Complete Memory Structures for Approximating Nonlinear Discrete-Time Mappings

Bryan Waitsel Stiles, *Member, IEEE,* Irwin W. Sandberg, *Life Fellow, IEEE,* and Joydeep Ghosh

***Abstract*—This paper introduces a general structure that is capable of approximating input–output maps of nonlinear discrete-time systems. The structure is comprised of two stages, a dynamical stage followed by a memoryless nonlinear stage. A theorem is presented which gives a simple necessary and sufficient condition for a large set of structures of this form to be capable of modeling a wide class of nonlinear discrete-time systems. In particular, we introduce the concept of a "complete memory." A structure with a complete memory dynamical stage and a sufficiently powerful memoryless stage is shown to be capable of approximating arbitrarily well a wide class of continuous, causal, time-invariant, approximately-finite-memory mappings between discrete-time signal spaces. Furthermore we show that any bounded-input bounded-output, time-invariant, causal memory structure has such an approximation capability if and only if it is a complete memory. Several examples of linear and nonlinear complete memories are presented. The proposed complete memory structure provides a template for designing a wide variety of artificial neural networks for nonlinear spatiotemporal processing.**

***Index Terms*— Approximation theory, discrete-time systems, functional analysis, modeling, multidimensional systems, neural networks, nonlinear systems, universal approximators.**

I. Introduction

A LARGE volume of theoretical work has been performed regarding the properties and capabilities of *memoryless* approximators. Many feedforward networks have been shown to be universal approximators of static maps in the sense of being able to approximate arbitrarily well any continuous real-valued function on a bounded subset of $\Re^n$ [1]–[3]. Other specific functional forms, for example those based on Bernstein polynomials [4] which can be used to produce a constructive proof of the Weierstrass theorem [5], or on the Kolmogorov formulation [6], [7] have also been studied. Powerful convergence rate results for sigmoidal networks have been obtained [8]. All these results pertain to *static* maps, where the value of the desired output at any particular point is determined solely by the current input at that point. This is in contrast with *dynamic* systems where the desired output also depends on the past history and hence some notion of memory must be invoked.

Until recently, most of the work in approximating dynamic systems has been empirical in nature [9]–[12]. A notable exception is a series of studies, starting from the work of McCulloch and Pitts [13], showing that certain recurrent networks could simulate various finite state machines or push-down automata. For example, both fully recurrent networks and NARX models are *at least* as powerful as Turing machines, and in this sense serve as universal computation devices [14], [15]. Turing computable discrete-time systems form an important class. However, they are restricted in that both the inputs and outputs are formed from (discrete) symbols taken from a finite alphabet.

In this paper *we are concerned with approximating input–output maps of nonlinear discrete-time systems in which both inputs and outputs can be continuous valued.* In this context certain two-stage structures have recently been shown to be capable of approximating arbitrarily well a wide class of continuous, causal, time-invariant approximately finite memory mappings between discrete-time signals [16]–[18].[1] These networks consist of a temporal encoding stage followed by a nonlinear memoryless stage. The memoryless stage consists of a neural network that is a universal approximator of static maps, such as a multilayer perceptron (MLP) [1], radial basis function network [2], or ridge polynomial network [3]. A general block diagram of such a two-stage structure is shown in Fig. 1.

Two-stage networks are interesting models for dynamic systems because they are typically much easier to train than recurrent networks, and are less sensitive to initial conditions. Also, recurrent networks are susceptible to the long-term dependency problem when a gradient descent based training algorithm is used [21], though we note that certain recent results somewhat alleviate this problem [22], [23]. The approximation results on two-stage networks are important, because when attempting to model an unknown system, often only a general knowledge of the system's characteristics (causal, time-invariant, etc.) is available. Based upon these characteristics, one must choose a structure that is capable of modeling the system. General approximation results such as [16]–[18], and the results in this paper are necessary to determine which structures have this capability. Until now, the specific structures for which this approximation ability have been established have contained linear temporal encoding stages. The main theorem in [16] does not restrict itself to linear temporal encoding stages, but the examples of specific structures to which this theorem has been applied are linear.

In this paper we determine necessary and sufficient properties of the temporal encoding stage (See Fig. 1) needed for such approximation capabilities. The resulting structures

Manuscript received April 14, 1996; revised January 22, 1997 and June 2, 1997. This work was supported in part by NSF Grant ECS 9307632 and ONR Contract N00014-92C-0232. B. W. Stiles was also supported by the Du Pont Graduate Fellowship in Electrical Engineering.

The authors are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712-1084 USA.

Publisher Item Identifier S 1045-9227(97)07520-6.

[1] See [19] and [20] for other results concerning the approximation of functionals.

