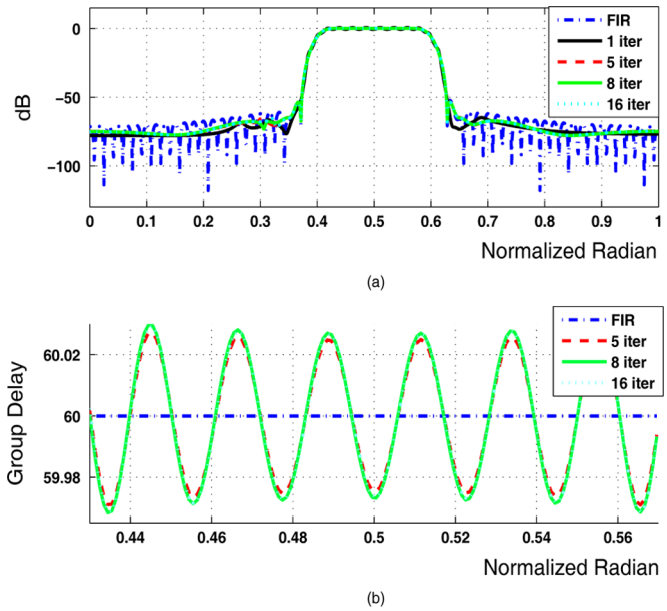

Fig. 6. Frequency responses of IIR approximants by VFz for different numbers of iterations (sampling points fixed at 100): (a) magnitude response and (b) passband group delays.

a simple rule of thumb is to set the number of sampling points N_s to be approximately equal to the FIR filter order L. Furthermore, an intuitive weighting scheme, which proves to work remarkably well as demonstrated in Examples 1 & 2, is to assign more points in the passband and transition band, and fewer in the stopband, thus resulting in better IIR shaping at the frequencies of higher importance.

Example 4: The influence of the number of iterations on VFz is investigated. The filter in Example 2 is used. Now we fix the number of sampling points to be 100 which are linearly spaced over $[0, \pi)$. Fig. 6 and Table IV show the results. The group delay curve of the one-iteration case is omitted since it does not give an accurate result. It can be observed in Fig. 6 that in general more iterations give rise to more accurate approximants, but the gain quickly tapers off beyond a certain number (as few as five in this test). Excessive iterations do not improve accuracy much but increase the CPU time. Moreover, if we put 80 linearly spaced sampling points in the passband and transition band, and 20 in the stopband, the passband ripples in Fig. 6 in the one-iteration curve will smoothen out. This further verifies the help of sensible sampling point allocation in enhancing the VFz convergence as well as accuracy. In practice, the number of VFz iterations required is filter- and initial-pole-dependent so the terminating condition should be one that checks for smaller relative pole update than a preset tolerance. We

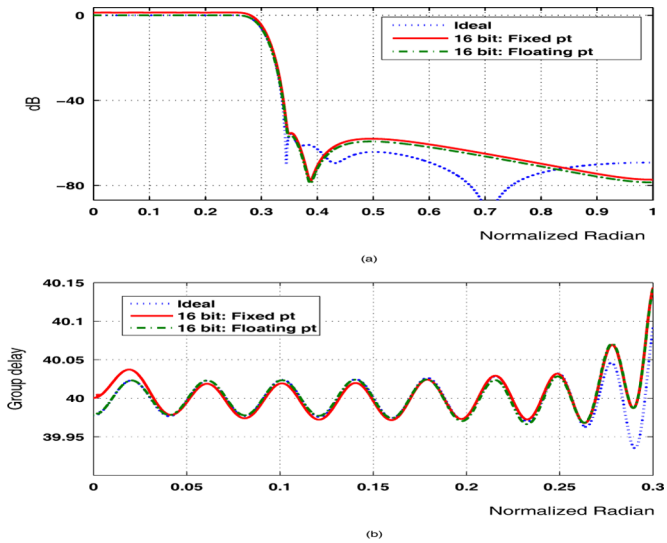

Fig. 7. Frequency responses of IIR approximants by VFz with pole radius constraint: (a) magnitude responses and (b) passband group delays.

have implemented this check and found that in all our experiments (including many not reported here), VFz consistently converges in 5–15 iterations with high robustness against distributions of initial poles and frequency sampling points. The practically bounded iterations in VFz thereby does not add to the $O(N^3)$ algorithmic complexity.

