The Identification of Nonlinear Discrete-Time Fading-Memory Systems Using Neural Network Models

Michael B. Matthews and George S. Moschytz, *Fellow, IEEE*

Abstract— **A fading-memory system is a system that tends to forget its input asymptotically over time. It has been shown that discrete-time fading-memory systems can be uniformly approximated arbitrarily closely over a set of bounded input sequences simply by uniformly approximating sufficiently closely either the external or internal representation of the system. In other words, the problem of uniformly approximating a fading-memory system reduces to the problem of uniformly approximating continuous real-valued functions on compact sets. The perceptron is a parametric model that realizes a set of continuous real-valued functions that is uniformly dense in the set of all continuous real-valued functions. Using the perceptron to uniformly approximate the external and internal representations of a discrete-time fading-memory system results, respectively, in simple finite-memory and infinite-memory parametric system models. Algorithms for estimating the model parameters that yield a best approximation to a given fading-memory system are discussed. An application to nonlinear noise cancellation in telephone systems is presented.**

I. Introduction

SYSTEM identification is the process of constructing a mathematical model of a given system based on observations of the system's behavior in response to past inputs. The model predicts the behavior of the system in response to future inputs. The model may not be exact in the sense that it always predicts the exact behavior of the system. In other words, the model output may be only an approximation of the system output in response to the same input. In many applications, one requires a model whose output remains uniformly close to the output of the system over all time, i.e., one requires a model that uniformly approximates the system. [Hereafter, when we write "approximate" we will mean "uniformly approximate," unless otherwise indicated.] For example, in the echo-cancellation application for telephone systems described in [1] and discussed in this paper, an accurate model of the near-end echo path enables one to accurately estimate and cancel over all time the near-end echo component in the incoming signal. Finally, the model should be such that the approximation can be made arbitrarily close simply by increasing the order of the model.

Manuscript received February 1, 1993; revised October 25, 1993. This paper was recommended by Associate Editor G. E. Ford.

M. B. Matthews is with MBARI, Pacific Grove, CA 93950 USA.

G. S. Moschytz is with the Signal and Information Processing Laboratory, Swiss Federal Institute of Technology, Zürich, Switzerland.

IEEE Log Number 9404751.

Not all systems can be approximated arbitrarily closely in the above sense. In feedback systems, for example, an inexact model produces errors that may manifest themselves in the state of the model; such errors can accumulate causing the model output to diverge from system output in response to the same input. As a specific example, consider a nonlinear autonomous system with multiple equilibrium states. An inexact model of this system may have regions of convergence in the state space that do not coincide exactly with those of the original system. Hence, if the initial state is close to the border of these regions, it is possible that the model will converge to an equilibrium state different from that of the original system.

One common class of systems that can in fact be approximated arbitrarily closely is the class of so-called fading-memory systems [4], [5], [20]. As their name implies, fading-memory systems are systems that "forget" their inputs asymptotically over time in a well-defined manner to be discussed shortly. It was shown in [14], that such systems can be approximated arbitrarily closely on a set of bounded input signals simply by approximating sufficiently closely either the external or internal representation of the system. In other words, the problem of approximating fading-memory systems reduces to the problem of approximating continuous functions on compact sets[1].

Since the asymptotic behavior of a stable linear time-invariant system is independent of its initial state, such a system is a good example of a fading-memory system. A stable linear discrete-time time-invariant system, for example, can be approximated arbitrarily closely simply by approximating sufficiently closely either its unit-sample response (external representation) or its poles and residues (internal representation). By approximating the unit-sample response of a linear discrete-time system, we mean constructing a linear finite-memory model, i.e., an FIR filter, whose unit-sample response is close in some sense to that of the system. Similarly, by approximating the poles of a linear system, we mean constructing a linear infinite-memory model, i.e., an IIR filter, whose poles are close in some sense to the dominant poles of the linear system. Other examples of fading-memory systems include systems composed of the concatenation of linear time-invariant systems and zero-memory nonlinearities[2], e.g. the

[1] A compact set of a finite-dimensional vector space is a set that is closed and bounded.

[2] A zero-memory nonlinearity is a system whose output at any point in time is a function of the input only at that point in time.

Wiener and Hammerstein models [19], [24]. Furthermore, any feedback system whose internal representation is characterized by a contraction mapping is also a fading-memory system [15].

In general, approximating the external representation of a fading-memory system results in a nonlinear finite-memory model, i.e., a model whose output at a particular point in time is a nonlinear function of a finite number of past inputs. Similarly, approximating the internal representation of a fading-memory system results in a nonlinear infinite-memory model, i.e., a model whose state at any point in time is a nonlinear function of its initial state. In this paper, we examine two such models based on neural networks.

The perceptron is a type of "neural network" composed of the linear combination of affinely-transformed saturating nonlinearities [17]. The perceptron is a parametric model that realizes a set of continuous real-valued functions that is uniformly dense in the set of all continuous real-valued functions [6]–[9]. In other words, every continuous function can be approximated arbitrarily closely on a compact set over the set of functions realized by the perceptron. This fact is not surprising; many parametric models (e.g., polynomials) realize dense sets of continuous functions. However, because of its saturating nature, using the perceptron to approximate the external and internal representations of a fading-memory system results in finite and infinite-memory parametric models, respectively, that possess attractive qualities in terms of analytical tractability (e.g., stability analysis) and the simplicity of parameter estimation algorithms.

The concept of using neural network models for the identification of nonlinear discrete-time dynamical systems has been studied by several investigators, most notably in the well-cited paper [18]. In that paper, however, the nonlinear system models were chosen *ad hoc* without reference to any practical systems. Furthermore, for the system models that were chosen, it is unclear how simply substituting a perceptron for the model nonlinearities results in a system approximation in any sense. Finally, no results regarding the stability of such models were provided.

This paper attempts to provide a unified approach to the problem of identifying a broad class of nonlinear systems, in particular fading-memory systems, using the perceptron. It presents an example of the application of such perceptron system models to nonlinear echo-cancelling in telephone systems. Because of the inherent nonlinearities in such systems in the form of data converters, amplifiers, and hybrid circuits, linear echo-cancellers of arbitrary order have an upper bound of echo rejection [1]. However, since telephone systems have inherent fading memory characteristics, they can in principle be approximated arbitrarily closely by the nonlinear models described herein, i.e. there is no upper bound on the echo rejection attainable with echo-cancellers based on such models.

II. Fading-memory System Approximation

We consider the class of discrete-time time-invariant systems of the form

$$\begin{aligned} x[k+1] &= f(x[k], u[k]) \\ y[k] &= h(x[k]) \end{aligned} \tag{1}$$

for all $k \in \mathbb{N} = \{0, 1, 2, \ldots\}$, where the input $u[k]$, the state $x[k]$, and the output $y[k]$ at time k belong to finite-dimensional real vector spaces U, X, and Y, respectively, where the *state map* $f : X \times U \to X$ and the *output map* $h : X \to Y$ are continuous, and where the initial state is x_0, i.e., $x[0] = x_0$. Let $U^k = U \times U \times \cdots \times U$ (k times) denote the product space of k-length sequences of elements of U, and let U^* denote the set of all finite-length sequences of elements of U including the empty sequence Λ, i.e., $U^* \triangleq \bigcup_{k \geq 0} U^k$. Hereafter, all sequences in U^* will be written in bold-face type, e.g., $\boldsymbol{u} \in U^*$. For sequences $\boldsymbol{u}, \boldsymbol{v} \in U^*$, $\boldsymbol{uv}$ denotes the concatenation of the two sequences. It is convenient to extend the state map f to the *reachability map* $f^* : X \times U^* \to X$ such that $x_0 = f^*(x_0, \Lambda)$ and $x[k+1] = f^*(x_0, u[0], u[1], \ldots, u[k])$ for all $k \in \mathbb{N}$. Similarly, we extend the output map h to the *response map* $h^* : X \times U^* \to Y$ such that $h^* = h \circ f^*$, *i.e.*, $y[k+1] = h^*(x_0, u[0], u[1], \ldots, u[k])$ for all $k \in \mathbb{N}$. Hereafter, when we write "system," we will mean a discrete-time system of the form in (1).

Functions f and h in (1) describe the system from a state-variable or *internal representation* point of view. To create this representation, one requires *a priori* knowledge of the actual physical system, i.e., one requires knowledge of its physical states and access to these states. Alternatively, the response map h^* provides an input-output or *external representation* of the system. To create an external representation, it suffices to make input-output measurements on the system.

Intuitively, we would expect a model to be a "good" approximation of a system if, in response to the same input, the model output remains uniformly close to the system output over all time. Indeed, in many applications one requires a model whose output remains within a fixed envelope around the system output over all time. In the echo-canceller application described in Section V, for example, it is desirable to have an upper bound on the uncancelled echo component, i.e., on the system/model error, in the received signal.

We say that a model *uniformly approximates* a system to within $\epsilon > 0$ on some subset K of the set of finite input sequences U^* if, for every sequence in K applied simultaneously to both model and system, the output sequences remain close to within ϵ over the entire length of the input sequence. More precisely, a model $\bar{\Sigma}$ with response map $\bar{h}^*$ uniformly approximates a system Σ with response map h^* to within ϵ on $K \subset U^*$ if

$$\sup_{\boldsymbol{u} \in K} \|h^*(x_0, \boldsymbol{u}) - \bar{h}^*(\bar{x}_0, \boldsymbol{u})\| < \epsilon \tag{2}$$

where $\|.\|$ denotes an arbitrary norm on the output space Y.

In general, feedback systems cannot be uniformly approximated in the above sense. In many systems with certain memory characteristics, an inexact model will produce errors that may tend to accumulate in the state of the model so as to cause the model output to diverge from the system output. As suggested in the introduction, systems with unique asymptotic properties, i.e., systems with fading memory, can in fact be approximated arbitrarily closely in the above sense simply by approximating sufficiently closely either the internal or external representation. To see this, we need to define explic-

itly what we mean by "fading memory." We will approach the definition of fading memory from the point of view of continuity of systems.