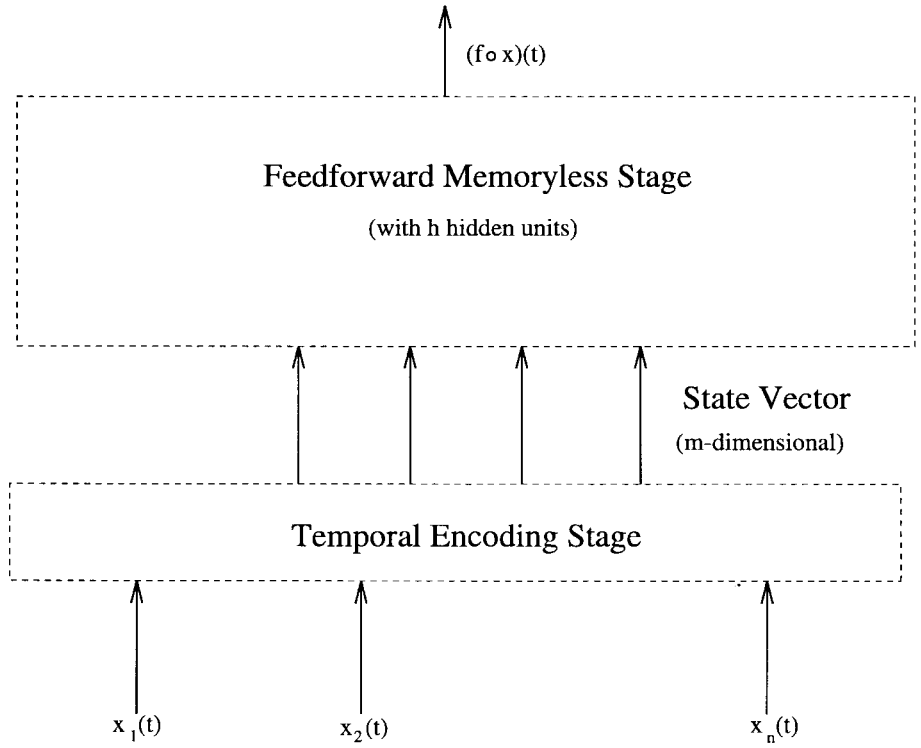

Fig. 1. Diagram of a generic two-stage structure for modeling discrete-time systems.

include examples of networks with nonlinear temporal encoding stages. Nonlinear temporal encoding schemes allow a richer variety of designs including several that are biologically plausible [24] and/or more efficient for certain applications [25]. In fact networks with linear temporal encoding stages are inappropriate for some problems because of a forced tradeoff between memory depth and memory resolution [26]. Certain nonlinear temporal encoders can avoid this problem, and this paper sets the framework for their design [27].

The next section summarizes the known results on properties of networks describable by Fig. 1. In Section III, we discuss structures in which the temporal encoding stage consists of functions which are elements of what we call a *complete memory*, and we demonstrate that such structures are capable of approximating arbitrarily well a wide class of continuous, causal, time-invariant, approximately-finite memory discrete-time systems. Additionally we exhibit a necessary condition such structures must satisfy in order to have this universal approximation capability. In Section IV we describe examples of sets of linear and nonlinear functions which are complete memories. One of these examples is the set of *habituation functions* which was used to generate the empirical results in [25]. Another example is the set of pattern search memory units which was used to generate the results in [27]. Section V discusses the contributions of this paper.

II. Two-Stage Dynamic Networks

In order to understand the history behind the proposed approach it is useful to examine previous work on two-stage networks involving linear temporal encoding mechanisms. A large of amount of theory in this area is given in [16]–[18]. These works include a proof of the universal approximation ability for time delay neural networks (TDNN's). TDNN's are simple two-stage architectures that use a tapped delay line to encode temporal information and an MLP as a feedforward stage. Under the weak assumption that the input set is uniformly bounded, it was shown that TDNN's can approximate arbitrarily well any continuous, causal, time-invariant, approximately finite memory mapping from one discrete-time sequence space to another [17]. In [16] a more general result of this kind is obtained by utilizing the concept of a fundamental set. Such a set is a family of mappings $\{F_\lambda\colon \lambda \in \Lambda\}$ associated with a given dynamic mapping G that satisfies certain properties with respect to G. In [16] it is shown that one can use such a fundamental set as a temporal encoding mechanism in order to approximate G. In the same paper, a structure is exhibited which can approximate arbitrarily well any G which is a continuous causal time-invariant, approximately finite memory mapping from one discrete-time sequence space to another. It was shown that such a G can be approximated arbitrarily well by a function F of the form

$$F(x) = \sum_l k_l \sigma\left[\sum_m \eta_{lm} h_m * x + \rho_l\right]. \tag{1}$$

Here h_m and x are functions of time, k_l, η_{lm}, and ρ_l are real constants, * denotes convolution, and σ is a sigmoid function. The overall approximation structure is an MLP feedforward stage, with linear operators used as a temporal encoding mechanism. It was proven that such an F can approximate G arbitrarily well by showing that a certain set of affine operators

is a fundamental set for G. Similar results for continuous time systems have also been obtained [16].