Example 5: The effects of maximum pole radius constraint and coefficient quantization are investigated (cf. Section III-D). The LP filter in Example 1 is used. A pole radius bound of 0.925 is imposed. It is found that VFz converges in only five iterations, and the maximum pole radius of the IIR approximant equals 0.925. The IIR filter coefficients are then quantized and realized using stable second-order sections. The frequency responses of the ideal, 16-bit floating- and fixed-point structures are compared in Fig. 7. It is shown that despite a minor degradation in the stopband attenuation, the passband group delays in the quantized cases are only slightly affected.

VI. Remarks

1) An interesting observation is that VFz fitting of the desired IIR response is consistently more accurate in the passband in both magnitude and phase. This can be explained by investigating the mechanism in solving the overdetermined linear (9) or (10). Take, for instance, the LP filter in Example 1, the rows A_k (in A) and b_k (in b) corresponding to the passband z_k's have bigger norms than those falling outside the passband. The LS solve of this system of equations then results in an automatic weighting and produces better "resolution" in the passband approximation.
2) In most non-model-reduction schemes, the determination of the poles, i.e., finding the denominator of $\hat{f}(z)$, constitutes the most important and difficult part in the algorithm [8], [11]. Because the strength of VFz lies in its *deterministic* refinement and explicit handling of poles, VFz is expected to outperform competing algorithms in terms of approximation accuracy. This has also been verified in our numerical examples.
3) VFz is simple in coding and concept since it is merely based on algebraic fitting of the prototype response with stability guarantee. As seen in the numerical examples, VFz exhibits high computational efficiency as in the iterative LS scheme, and comparable or even better optimality than the BMR approach (cf. Table II for the L_2 and L_∞ IIR approximation errors versus the FIR prototype). From Section IV, the convergence of VFz tracks that of SM iteration. To summarize the recommended settings for VFz in the IIR approximation exercise, we may choose $N \approx L/4$ with standard initial pole placement as in Fig. 1, $N_s \approx L$ and with $(z_k, f(z_k))(k = 1, \ldots, N_s)$ evenly distributed across $[0, \pi)$ or concentrated towards the passband and transition band, and then run VFz for 10–15 iterations or until the pole update is negligible.
4) The weighted LS version of VFz, paralleling that in the s-domain VF [3], can be formally developed. However, this is beyond the central theme of this correspondence and is not further elaborated.

VII. Conclusion

This correspondence has generalized the VF algorithm to its discrete-time counterpart called VFz. The novel application of VFz in IIR approximation of FIR filters has been investigated in depth. Starting with a set of prescribed initial poles, VFz uses linear solves and eigenvalue computations to iteratively relocate the poles for improved approximation. Modification of VFz to allow complex poles and the effects of pole flipping and finite wordlength consideration have been described. Numerical examples have confirmed that IIR approximation by VFz exhibits fast convergence and excellent accuracy in terms of both magnitude and phase.