A function is said to be continuous if elements in the domain that are close have images that are close. Similarly, a system is said to be continuous if input sequences that are close in some sense produce output sequences that are close in some sense. To be precise, a system Σ is said to be continuous on some subset K of U^* if, for every $\epsilon > 0$ and $\boldsymbol{u} \in K$, there is a $\delta > 0$ depending on both ϵ and $\boldsymbol{u}$ such that, for every $\boldsymbol{v} \in K$,

$$\|\boldsymbol{u} - \boldsymbol{v}\|_* < \delta \implies \|h^*(x_0, \boldsymbol{u}) - h^*(x_0, \boldsymbol{v})\| < \epsilon \quad (3)$$

where $\|.\|_*$ denotes an arbitrary norm on U^* and where $\|.\|$ denotes an arbitrary norm on the output space Y. The distance measure between two sequences in U^*, i.e., the norm $\|.\|_*$, can be defined in many ways. It is convenient to define a *weighted norm* $\|.\|_w$ on U^* with respect to a nonnegative real sequence w in the manner that

$$\|\boldsymbol{u}\|_w = \max_{n \in [0, |\boldsymbol{u}|]} \|u[n]\| \, w[\,|\boldsymbol{u}| - n] \quad (4)$$

where $|\boldsymbol{u}|$ denotes the length of the sequence $\boldsymbol{u}$ and where $\|.\|$ denotes an arbitrary norm on the input space U. A *fading-memory system* is a system that tends to 'forget' its input in a well-defined manner over time. In other words, a fading-memory system is one in which, at any given time, input sequences that are close in the *recent past* produce output sequences that are close at that time. The following definition of a fading-memory system is based on that in [4] for continuous-time systems.

Definition: A system $\Sigma = (U, X, Y, f, h, x_0)$ has *fading memory on a subset* K of U^* if there exists a nonincreasing positive sequence w with $\lim_{k \to \infty} w[k] = 0$ such that, for every $\epsilon > 0$ and every $\boldsymbol{u} \in K$, there exists a $\delta = \delta(\epsilon, \boldsymbol{u}) > 0$ such that, for all $\boldsymbol{v} \in K$ with $|\boldsymbol{v}| = |\boldsymbol{u}|$,

$$\|\boldsymbol{u} - \boldsymbol{v}\|_w < \delta \implies \|h^*(x_0, \boldsymbol{u}) - h^*(x_0, \boldsymbol{v})\| < \epsilon \quad (5)$$

where $\|.\|_w$ denotes the weighted norm on U^* with respect to the sequence w . □

In other words, at some point in time k, if the k-length input sequences $\boldsymbol{u}$ and $\boldsymbol{v}$ remain within the envelope $\delta / w[k - n]$ for $n \in [0, k]$, then the outputs will be within ϵ at time $k + 1$. According to Definition 1, a fading-memory system is a system whose response map h^* is *continuous with respect to a weighted norm* $\|.\|_w$ *on* U^* *where the sequence* w *is positive, nonincreasing, and converges to zero.*

From Definition 1, it is clear that a fading-memory system has the asymptotic property that, when the same input sequence is applied to the system at distinct initial states, the output trajectories of the system will converge. To be more precise, given a system Σ with fading memory on some subset K of U^*, for every $\epsilon > 0$, there is a positive integer n such that for all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{s} \in K$ with $|\boldsymbol{s}| \geq n$, it follows that $\|h^*(x_0, \boldsymbol{us}) - h^*(x_0, \boldsymbol{vs})\| < \epsilon$. The proof of this result, which is similar to that in [4] for continuous-time systems, can be found in [15].

The above asymptotic property of fading-memory systems implies that if x_0 is an equilibrium state of system Σ, i.e., if $f(x_0, 0) = x_0$, then, for all sequences $\boldsymbol{u} \in K$

$$\| h^*(x_0, u[0], u[1], \ldots, u[k]) - h^*(x_0, 0, \ldots, 0, u[k - n + 1], \ldots, u[k]) \| < \epsilon \,. \quad (6)$$

Hence, any system Σ with an equilibrium state at x_0 and with fading memory on a set of bounded input sequences can be approximated arbitrarily closely by a model $\bar{\Sigma}$ that realizes an arbitrary continuous function of a finite number of past input samples. A discrete-time system whose output is a function of a finite number of past input samples will be referred to as a *finite-memory system.* It is clear from Definition 1 that such systems are fading-memory systems. Systems that are not finite-memory systems will be referred to as *infinite-memory systems.* The model $\bar{\Sigma}$ then consists of an nth-order "vector tapped-delay-line" (i.e., a linear system whose output in U^n consists of the last n input samples) concatenated with a structure that realizes a continuous mapping $\bar{h} : U^n \to Y$ of the form $\bar{h}(\boldsymbol{u}) = h^*(x_0, \boldsymbol{u})$. In fact, it is neither practical nor necessary to realize $\bar{h}$ exactly; it suffices simply to approximate $\bar{h}$ arbitrarily closely on a compact subset of U^n. Hence, the problem of approximating arbitrarily closely a fading-memory system reduces to the problem of approximating arbitrarily closely continuous real-valued functions on compact sets.

The concept of fading memory is also closely related to that of a contraction mapping[3] in the sense that if the internal representation of a system is defined by a contraction mapping on a subset of the input space, then the system has fading memory on sequences of elements in that subset. To be more precise, given a system Σ such that, for every u in a compact subset K of U, the state mapping $f(\,.\,, u) : X \to X$ is a contraction mapping, then Σ has fading memory on K^*. This result, whose proof can again be found in [15], implies that we can approximate a fading-memory system arbitrarily closely on a set of bounded input sequences simply by approximating arbitrarily closely its state-space or internal representation. By this we mean that, a fading-memory system Σ defined in (1) by the state map f and the output map h can be approximated arbitrarily closely by a model $\bar{\Sigma}$ defined by the state map $\bar{f}$ and the output map $\bar{h}$, where $\bar{f}$ is an arbitrarily-close approximation of f on a compact subset of $X \times U$, and where $\bar{h}$ is an arbitrarily-close approximation of h on a compact subset of X. A precise statement of this result can be found in [15].

III. The Perceptron as a Universal Approximator

The important result in the previous section is that the problem of approximating a fading-memory system on a set of bounded input sequences reduces to the problem of approximating continuous real-valued functions on compact sets. We denote by $C(U)$ the set of all continuous real-valued functions on U. Any subset of $C(U)$ is said to be *uniformly dense* in $C(U)$ if, for every element f in $C(U)$ and every small $\epsilon > 0$, there is an element $\bar{f}$ in that subset that uniformly approximates f to within ϵ on U, that is $\sup_{u \in U} |f(u) - \bar{f}(u)| < \epsilon$. There are many parametric models

[3] A contraction mapping is a mapping $f : X \to X$ such that there is an $\alpha \in (0, 1)$ such that $\rho(f(x), f(\bar{x})) \leq \alpha \, \rho(x, \bar{x})$ for all $x, \bar{x} \in X$, where ρ is a metric on X.

that realize sets of continuous real-valued functions that are uniformly dense in $\mathcal{C}(U)$. One common set is the set of polynomials $\mathcal{P}(U)$ of the form

$$\nu(u) = \sum_{|\alpha| \leq m} a_\alpha u^\alpha \tag{7}$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ is a multiindex, where $|\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_n$, where $a_\alpha \in \mathbb{R}$ and where $u^\alpha = u_1^{\alpha_1} u_2^{\alpha_2} \cdots u_n^{\alpha_n}$ for $n = \dim U$. That $\mathcal{P}(K)$ is uniformly dense in $\mathcal{C}(K)$ for any compact subset K of U, i.e., that any function in $\mathcal{C}(K)$ can be approximated arbitrarily closely over the set of polynomials, is the classic result from Weierstrass [11]. In fact, for any finite-dimensional output space Y, we can approximate arbitrarily closely any continuous mapping f from U into Y simply by approximating sufficiently closely each component function f_i of f by a polynomial ν_i, $i = 1, 2, \ldots, \dim Y$.

Approximating the *external representation* of a fading-memory system over the set of polynomials on U^n, i.e., approximating the mapping $\bar{h}(\boldsymbol{u}) = h^*(x_0, \boldsymbol{u})$ over $\mathcal{P}(U^n)$, results in a finite-memory polynomial system model called a *polynomial filter* [14]. Such a system can also be arrived at in the context of truncated Volterra series representations of nonlinear systems [4]. The polynomial filter has the advantage of being linear in its parameters; hence, a parameter estimation algorithm based on a quadratic cost function will converge to a unique solution, provided that the stability of the algorithm is ensured. The disadvantage of the polynomial filter lies in the inordinate number of parameters required to specify a polynomial that well-approximates even the simplest fading-memory systems: an mth-degree polynomial on $\mathbb{R}^n$ requires

$$\sum_{i=0}^{n} \binom{m-i+1}{i}$$

parameters.

Aside from the inordinate number of parameters required to effect a good approximation of a given fading-memory system, the polynomial filter also exhibits a pathological behavior (the output increases without bound) for inputs outside the compact set on which the approximation is defined. In other words, outliers in the input signal are amplified exponentially and must therefore be clipped either at input or at the output.

Alternatively, approximating the *internal representation*, i.e., the state and output mappings f and h, of a fading-memory system over the set of polynomials on $X \times U$ and X, respectively, results in an infinite-memory polynomial system model [19], [22]. The algebraic properties of polynomial systems were studied in [23]. Unfortunately, polynomial models are not inherently stable, and no general stability criterion exists. Even though such a model may be stable for inputs inside the compact set on which the approximation is defined, it may be unstable for inputs outside this set. Furthermore, assuring the stability of a polynomial system model is necessary for implementing an on-line parameter-estimation algorithm. Because of the aforementioned problems with both finite-memory and infinite-memory polynomial system models, their practical application has been limited to well-behaved mildly nonlinear systems, [1], [3], [10].