Subsequently it was shown that several different specific forms for the temporal encoding stage are sufficiently general in order to have the same approximation power [18]. One example of such a temporal encoding stage is similar to the gamma memory structure first studied by de Vries and Principe [10].

The various results mentioned above are "existence" results, and do not prescribe the complexity of the temporal encoding stage or feedforward network required (e.g. for TDNN's, the number of delays, and number of hidden units in the MLP) for a certain degree of approximation or a method for determining the network parameters.

In [25], a particular structure with a nonlinear temporal encoding stage, when compared with TDNN's, generated less complex classifiers with improved performance on several signal classification problems involving artificial Banzhaf sonograms. This empirical evidence along with the expectation that considering a more general family of structures may lead to improved performance on some problems, motivated us to develop structures with nonlinear temporal encoding stages. Later other nonlinear memory structures were shown to have theoretical advantages over linear memory structures as well [26], and showed superior performance in several experiments [27]. All these studies motivate this present paper which 1) precisely characterizes the desirable properties of the temporal encoding stage and 2) provides a guide to the design of nonlinear memory units.

III. Complete Memory Structure Theorems

Let R be the set of all mappings from the set Z_+ of nonnegative integers, to the set $\Re$ of real numbers. (Typically elements of Z_+ are used to reference discrete-time steps.) Let X be the subset of R for which $x \in X$ implies $x(t) \in [0,1]$ for all t.

Similarly with n any positive integer let R^n be the set of mappings from Z_+ to $\Re^n$, and let X^n be the subset of R^n for which $x \in X^n$ implies $x(t) \in [0,1]^n$ for all t. The delay operator T_β from R^n to R^n is defined by

$$(T_\beta \circ x)(t) = \begin{cases} \theta & \text{if } t < \beta \\ x(t-\beta) & \text{otherwise} \end{cases}$$

with θ the zero element in R^n. When dealing with operators such as T_β which operate on sequences, we adopt the notation $(T_\beta \circ x)(t)$ which should be read as T_β operating on x at time t. Moreover, if $x \in X^n$ then x_i denotes the element of X such that for all $t \in Z_+, x_i(t)$ is equal to ith component of $x(t)$.

Now we define what is precisely meant by the terms, *causal*, *time-invariant*, and *continuous*. A mapping M from X^n to R is *time-invariant* if for each nonnegative integer β and each $x \in X^n$

$$(M \circ T_\beta \circ x)(t) = \begin{cases} 0 & \text{if } t < \beta \\ (M \circ x)(t-\beta) & \text{otherwise.} \end{cases}$$

For each nonnegative integer a, let the truncation operator Q_a from R^n to R^n be defined by

$$(Q_a x)(t) = \begin{cases} \theta & \text{if } t > a \\ x(t) & \text{otherwise.} \end{cases}$$

Let C_a denote the intersection of the sets $[0,a]$ and Z_+. A mapping M from X^n to R is *causal* if $\forall a \in Z_+$ the statement $Q_a(x) = Q_a(y)$ implies $(M \circ x)(t) = (M \circ y)(t)$ for all $t \in C_a$. For a causal M, the value of the sequence $M(x)$ at any instant t is independent of the future values of x.

A mapping M from X^n to R is *continuous* if for each positive ϵ and any $x \in X^n$ there exists a positive δ such that $y \in X^n$ and $\|x(t) - y(t)\| < \delta$ for all t implies $|(M \circ x)(t) - (M \circ y)(t)| < \epsilon$ for all t. (Here $\|\cdot\|$ denotes the Euclidean norm on $\Re^n$.)