References

[1] N. Wong and C. U. Lei, "FIR filter approximation by IIR filters based on discrete-time vector fitting," in *Proc. IEEE Symp. Circuits and Systems*, May 27–30, 2007, pp. 2343–2346.
[2] B. Gustavsen and A. Semlyen, "Rational approximation of frequency domain responses by vector fitting," *IEEE Trans. Power Del.*, vol. 14, no. 3, pp. 1052–1061, Jul. 1999.
[3] B. Gustavsen, "Computer code for rational approximation of frequency dependent admittance matrices," *IEEE Trans. Power Del.*, vol. 17, no. 4, pp. 1093–1098, Oct. 2002.
[4] B. Gustavsen and A. Semlyen, "Simulation of transmission line transients using vector fitting and modal decomposition," *IEEE Trans. Power Del.*, vol. 13, no. 2, pp. 605–614, Apr. 1998.
[5] D. Ioan, G. Ciuprina, M. Radulescu, and E. Seebacher, "Compact modeling and fast simulation of on-chip interconnect lines," *IEEE Trans. Magn.*, vol. 42, no. 4, pp. 547–550, Apr. 2006.
[6] A. Betser and E. Zeheb, "Reduced order IIR approximation to FIR digital filters," *IEEE Trans. Signal Process.*, vol. 39, no. 11, pp. 2540–2544, Nov. 1991.
[7] B. Beliczynski, I. Kale, and G. D. Cain, "Approximation of FIR by IIR digital filters: An algorithm based on balanced model reduction," *IEEE Trans. Signal Process.*, vol. 40, no. 3, pp. 532–542, Mar. 1992.
[8] H. Brandenstein and R. Unbehauen, "Least-squares approximation of FIR by IIR digital filters," *IEEE Trans. Signal Process.*, vol. 46, no. 1, pp. 21–30, Jan. 1998.
[9] R. W. Aldhaheri, "Design of linear-phase IIR digital filters using singular perturbational model reduction," *Proc. Inst. Elect. Eng.—Vision, Image, Signal Process.*, vol. 147, no. 5, pp. 409–414, Oct. 2000.
[10] H. Brandenstein and R. Unbehauen, "Weighted least-squares approximation of FIR by IIR digital filters," *IEEE Trans. Signal Process.*, vol. 49, no. 3, pp. 558–568, Mar. 2001.
[11] H. K. Kwan and A. Jiang, "Recent advances in FIR approximation by IIR digital filters," in *Proc. Int. Conf. Communications, Circuits, Systems*, Jun. 2006, pp. 185–190.
[12] W. Hendrickx and T. Dhaene, "A discussion of "Rational" approximation of frequency domain responses by vector fitting," *IEEE Trans. Power Syst.*, vol. 21, no. 1, pp. 441–443, Feb. 2006.
[13] C. Sanathanan and J. Koerner, "Transfer function synthesis as a ratio of two complex polynomials," *IEEE Trans. Autom. Control*, vol. 8, no. 1, pp. 56–58, Jan. 1963.
[14] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 1st ed. Englewood Cliffs, NJ: Prentice-Hall, 1989.
[15] M. C. Lang, "Least-squares design of IIR filters with prescribed magnitude and phase responses and a pole radius constraint," *IEEE Trans. Signal Process.*, vol. 48, no. 11, pp. 3109–3121, Nov. 2000.
[16] S. Grivet-Talocia and M. Bandinu, "Improving the convergence of vector fitting for equivalent circuit extraction from noisy frequency responses," *IEEE Trans. Electromagn. Compat.*, vol. 48, no. 1, pp. 104–120, Feb. 2006.

[17] K. Steiglitz and L. McBride, "A technique for the identification of linear systems," *IEEE Trans. Automat. Control*, vol. 10, no. 4, pp. 461–464, Oct. 1965.
[18] P. A. Regalia and M. Mboup, "Undermodeled adaptive filtering: An *a priori* error bound for the Steiglitz–McBride method," *IEEE Trans. Circuits Syst. II*, vol. 43, no. 2, pp. 105–116, Feb. 1996.
[19] W. S. Lu, S. C. Pei, and C. C. Tseng, "A weighted least-squares method for the design of stable 1-D and 2-D IIR digital filters," *IEEE Trans. Signal Process.*, vol. 46, no. 1, pp. 1–10, Jan. 1998.
[20] C. C. Tseng and S. L. Lee, "Minimax design of stable IIR digital filter with prescribed magnitude and phase responses," *IEEE Trans. Circuits Syst. I*, vol. 49, no. 4, pp. 547–551, Apr. 2002.

Quickest Detection and Tracking of Spawning Targets Using Monopulse Radar Channel Signals

Atef Isaac, Peter Willett, and Yaakov Bar-Shalom

***Abstract*—Recent advances have been reported in detecting and estimating the location of more than one target within a single monopulse radar beam. Successful tracking of those targets has been achieved with the aid of nonlinear filters that approximate the targets' states' conditional pdf, bypassing the measurement extraction stage, and operating directly on the monopulse sum/difference data, i.e., without measurement extraction. The problem of detecting a target spawn will be tackled in this paper. Particle filters will be employed as nonlinear tracking filters to approximate the posterior probability densities of the targets' states under different hypotheses of the number of targets, which in turn can be used to evaluate the likelihood ratio between two different hypotheses at subsequent time steps. Ultimately, a quickest detection procedure based on sequential processing of the likelihood ratios will be used to decide on a change in the underlying target model as an indication of a newly spawning target. Radar signal processing, data association, and target tracking are handled simultaneously.**

***Index Terms*—Monopulse radar, particle filter, target tracking, unresolved targets.**

I. Introduction

There has been a growing interest in the early detection of missiles that separate from a re-entry platform. Due to the limited resolution of monopulse radar and the fact that the separating missiles were in essence (*just moments back in time*) parts of the same platform, return signals from those objects merge altogether for a single radar measurement (a matched filter sample). A simple monopulse ratio [4] estimated DOA (direction of arrival) will erroneously correspond to a single object that best matches the observation, nonindicative of the true number of those separate objects, their locations, and, more important, the time they were set apart.