Consider instead the set of real-valued functions on U of the form

$$u \mapsto \sum_{j=1}^{m} c_j \, \phi(\alpha_j(u)) \tag{8}$$

where $c_j \in \mathbb{R}$, where $\phi : \mathbb{R} \to \mathbb{R}$ is a nonconstant Borel measurable function[4], and where α_j is a real-valued affine mapping on U of the form $\alpha_j(u) = A_j u + b_j$ with $A_j : U \to \mathbb{R}$ a linear mapping and $b_j \in \mathbb{R}$. That the set of functions of the form in (8), where ϕ is *discriminatory*, is uniformly dense in $\mathcal{C}(K)$ for any compact subset K of U was shown in [6]. That every nonconstant Borel measurable function is discriminatory was shown in [8].

The power of the general model in (8) lies in our ability to choose virtually any nonconstant continuous function ϕ that best facilitates the practical implementation of (8) in a system model. By this we mean choosing ϕ for both its efficiency in terms of the number m of elements in (8) required to effect a good approximation of a given function, and for the utility of ϕ in terms of making (8) both analytically tractable and well-behaved when used in a system model.

Few physical systems are inherently "polynomial" in nature in the sense that the output of the system increases without bound as the input increases. Most physical systems are finite in the sense that they are composed of components whose outputs saturate as their inputs increase without bound. A multistage amplifier, for example, consists of several concatenated amplifier stages each with a limited dynamic range. Furthermore, in many cases, it is this very saturating property that characterizes the nonlinear behavior of the system which we wish to approximate. Would it not be logical to choose for the general model in (8) a function ϕ that saturates for large inputs? In this regard, we define a *sigmoid function* to be a real-valued strictly-increasing continuous function $\phi : \mathbb{R} \to (0, 1)$ with $\lim_{x \to \infty} \phi(x) = 1$ and $\lim_{x \to -\infty} \phi(x) = 0$. That such a function is discriminatory was shown in [6].

We define a *perceptron* to be a 6-tuple $(U, X, Y, A, C, \boldsymbol{\phi})$, where U (input space), X (hidden-layer space) and Y (output space) are finite-dimensional real vector spaces, where $A : U \to X$ is an affine mapping, where $C : X \to Y$ is a linear mapping, and where $\boldsymbol{\phi} : X \to X$ is a vector-valued sigmoid function, i.e., $\boldsymbol{\phi}(x_1, x_2, \ldots, x_m) = (\phi_1(x_1), \phi_2(x_2), \ldots, \phi_m(x_m))$ with each ϕ_i, $i = 1, 2, \ldots, m$, $m = \dim X$, a sigmoid function. We define the *order* of a perceptron to be the dimension of the hidden-layer space X. We interpret an mth-order perceptron as realizing a continuous function $\psi : U \to Y$ of the form $\psi = C \circ \boldsymbol{\phi} \circ A$, which is simply a multidimensional form of (8). In other words, for $\dim Y = l$, the component functions $(\psi_1, \psi_2, \ldots, \psi_l)$ of ψ can be written as

$$\psi_i(u) = \sum_{j=1}^{m} c_{ij} \, \phi_j(\alpha_j(u)) \qquad i = 1, 2, \ldots, l \tag{9}$$

[4] A Borel measurable function is a function such that the inverse image of a Borel set, i.e., a set composed of the countable union and intersection of sets of the form $\{x : x \leq a\}$, is a Borel set. In practical terms, a Borel measurable function is a nonpathological function.

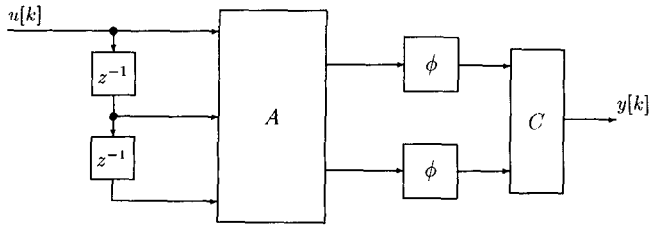


Fig. 1. A (3,2)th-order perceptron filter.

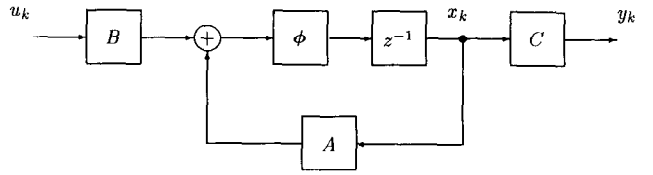


Fig. 2. Recurrent network structure.

with $A = (\alpha_1, \alpha_2, \ldots, \alpha_m)$, where $\alpha_j : U \to \mathbb{R}$, $j = 1, 2, \ldots, m$, is an affine mapping. It follows from the result for the scalar-valued model in (8) that the set of continuous mappings $\psi : U \to Y$ realized by a perceptron is uniformly dense in the set of all continuous functions from U into Y. To be more precise, for every $\epsilon > 0$, for every function $f : U \to Y$, and for every compact subset K of U, there is a positive integer m and an mth-order perceptron which realizes the mapping $\psi : U \to Y$ such that $\sup_{u \in K} \|f(u) - \psi(u)\| < \epsilon$ where $\|.\|$ is an arbitrary norm on Y.

Approximating the external representation of a fading-memory system over the set of continuous real-valued functions on U^n realized by the perceptron, i.e., approximating the mapping $\bar{h}(\boldsymbol{u}) = h^*(x_0, \boldsymbol{u})$ by the mapping ψ, results in the finite-memory system shown in Fig. 1 which we call the *perceptron filter* [13]. We define an (n, m)th-order perceptron filter to be a model consisting of an nth-order tapped-delay-line followed by an mth-order perceptron. Clearly, the perceptron filter realizes a set of finite-memory systems that is uniformly dense in the set of fading-memory systems.

Approximating the internal representation of a fading-memory system, i.e., approximating the state and output mappings f and h over the set of continuous functions on $X \times U$ and X, respectively, realized by the perceptron, results in the infinite-memory model shown in Fig. 2 of the form

$$\begin{aligned} x[k+1] &= \boldsymbol{\phi}(Ax[k] + Bu[k]) \\ y[k] &= Cx[k] \end{aligned} \tag{10}$$

where the input $u[k]$, the state $x[k]$, and the output $y[k]$ at time k belong to the finite-dimensional real spaces U, X, and Y, respectively, where $A : X \to X$ is a linear mapping, where $B : U \to X$ and $C : X \to Y$ are affine mappings, and where $\boldsymbol{\phi} : X \to X$ is a vector-valued sigmoid function, i.e., $\boldsymbol{\phi}(x_1, x_2, \ldots, x_n) = (\phi_1(x_1), \phi_2(x_2), \ldots, \phi_n(x_n))$, $n = \dim X$.

We call such a model a *recurrent network* and define it by the 8-tuple $(U, X, Y, A, B, C, \boldsymbol{\phi}, x_0)$ where $x_0 \in X$ is the initial state. We define the order of a recurrent network to be the dimension of the state space X. That the form in Fig. 2 results from the approximation of f and h by a perceptron is shown in the Appendix. That the recurrent network realizes a set of infinite-memory systems that is uniformly dense in the set of fading-memory systems is discussed in [15].

The recurrent network has a very simple structure in that the sigmoid function $\boldsymbol{\phi}$ is "decoupled," i.e., each component function ϕ_i of $\boldsymbol{\phi}$ is a function only of its respective component x_i for $i = 1, 2, \ldots, \dim X$. Because of this unique decoupled structure, the recurrent network has a simple stability criterion related to its underlying linear system. To be precise, an nth-order recurrent network is globally asymptotically stable at an equilibrium state if

$$\|A\| < \max_{i \in [1,n]} \sup_{x \in \mathbb{R}} \dot{\phi}_i(x) \tag{11}$$

where $\dot{\phi}_i$ denotes the derivative of ϕ_i [14]. In other words, if the eigenvalue of A with the largest magnitude is less that the maximum value of derivative of the sigmoid functions, then, for any initial state x_0, the state of the recurrent network will, for zero input, converge to an equilibrium state. The key to this result is the fact that the recurrent network belongs to the well-studied class of so-called *sector-nonlinear systems* [2], [16], [21], [25], and [26]. Finally, we note that the linear stability criterion in (11) is only a sufficient condition for global asymptotic stability of the recurrent network.

IV. State and Parameter Estimation

In the preceding section, we introduced two system models, the perceptron filter and the recurrent network, that realize sets of finite-memory and infinite-memory systems, respectively, that are uniformly dense in the set of fading-memory systems. Furthermore, for every positive integer n, and every sigmoid function $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_n)$, there is a (not necessarily unique) best approximation for every fading-memory system Σ out of each set of finite-memory and infinite-memory systems realized by the nth-order perceptron and recurrent network, respectively, over the set of bounded mappings A, B, and C.

Let Σ be a system with fading memory on a subset K of U^*. We will consider the problem of determining the set of mappings $\{A, B, C\}$ that specify a recurrent network that is a best approximation to Σ. A similar treatment of the perceptron filter can be found in [14]. We will consider the mappings $\{A, B, C\}$ to be parametric functions on a real parameter space Θ of dimension $p = n^2 + nm + ln + n + l$, where $n = \dim X$, $m = \dim U$, and $l = \dim Y$. We will denote this by the notation $\{A(\theta), B(\theta), C(\theta)\}$ with $\theta \in \Theta$. In other words, for given bases for U, X, and Y, the parameter vector θ consists of the elements of the corresponding matrices A, B, and C.

Let $\Sigma(\bar{\theta})$ be a nth-order recurrent network, specified by the parameter vector $\bar{\theta} \in \Theta$, such that $\Sigma(\bar{\theta})$ approximates Σ to within ϵ on $K \subset U^*$, and such that $\Sigma(\bar{\theta})$ is a best approximation of Σ out of the set of systems realized by the nth-order recurrent network. We will make the hypothesis that, for any input random process u with realizations in K applied to both Σ and $\Sigma(\bar{\theta})$, the difference in their outputs $e = y - \bar{y}$ is a zero-mean i.i.d. random process whose distribution has support on a subset of the output space Y given by $\{y : y \in Y, |y_i| < \epsilon, i = 1, 2, \ldots, l\}$ and which

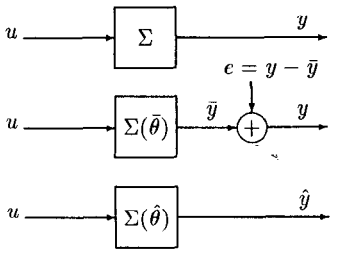


Fig. 3. The parameter estimation problem.

is independent of the input process u. This hypothesis, which we will call the "sufficient approximation hypothesis," is most likely untrue in a strict sense; however, it is useful insofar as it allows us to conceptually recast the problem of identifying Σ into one of estimating the parameter and state vectors of $\Sigma(\bar{\theta})$ given "noisy" observations $y = \bar{y} + e$ as illustrated in Fig. 3.

Referring to the form in (10), let $V_{k+1}(\theta)$ be a conditional mean-square cost function of the form

$$V_{k+1}(\theta) = \frac{1}{2} E\{e[k+1]^T \Lambda\, e[k+1] \mid \boldsymbol{y}[k], \boldsymbol{u}[k]\} \tag{12}$$

where Λ is an $l \times l$ diagonal weighting matrix, where

$$\boldsymbol{y}[k] = (y[0], y[1], \ldots, y[k])$$

is the observed sequence, where

$$\boldsymbol{u}[k] = (u[0], u[1], \ldots, u[k])$$

is the input sequence, and where

$$e[k+1] = y[k+1] - \hat{y}[k+1]$$

with

$$\hat{y}[k+1] = C(\theta)\hat{x}[k+1]$$

where $\hat{x}[k+1]$ is the *a priori* estimate of $x[k+1]$, i.e.,

$$\hat{x}[k+1] = E\{x[k+1] \mid \boldsymbol{y}[k], \boldsymbol{u}[k]\}.$$

Here $e[k+1]^T$ denotes the transpose of $e[k+1]$. Because the recurrent network is nonlinear in its state, we will choose for the *a priori* estimate $\hat{x}[k+1]$ the first order approximation of the optimal mean-square estimate in the usual manner [14],

$$\hat{x}[k+1] = \boldsymbol{\phi}(A(\theta)\bar{x}[k] + B(\theta)u[k])$$

where $\bar{x}[k] = \hat{x}[k] + K(\theta)(y[k] - C(\theta)\hat{x}[k])$ is the *a posteriori* estimate of $x[k]$. Here we include the linear "gain" operator $K(\theta) : Y \rightarrow X$ among the parametric functions in the parameter space Θ.

Based on the "sufficient approximation hypothesis," it is clear that $\min_{\theta \in \Theta} V_{k+1}(\theta) = V_{k+1}(\bar{\theta})$. Let $\hat{\theta}[k]$ denote the parameter estimate of $\bar{\theta}$ at time k and expand $V_{k+1}(\theta)$ in a Taylor series about $\hat{\theta}[k]$,

$$\begin{aligned} V_{k+1}(\theta) = {} & V_{k+1}(\hat{\theta}[k]) + \acute{V}_{k+1}(\hat{\theta}[k])\,(\theta - \hat{\theta}[k]) \\ & + \frac{1}{2}\,(\theta - \hat{\theta}[k])^T\, \acute{\acute{V}}_{k+1}(\hat{\theta}[k])\,(\theta - \hat{\theta}[k]) \\ & + O_{k+1}(\theta, \hat{\theta}[k]) \end{aligned} \tag{13}$$

where $\acute{V}_{k+1} = \frac{\partial V_{k+1}}{\partial \theta}$ and $\acute{\acute{V}}_{k+1} = \frac{\partial^2 V_{k+1}}{\partial \theta^2}$. If we neglect the higher-order term $O_{k+1}(\theta, \hat{\theta}[k])$, we can differentiate (13) with respect to θ, set the result equal to zero, and solve for an approximate minimum for $V_{k+1}(\theta)$. We may take this value to be the estimate of $\bar{\theta}$ at time $k+1$, i.e.

$$\hat{\theta}[k+1] = \hat{\theta}[k] - [\acute{\acute{V}}_{k+1}(\hat{\theta}[k])]^{-1}\, \acute{V}_{k+1}(\hat{\theta}[k])^T\,. \tag{14}$$

The expression in (14) is consistent since $V_{k+1}(\theta)$ is an ensemble average of the *a priori* output squared error $e[k+1]^T \Lambda\, e[k+1]$. We proceed to derive the recursive parameter estimation algorithm for the recurrent network in a manner similar to the Recursive Prediction Error (RPE) Algorithm discussed in [12] for linear state-space systems. This involves approximating in a recursive manner the expectation in the quantities $\acute{V}_{k+1}$ and $\acute{\acute{V}}_{k+1}$ by the time average. This results in the following algorithm [14].

Algorithm 1 (Recurrent Network RPE Algorithm):

$$\hat{x}[k+1] = A(\hat{\theta}[k])\,\boldsymbol{\phi}(\bar{x}[k]) + B(\hat{\theta}[k])\,u[k] \tag{15}$$

$$e[k+1] = y[k+1] - C(\hat{\theta}[k])\,\boldsymbol{\phi}(\hat{x}[k+1]) \tag{16}$$

$$\begin{aligned} \Gamma[k+1] = {} & A(\hat{\theta}[k])\,J(\bar{x}[k]) \\ & \Big[I - K(\hat{\theta}[k])\,C(\hat{\theta}[k])\,J(\hat{x}[k])\Big]\Gamma[k] + N[k] \end{aligned} \tag{17}$$

$$\begin{aligned} N[k] = {} & A(\hat{\theta}[k])\,J(\bar{x}[k]) \\ & \left[\left.\frac{\partial K(\theta)}{\partial \theta}\right|_{\hat{\theta}[k]}[k] - K(\hat{\theta}[k])\left.\frac{\partial C(\theta)}{\partial \theta}\right|_{\hat{\theta}[k]}\boldsymbol{\phi}(\hat{x}[k])\right] \\ & + \left.\frac{\partial A(\theta)}{\partial \theta}\right|_{\hat{\theta}[k]}\boldsymbol{\phi}(\bar{x}[k]) + \left.\frac{\partial B(\theta)}{\partial \theta}\right|_{\hat{\theta}[k]}u[k] \end{aligned} \tag{18}$$

$$\begin{aligned} \psi^T[k+1] = {} & C(\hat{\theta}[k])\,J(\hat{x}[k+1])\,\Gamma[k+1] \\ & + \left.\frac{\partial C(\theta)}{\partial \theta}\right|_{\hat{\theta}[k]}\boldsymbol{\phi}(\hat{x}[k+1]) \end{aligned} \tag{19}$$

$$R[k+1] = \lambda[k]\,R[k] + \beta[k]\,\psi[k+1]\,\Lambda\,\psi[k+1]^T \tag{20}$$

$$\hat{\theta}[k+1] = \hat{\theta}[k] + \beta[k]\,R[k+1]^{-1}\,\psi[k+1]\,\Lambda\,e[k+1] \tag{21}$$

$$[k+1] = y[k+1] - C(\hat{\theta}[k+1])\,\boldsymbol{\phi}(\hat{x}[k+1]) \tag{22}$$

$$\bar{x}[k+1] = \hat{x}[k+1] + K(\hat{\theta}[k+1])\,[k+1]. \tag{23}$$

The quantity $J(x)$ is the $n \times n$ diagonal Jacobian matrix corresponding to $\boldsymbol{\phi}(x)$ of the form $J(x_1, x_2, \ldots, x_n) = \text{diag}\{\acute{\phi}_1(x_1), \acute{\phi}_2(x_2), \ldots, \acute{\phi}_n(x_n)\}$. The quantities $\Gamma[k]$ and $N[k]$ are both $n \times p$ matrices, and $\psi[k+1]$ and $R[k]$ are $p \times l$ and $p \times p$ matrices, respectively. The sequence $\{\beta[k]\}$ is a decreasing weighting sequence and $\lambda[k]$ is a number close to one; a possibility is $\beta[k] = \beta[k-1](1 + \beta[k-1])^{-1}$ and $\lambda[k] = 1 - \beta[k]$ for all $k > 0$ with $\beta[0] = 1$.

In a similar approach, also based on the "sufficient approximation hypothesis," to the problem of estimating the parameter vector $\bar{\theta}$ which specifies a best approximation to a fading-memory system Σ out of the set of systems realized by the (n, m)th-order perceptron filter, we derived the following Least-Mean-Squares (LMS) algorithm [14].

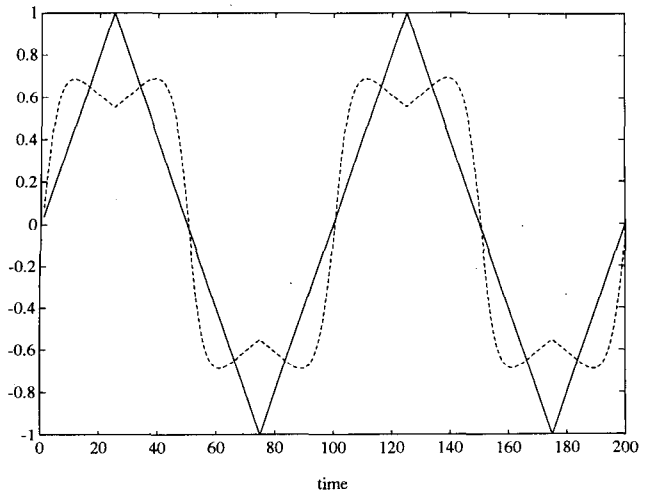


Fig. 4. 5th-order recurrent network fading-memory system response to a triangular wave.