Manuscript received August 16, 2006; revised August 6, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Steven M. Kay. This research was supported by the Office of Naval Research (ONR) and by the Army Research Office (ARO).

A. Isaac was with the Electrical and Computer Engineering Department, University of Connecticut, Storrs, CT 06269 USA. He is now with Agilent Technologies, Inc., Santa Clara, CA 95051 USA (e-mail: atef_isaac@agilent.com).

P. Willett and Y. Bar-Shalom are with the Electrical and Computer Engineering Department, University of Connecticut, Storrs, CT 06269 USA (e-mail: willett@engr.uconn.edu; ybs@engr.uconn.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2007.908982

In principle, a monopulse radar provides excellent sub-beam angular estimation by comparing energy returns of four squinted sub-beams steered symmetrically around the expected single target location [13]. Work has been done to overcome the aforementioned problem, i.e., to resolve the multiple objects lying within a single monopulse radar beam. In [5], the authors extended the complex monopulse ratio processing concept to the two-target case by using the in-phase and quadrature components of the complex monopulse ratio. They developed the monopulse ratio conditional pdf of the amplitude of the sum signal in [6] and then used it to develop the CRLB for the DOA estimator. In [4], they developed an angle estimator based on the in-phase and quadrature components of the complex monopulse ratio and the observed signal strengths for two unresolved Rayleigh targets. A maximum-likelihood (ML) angle estimator for both the Swerling I and Swerling III targets [13] was developed in [17]. A closed-form ML solution that replaces the numerical search of [17] was given in [18]. By correlating the radar's consecutive matched filter samples on its three channels (the sum, horizontal difference and the vertical difference channels) and utilizing the models developed in [5], the authors in [20] upper bounded the identifiability of the number of targets to five, and imposed a minimum description length (MDL) penalty to help in this discrimination.

However, when the ML solution of [20] was coupled to Kalman filters in [11], its tracking results proved inferior to what was obtained when using a joint particle filter that integrated the tracking with the measurement extraction tasks. This was done by having the particle filter operate directly on the radar channel signals, i.e., without target position extraction techniques based on monopulse ratio processing. In addition, it was asserted in [19] that the techniques in [20] are ineffective at deciding the number of targets on a scan-by-scan basis when the targets are close (which is the case when one target is spawning from the other); this was due to the fact that the likelihood function is often maximized by two collocated targets [6], [17], and an incorrect preference for a single target decision is often exhibited by the MDL criterion. Hence, some sort of memory is needed to accumulate the confidence that an underlying system model change is taking place. For this purpose, and influenced by the good tracking results in [11], we will additionally call upon the ideas from [7] to construct sequentially the likelihood ratio using the particles' un-normalized weights. The likelihood ratio at any given time step is central to classical threshold-based tests, such as the Neyman–Pearson test that maximizes the probability of detection for a given false alarm probability. However, our concern in this paper is to detect a target spawn event in the shortest time possible, i.e., to minimize the stopping time (the time at which a final detection decision is taken). To this purpose, we will adopt Page's CUSUM (cumulative sum) procedure as a change detection scheme that does not require the knowledge of the starting point of a model change, to process the sequential likelihood ratio functions as they arrive and to declare a target spawn event whenever the CUSUM exceeds a threshold. Data association is implicitly handled by this new algorithm whenever a decision is made on the number of targets.

The paper describes the measurement model in Section II, in which the nonlinearity of the filtering model is stressed. If the filtering problem is nonlinear, then a particle filter offers a reasonable choice, and in Section III we discuss the version we use: the *auxiliary* particle filter. Particle filters work well for estimation (location), but we have already seen that for the closely-spaced monopulse problem in [11]. What we offer here is an integrated determination of the *number* of targets. Hypothesis testing and the "quickest-detector" Page test (see Section V) can directly use information from the particle filters, as described in Section IV. The paper concludes with simulation results in Section VI.