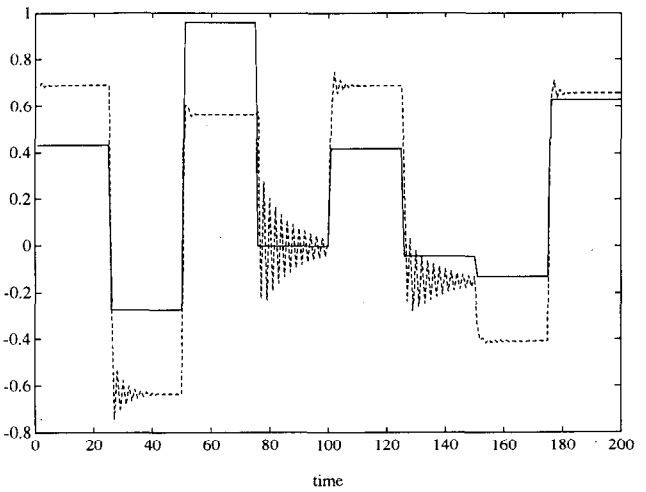


Fig. 5. 5th-order recurrent network fading-memory system response to a random step sequence.

Algorithm 2 (Perceptron Filter LMS Algorithm):

$$\hat{\theta}[k+1] = \hat{\theta}[k] - \beta[k]\,\dot{\psi}(\boldsymbol{u}[k]\,;\hat{\theta}[k])\,\Lambda^{-1} \cdot (y[k] - \psi(\boldsymbol{u}[k]\,;\hat{\theta}[k]))$$

$$\psi(\boldsymbol{u}[k];\hat{\theta}[k]) = C(\hat{\theta}[k]) \circ \boldsymbol{\phi} \circ A(\hat{\theta}[k])(\boldsymbol{u}[k])$$

$$\dot{\psi}(\boldsymbol{u}[k]\,;\hat{\theta}[k]) = \frac{\partial}{\partial\theta}\psi(\boldsymbol{u}[k];\theta)\bigg|_{\theta=\hat{\theta}[k]}$$

$$\boldsymbol{u}[k] = (u[k], u[k-1]\ldots., u[k-n+1]).$$

Example 1 (Recurrent Network Identification): In this example, a 5th-order scalar-input, scalar-output recurrent network is identified by a second, randomly-initialized recurrent network of the same order using Algorithm 1. The system exhibits severe nonlinear input-output distortion as shown in Fig. 4, as well as an amplitude-dependent underdamped low-pass characteristic as shown in Fig. 5. Each sigmoid function ϕ_i in $\boldsymbol{\phi}$ was chosen to be the same with a maximum derivative of one. The eigenvalues of the A matrix were chosen close to unity. Algorithm 1 was used to adapt the parameters of a second 5th-order recurrent network for a 500-sample zero-mean i.i.d. unity-variance Gaussian random input sequence. Fig. 6 shows performance of the 5th-order randomly-initialized recurrent network during learning of the fading-memory system; the first trace represents the random input sequence; the second trace represents the output of the test recurrent network; the third trace represents the output of the second recurrent network, and the fourth trace is the system-model output error. The corresponding single-ensemble squared-error learning curve is shown in Fig. 7.

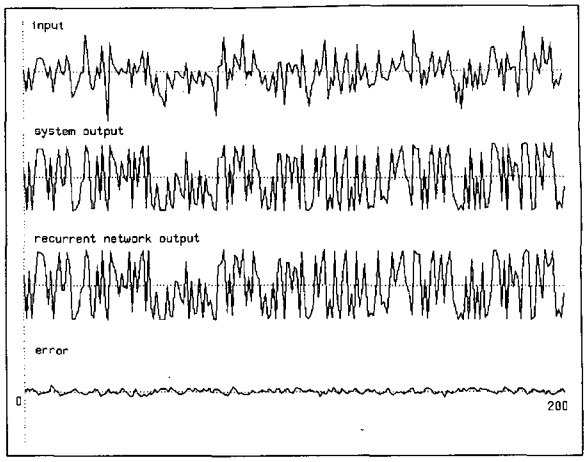


Fig. 6. 5th-order recurrent network performance during learning of a 5th-order recurrent network using Algorithm 1.

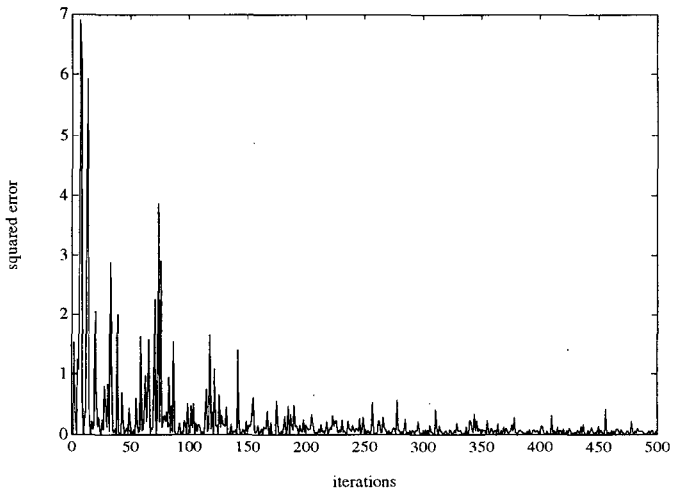


Fig. 7. 5th-order recurrent network single-ensemble squared-error learning curve for the fading-memory system.

V. A Nonlinear Echo-Cancellation Application

An example of the application of both the perceptron filter and the recurrent network is in near-end echo cancellation in telephone systems of the type shown in Fig. 8.

The near-end digital signal $r[k]$ is converted to analog by the digital-to-analog converter f_1 and shaped by the linear low-pass transmit filter H_1 to produce signal $\tilde{r}(t)$. The hybrid circuit separates $\tilde{r}(t)$ from the received far-end signal $s(t)$, placing $\tilde{r}(t)$ on the transmission-line, while routing $s(t)$

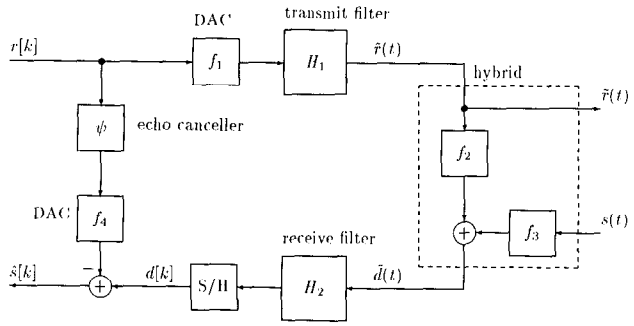


Fig. 8. Echo-cancelling example in telephone systems.

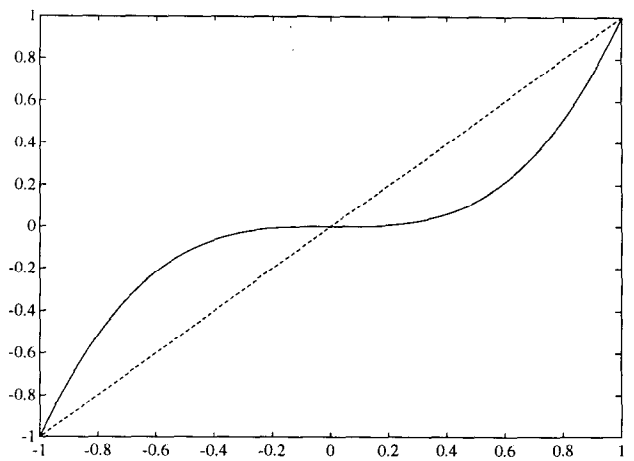


Fig. 9. DAC nonlinearity f_1.

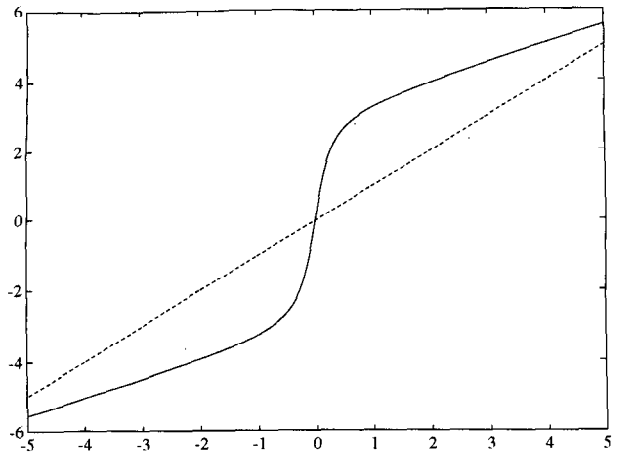


Fig. 10. Hybrid nonlinearity f_2.

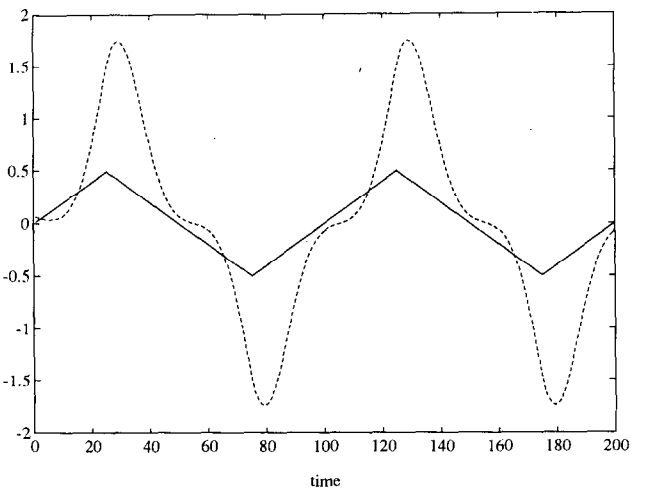


Fig. 11. Echo path model response to a triangular wave input.

through the linear receive filter H_2 and the sample-and-hold device (S/H). We will assume that the sample-and-hold device is ideal. Because of "leakage" in the hybrid circuit, the near-end output $\tilde{d}(t)$ contains contributions from both $s(t)$ and an attenuated 'echo' term from $\tilde{r}(t)$. The echo path through the hybrid is denoted by f_2. The far-end path through the hybrid is denoted by f_3. We will assume that f_1, f_2, f_3, and f_4 are all "zero-memory" nonlinearities, i.e., systems whose outputs at a point in time are nonlinear functions of the inputs at that point in time. We will also assume that f_1, f_2, f_3, and f_4 exhibit no hysteresis.

The task of the echo canceller ψ is to estimate the echo component in $d[k]$ and cancel it in the analog domain. While it is possible to do the cancellation in the digital domain, nonlinearity in the required analog-to-digital converter at the signal $d[k]$ would create a nonlinear function of the sum of two signals, resulting in an intermodulation term that could not be cancelled.

A problem arises with such a system in that the attenuation of the near-end signal $\tilde{r}(t)$ through the hybrid path f_2 can be as low as 10 dB, while the transmission-line attenuation of the far-end signal $s(t)$ can be as high as 40 dB. Therefore, a canceller that achieves at least 60 dB near-end rejection is required. The nonlinearities in the data converters f_1 and f_4 and the nonlinearities f_2 and f_3 in the hybrid itself limit the performance of a linear echo canceller to about 60 dB with 1% differential nonlinearity [1]. Neglecting the effect of the far-end signal, the task of a nonlinear echo canceller ψ would be to realize the system $f_4^{-1} \circ \tilde{f}$, where $\tilde{f}$ is the cascade of linear and nonlinear systems f_1, H_1, f_2, and H_2. In the absence of hysteresis, it is clear that such a system has fading memory on any set of input sequences. In the following examples, we use both the perceptron filter and the recurrent network to approximate the system $f_4^{-1} \circ \tilde{f}$.

Example 2 (Perceptron Filter—finite-memory echo-path model): In this example, we demonstrate the performance of the perceptron filter using Algorithm 2 in identifying a finite-memory echo-path model given by the concatenation of systems f_1, H_1, f_2, and H_2 shown in Fig. 8. The DAC transfer function f_1 was modeled as a third-degree polynomial zero-memory nonlinearity of the form $f_1(x) = x^3$ as shown in Fig. 9. The hybrid transfer function f_2 was modeled as a zero-memory nonlinearity of the form $f_2(x) = 2\tan^{-1}(5x) + 0.25x$ as shown in Fig. 10. The linear transmit and receive filters H_1 and H_2 were modeled as finite-memory linear systems with finite unit-sample responses $h_1[n] = h_2[n] = 0.7(n+1)^{-1}$, $n = 0, 1, \ldots, 9$. The response of this system to a low-frequency triangular wave is shown in Fig. 11. The response of the system to a random step sequence is shown in Fig. 12. In the following experiments using Algorithm 2, both the

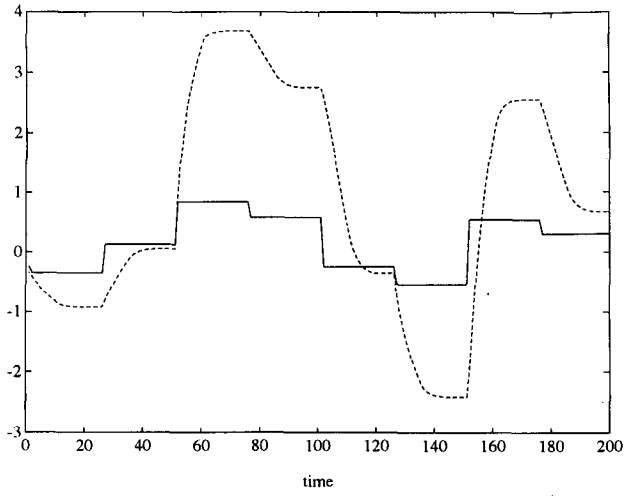


Fig. 12. Echo path model response to a random step input.

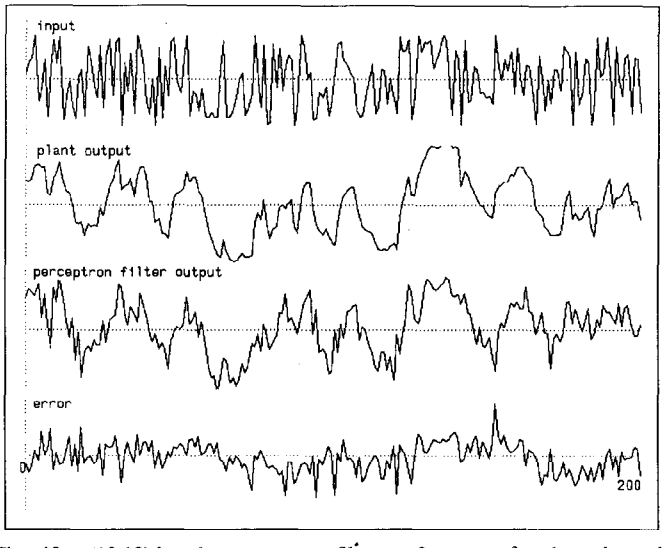


Fig. 13. (10,10)th-order perceptron filter performance for the echo-path model showing approximately 15 dB noise rejection.

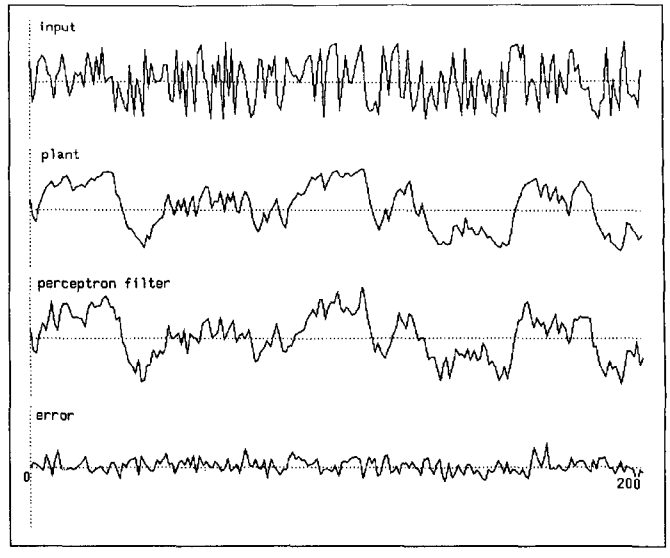


Fig. 14. (20,50)th-order perceptron filter performance for the echo-path model showing approximately 21 dB noise rejection.

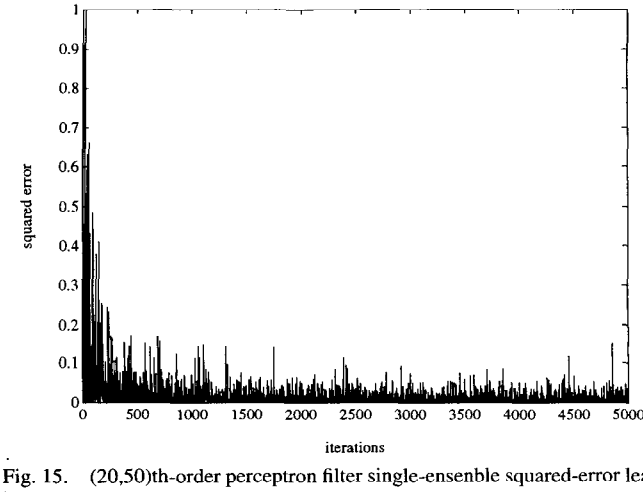


Fig. 15. (20,50)th-order perceptron filter single-ensenble squared-error learning curve.

system and the perceptron filter were fed a zero-mean i.i.d. random sequence uniformly distributed on the interval $[-1, 1]$. The sequence of learning constants $\{\beta[k]\}$ was chosen to be $\beta[k] = 0.025(0.1)^{k/5000}$ for all $k \geq 0$.

Fig. 13 shows the performance of the $(10, 10)$th-order randomly-initialized perceptron filter after 5000 iterations. The uppermost trace is the input signal, the second trace is the output of the echo-path model, the third trace is the output of the perceptron filter, and the fouth trace is the output error between the echo-path model and the perceptron filter. This example shows approximately 15 dB of noise rejection. Figs. 14 and 15 show the performance of a (20,50)th-order perceptron filter and its single-ensemble squared error learning curve, respectively, with approximately 21 dB of noise rejection.

Example 3 (Recurrent Network—infinite-memory echo-path model): In this example, we demonstrate the performance of the recurrent network using Algorithm 1 in identifying an infinite-memory echo-path model given by the concatenation of systems f_1, H_1, f_2, and H_2 shown in Fig. 8. The DAC transfer function f_1 was modeled as a third-degree polynomial zero-memory nonlinearity of the form $f_1(x) = x^3$ as shown in Fig. 9. The hybrid transfer function f_2 is modeled as a zero-memory nonlinearity of the form $f_2(x) = 2\tan^{-1}(5x) + 0.25x$ as shown in Fig. 10. The linear transmit and receive filters H_1 and H_2 are modeled as infinite-memory linear systems with transfer functions $H_1(s) = (s + 0.75)^{-1}$ and $H_2(s) = 0.2(s + 0.75)^{-1}$. The response of this system to a low-frequency triangular wave is shown in Fig. 16. The response of the system to a random step sequence is shown in Fig. 17. Algorithm 1 was used to adapt a recurrent network using a zero-mean i.i.d. uniformly distributed random input sequence on the interval $[-1, 1]$.

Fig. 18 shows the performance of a 3rd-order randomly-initialized recurrent network during learning of the infinite-memory echo-path model; it shows approximately 29 dB of noise rejection. Fig. 19 shows the performance of a 5th-order

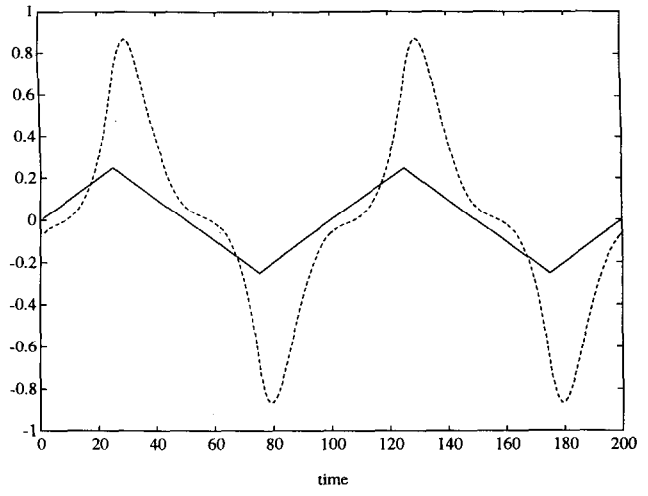


Fig. 16. (20,50)th-order perceptron filter single-ensemble squared-error learning curve.

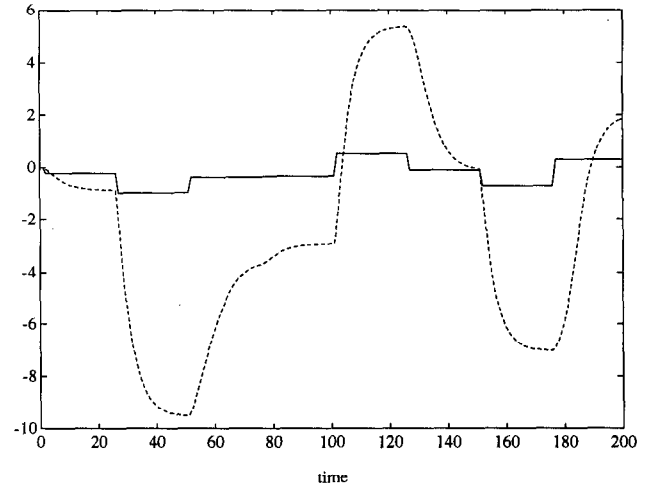


Fig. 17. Echo path model response to a triangular wave input.

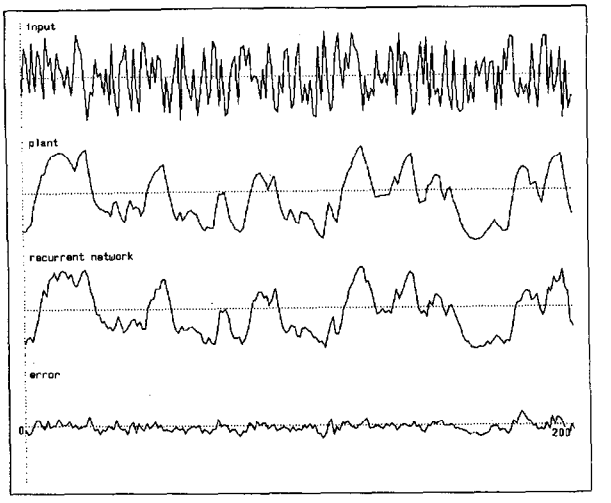


Fig. 18. Echo path model response to a random step input.

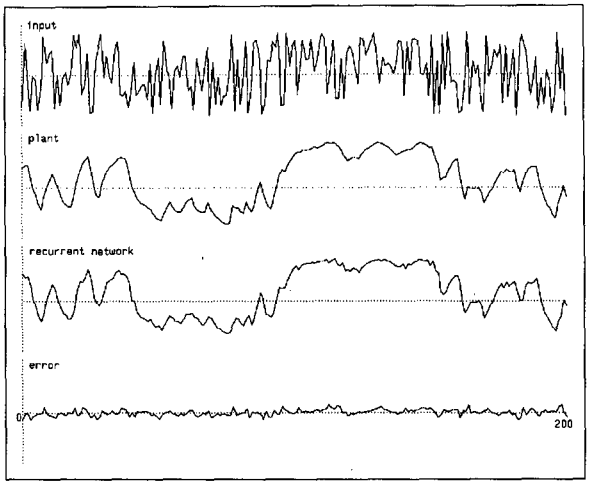


Fig. 19. Third-order recurrent network performance during learning of the echo-path model.

recurrent network during learning of the echo-path model with approximately 36 dB of noise rejection. Fig. 20 shows the 5th-order recurrent network single-ensemble squared-error learning curve.

VI. Conclusion

In this paper, we have considered the problem of identifying nonlinear systems. We defined "identification" as the process of creating a mathematical model of a system based on observation of the system's behavior in response to an input. An exact model of a physical system is often both impractical and unnecessary; an inexact model that uniformly approximates the system arbitrarily closely over a set of bounded inputs suffices in many practical applications. In this regard, it is convenient to choose a parametric model that realizes a set of systems that is uniformly dense in the set of systems to be modeled. A parametric model of a given order realizes over a bounded subset of the parameter space, a closed set of systems that has the best approximation property in the sense that there is a system in this set that is a best approximation to the system to be modeled. The identification problem becomes one of finding the parameter vector that specifies a best approximation. The "sufficient approximation hypothesis" states that, for a sufficiently-close approximation, the model-system output error is a zero-mean i.i.d. process uncorrelated with the input. Based on this hypothesis, we may recast the identification problem into one of estimating the parameter vector of the model based on "noisy" observations of the model output.

In general, feedback systems cannot be approximated in the above sense. One common class of systems that can indeed be approximated is the class of fading-memory systems. Fading-memory systems are systems that "forget" their inputs in a well-defined manner over time. It was shown that such systems can be approximated arbitrarily closely simply by approximating sufficiently closely either the external or internal representation of the system. In other words, approximating fading-memory systems is tantamount to approximating continuous functions.

In this regard, the perceptron is a parametric model that realizes a set of continuous functions that is uniformly dense in

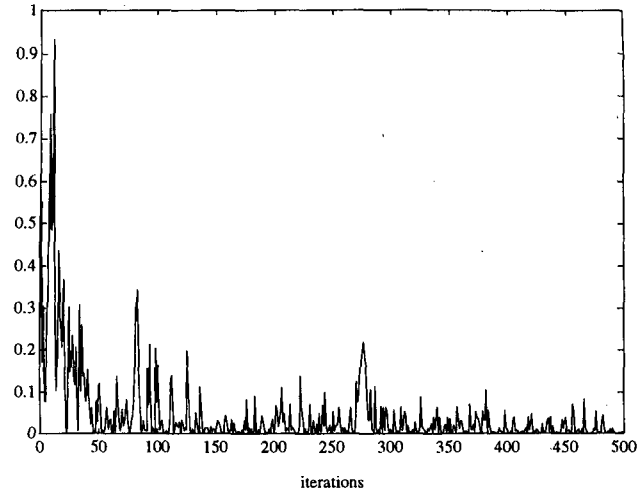


Fig. 20. Fifth-order recurrent network performance during learning of the echo-path model.

the set of all continuous functions. The perceptron is preferable to other parametric models such as the polynomial model, in that it is composed of saturating nonlinearities which more closely resemble the physical processes in most real systems. Approximating the external and internal representations of a fading-memory system over the set of continuous functions realized by the perceptron results in finite and infinite-memory models, respectively, called the perceptron filter and the recurrent network, respectively. These models possess attractive qualities in terms of analytical tractability, and simplicity. We envision the perceptron filter and the recurrent network being used for approximating fading-memory systems in a manner similar to the use of FIR and IIR filters to approximate linear systems: For systems with finite or rapidly-fading memory, a perceptron filter may suffice as a good approximation; for systems with slowly-fading memory, a recurrent network may be more efficient in terms of the order of the model required to effect a sufficient approximation.

We have suggested a recursive parameter-estimation algorithm for the perceptron filter and for the recurrent network based on a conditional mean-square cost function, and we have shown an application for echo-cancellation in telephone systems. The echo-path model, composed of the concatenation of linear systems and zero-memory nonlinearities, is a good example of a nonlinear fading-memory system. As expected, the perceptron filter, being a finite-memory model, performs well in approximating an echo-path model with rapidly-fading memory or finite memory; however, as the length the memory increases, the order of perceptron filter needed to effect a good approximation increases rapidly. Furthermore, also as expected the recurrent network performs well in approximating echo-path models with infinite memory.

APPENDIX

In this appendix, we show how the recurrent network of the form in (10) results by approximating with a perceptron the continuous mappings $f : X \times U \to X$ and $h : X \to Y$ in the system $\Sigma = (U, X, Y, f, h, x_0)$ whose behavior is given by the equations $x[0] = x_0$ and

$$x[k+1] = f(x[k], u[k])$$
$$y[k] = h(x[k])$$

for all $k \in \mathbb{N}$, and where the input space U, the state space X, and the output space Y are real vector spaces. Assume that the system Σ has fading memory on a subset K_u^* of U^* for some compact subset K_u of U. Let K_x be the set of reachable states for sequences in K_u^*, *i.e.*, $K_x = f^*(x_0, K_u^*)$. Since $f^*(x_0, .)$ is continuous, K_x is compact. We wish to uniformly approximate system Σ by uniformly approximating both f and h on $K_x \times K_u$ and K_x, respectively, using the perceptron.

First, let us consider the case where the output function h is the identity map, i.e., $h(x) = x$ for all $x \in X$. Let $(X \times U, V, X, D, C, \boldsymbol{\phi})$ be a perceptron, with hidden-layer space V, that realizes the function $\bar{f} : X \times U \to X$ where $\bar{f} = C \circ \boldsymbol{\phi} \circ D$ is such that $\bar{f}$ uniformly approximates f sufficiently closely on $K_x \times K_u$. Assume that $\dim V > \dim X$. Let $\bar{\Sigma} = (U, X, Y, \bar{f}, h, x_0)$ be the system associated with the mapping $\bar{f}$.

Consider the affine operator $D : X \times U \to V$. Clearly, there exists an affine operator $E : U \to V$ and a linear operator $F : X \to V$ such that $D(x, u) = Fx + Eu$ for all $x \in X$ and $u \in U$. System $\bar{\Sigma}$ has dynamics $x[0] = x_0$ and

$$x[k+1] = C\,\boldsymbol{\phi}(Fx[k] + Eu[k])$$
$$y[k] = x[k] \tag{24}$$

for all $k \in \mathbb{N}$. Now consider the affine operator $C : V \to X$; assume that C is surjective. Then there exists an affine mapping $C^* : X \to V$ such that $CC^*x = x$, for all $x \in X$. The range of C^* is an affine subspace V^* of V, where $\dim V^* = \dim X$. With the aid of the mapping C^*, we may extend the state space X of system $\bar{\Sigma}$ to the entire hidden-layer space V such that $v[0] = C^*x[0]$ and

$$v[k+1] = C^*C\,\boldsymbol{\phi}(FCv[k] + Eu[k])$$
$$y[k] = Cv[k] \tag{25}$$

for all $k \in \mathbb{N}$. Note that C^*C is a projection operator (not necessarily orthogonal) onto the subspace V^*. Note also that $\ker C^*C = \ker C$. Since $V = V^* \oplus \ker C$, every $p \in V$ can be uniquely represented as $p = v + z$ for some $v \in V^*$ and $z \in \ker C$. Hence, from (25) it follows that

$$v[k+1] + z[k+1] = \boldsymbol{\phi}(FC(v[k] + z[k]) + Eu[k])\,,$$
$$y[k] = C(v[k] + z[k]) \tag{26}$$

for $z[k+1], z[k] \in \ker C$. Finally, let $A : V \to V$ and $B : U \to V$ be linear and affine operators, respectively, such that $FCv + Eu = Av + Bu$ for all $v \in V$ and $u \in U$. Therefore, the system $\hat{\Sigma} = (U, V, Y, \hat{f}, \hat{h}, v_0)$, where $Cv_0 = x_0$, where $\hat{f}(v, u) = \boldsymbol{\phi}(Av + Bu)$ and $\hat{h}(v) = Cv$, with dynamics $v[0] = v_0$ and

$$v[k+1] = \boldsymbol{\phi}(Av[k] + Bu[k])\,, y[k] = Cv[k] \tag{27}$$

for all $k \in \mathbb{N}$, has a response map identical to that of system $\bar{\Sigma}$.

REFERENCES

[1] O. Agazzi, D. G, Messerschmitt, D. A. Hodges, "Nonlinear echo cancellation of data signals," *IEEE Trans. Commun.*, vol. COM-30, pp. 2421-2433, Nov. 1982.

[2] B. D. O. Anderson, "A simplified viewpoint of hyperstability," *IEEE Trans. Automat. Contr.*, vol. AC-13, pp. 292-294, June 1968.

[3] E. Biglieri and A. Gersho, "Adaptive cancellation of nonlinear intersymbol interference for voiceband data transmission," *IEEE J. Select. Areas Commun.*, vol. SAC-2, no. 5, pp. 765-777, Sept. 1984.

[4] S. Boyd and L. O. Chua, "Fading memory and the problem of approximating nonlinear operators with Volterra series," *IEEE Trans. Circuits Syst.*, vol. CAS-32, no. 11, pp. 1150–1161, Nov. 1985.

[5] B. D. Coleman and V. J. Mizel, "On the general theory of fading memory," *Arch. Rational Mech. Anal.*, vol. 29, pp. 18–31, 1968.

[6] G. Cybenko, "Approximation by superposition of a sigmoid function," *Mathemat. Control, Sig., Syst.*, vol. 2, pp. 303-314, 1989.

[7] K. Funahashi, "On the approximate realization of continuous mappings by neural networks," *Neural Networks*, vol. 2, pp. 183-192, 1989.

[8] K. Hornik, "Approximation capabilities of multilayer feedback networks," *Neural Networks*, vol. 4, pp. 251-257, 1991.

[9] ——, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359-366, 1989.

[10] T. Koh and E. J. Powers, "Second-order Volterra filtering and its application to nonlinear system identification," *IEEE Trans. Acoust., Speech, and Sig. Proc.*, vol. ASSP-33, no. 6, pp. 1445-1455, Dec. 1985.

[11] E. Kreyszig, *Introductory Functional Analysis with Applications*. New York: Wiley, 1979. .

[12] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*. Cambridge: MIT Press, 1983.

[13] M. B. Matthews, "An adaptive nonlinear filter structure," *Proc. IEEE Symp Circuits Syst.*, Singapore, pp. 694-697, June 1991.

[14] ——, "On the uniform approximation of nonlinear discrete-time fading-memory systems using neural network models," Dr. Sc. Tech. dissertation number 9635, Institute for Signal and Information Processing, Swiss Federal Institute of Techology, Zürich, Switzerland.

[15] ——, "Approximating nonlinear fading-memory operators using neural network models," *Circuits, Syst., Sig. Proc.*, vol. 12, no. 2, pp. 279-307, 1993.

[16] D. Mitra, "The absolute stability of high-order discrete-time systems utilizing the saturation nonlinearity," *IEEE Trans. Circuits Syst.*, vol. CAS-25, no. 6, pp. 365-370, June 1978.

[17] M. L. Minsky and S. A. Papert, *Perceptrons*. Cambridge: MIT Press, 1990.

[18] K. S. Narendra and K. Parthasarathy, "Identification and control of dynamical systems using neural networks," *IEEE Trans. Neural Networks*, vol. 1, no. 1, pp. 4–26, Mar. 1990.

[19] W. J. Rugh, *Nonlinear Systems: A Volterra/Wiener Approach*. Baltimore, MD: Johns Hopkins Press, 1981.

[20] I. W. Sandberg, "Approximate-finite memory and input-output maps," *IEEE Trans. Circuits Syst. – I*, vol. 39, no. 7, pp. 549–556, July 1992.

[21] V. Singh, "A generalized approach for the absolute stability of discrete-time systems utilizing the saturation nonlinearity, based on passivity properties," *IEEE Trans. Circuits Syst.*, vol. 37, no. 3, pp. 444-447, Mar. 1990.

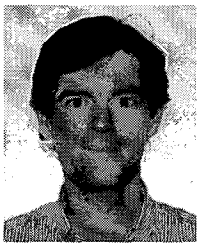
[22] E. Sontag, *Polynomial Response Maps*. New York: Academic, 1979.

[23] E. Sontag, "Realization theory of discrete-time nonlinear systems: Part I—the bounded case," *IEEE Trans. Circuits Syst.*, vol. CAS-26, no. 4, pp. 342-356, Apr. 1979.

[24] T. Wigren, "Recursive identification-based on the nonlinear Wiener model," Ph.D. dissertation, Uppsala University, 1990.

[25] G. Zames, "On the input-output stability of time-varying nonlinear feedback systems—part I: conditions derived using concepts of loop gain, conicity, and passivity," *IEEE Trans. Automat. Contr.*, vol. AC–11, no. 2, pp. 229-238, Apr. 1966.

[26] ——, "On the input-output stability of time-varying nonlinear feedback systems—part II: conditions involving circles in the frequency plane and sector nonlinearities," *IEEE Trans. Automat. Control*, vol. AC-11, no. 3, pp. 465-477, July 1966.



Michael B. Matthews received the B.Sc degree from the University of California, Irvine in 1981; the M.Sc. degree from the University of Washington, Seattle, in 1986; and the D.Sc. degree in 1992 from the Institute for Signal and Information Processing at the Swiss Federal Institute of Technology (ETH), Zürich, all in electrical engineering.

Since 1992, he has been with the Monterey Bay Aquarium Research Institute, Pacific Grove, CA, where his research is concentated on problems on problems in nonlinear estimation, sampling, and approximation theory as applied to submarine navigation and oceanographic data analysis.

George S. Moschytz (S'65–SM'77–F'78) received the E.E. Diploma in 1958 and the Ph.D. degree in optical scanning of code-addressed envelopes in 1962, both from the Swiss Federal Institute of Technology (ETH) in Zürich.

From 1960 to 1962 he was at RCA Laboratories in Zürich where he worked on envelope-delay measurement techniques for the transmission of color TV signals, and on conversion techniques for NTSC, PAL, and SECAM TV signals. From 1963 to 1972 he was at Bell Labs, Holmdel, NJ, where he developed and later supervised methods of designing hybrid-integrated active RC filters, phase-locked loops and oscillators, as well as silicon-integrated logic circuits and modulators for use in data transmission equipment and modems. Since 1973 he has been Professor for Network Theory and Signal Processing, and Director of the Laboratory for Signal and Information Processing, at the Swiss Federal institute of Technology (ETH), Zürich, Switzerland. He has authored many papers in the field of network theory, active and switched-capacitor filter and network design, and sensitivity theory, and holds several patents in these areas. He is author of *Linear Integrated Networks: Fundamentals* (Van Nostrand Reinhold, 1974), *Linear Integrated Networks: Design* (Van Nostrand Reinhold, 1975), co-author of *Active Filter Design Handbook* (Wiley, 1981), and Editor of *MOS Switched-Capacitor Filters: Analysis and Design* (IEEE Press, 1984). His present interests are digital, switched-capacitor, and adaptive filters, neural networks for signal processing, and the application of signal processing techniques to medical problems (electromyography and hearing aids).

Professor Moschytz is President of the IEEE Swiss Chapter on Digital Communication Systems and a member of the Swiss Electrotechnical Society. From 1981 to 1982, he was President of the Swiss Section of the IEEE. He is also the Swiss representative on the Commission for the Development of European Science and Technology (CODEST) in Brussels. He has held several terms at the Adcom of the IEEE Circuits and Systems Society as well as on the Editorial Board of the PROCEEDINGS OF THE IEEE and has been an Associate Editor of the *IEEE Circuits and Systems Magazine*. He is also an Associate Editor of several other technical journals. He is an elected member of the Swiss Academy of Engineering Sciences, winner of the Best Paper Award (for a paper on active filter design using tantalum thin-film technology) and a member of the Eta Kappa Nu Honor Society.