Theorems are presented below concerning the ability of a general family of structures to approximate arbitrarily well any continuous, causal, time-invariant mapping f from X^n to R. The structures can also be slightly modified in order to approximate functions on more general input domains. To see this, let a and b be any real numbers such that $a < b$. Let $X^n_{a,b}$ be the subset of R^n for which $x \in X^n_{a,b}$ implies $x(t) \in [a,b]^n$ for all t. Let u be the element of $\Re^n$ with all its components equal to one, and s be the function from $X^n_{a,b}$ to X^n defined by

$$(s \circ x)(t) = \frac{x(t) - au}{b-a}. \tag{2}$$

For any function g from $X^n_{a,b}$ to R, there is a unique f from X^n to R such that

$$(g \circ x)(t) = (f \circ s \circ x)(t) \tag{3}$$

for all x and t. Clearly, if g is continuous, causal, and time-invariant so is f. If a function $\tilde{f}$ exists that approximates f with tolerance ϵ at time t in the sense that

$$\|(\tilde{f} \circ x)(t) - (f \circ x)(t)\| < \epsilon \tag{4}$$

for all $x \in X^n$ then $\tilde{g} = (\tilde{f} \circ s)$ similarly approximates g so that

$$\|(\tilde{g} \circ x)(t) - (g \circ x)(t)\| < \epsilon \tag{5}$$

for all $x \in X^n_{a,b}$.

Approximations of functions from X^n to R can thus be easily used to generate approximations of functions from $X^n_{a,b}$ to R. Therefore, for the sake of conciseness and simplicity of notation we focus attention on X^n. We show that certain structures can approximate arbitrarily well any continuous, causal, time-invariant function f from X^n to R.

The key to the proof is to show that the memory structure realized by the temporal encoding stage is a *complete memory*. Then, provided the feedforward stage is capable of approximating continuous functions from compact subsets of $\Re^n$ to $\Re$, the overall network will be capable of approximating f. Theorem 1 states that a two-layer neural network with an exponential activation function and a particular structure for processing the inputs can approximate f arbitrarily well.

Before presenting the theorem we define the concept of a *complete memory*.

Definition 1: Let B be a set of *continuous* mappings from X to R. B is a *complete memory* if it has the following four

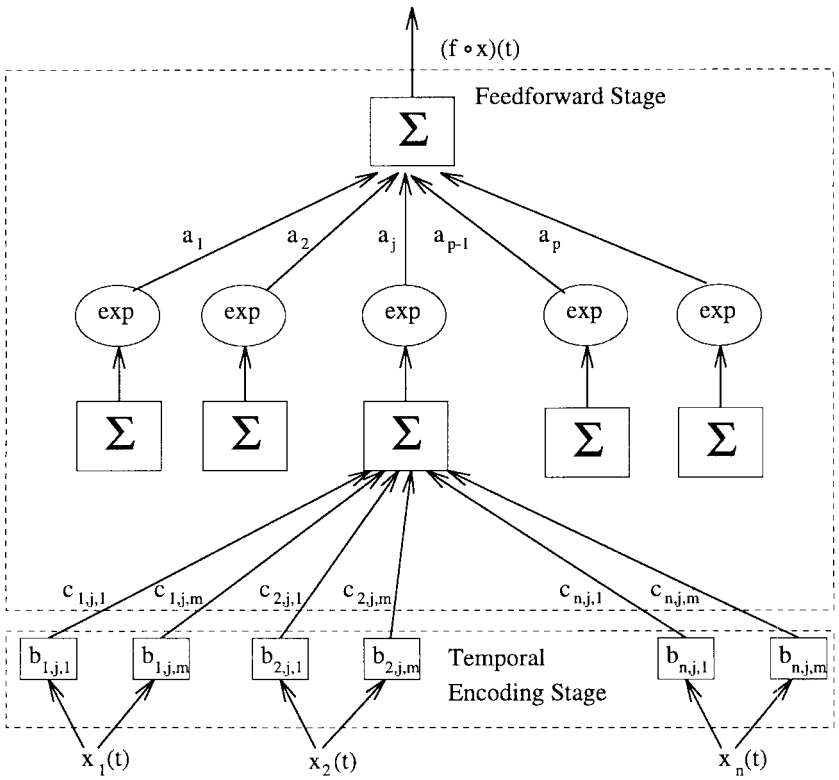

Fig. 2. Approximation structure in Theorem 1.

properties. First, there exist real numbers a and c such that $(b \circ x)(t) \in (a, c)$ for all $t \in Z^+, x \in X$, and $b \in B$. Second, for any $t \in Z_+$ and any t_1 such that $0 \leq t_1 \leq t$, the following is true. If x and y are elements of X and $x(t_1) \neq y(t_1)$, then there exists some $b \in B$ such that $(b \circ x)(t) \neq (b \circ y)(t)$. Third, if $b \in B$ then $(b \circ T_\beta \circ x)(t) = (b \circ x)(t - \beta)$ for all $t \in Z_+$, all $x \in X$ and any β such that $0 \leq \beta \leq t$. Fourth, every $b \in B$ is causal.

The following theorem shows the approximation ability of a structure comprised of a complete memory dynamical stage followed by a summation of exponential functions. This structure is shown to be capable of approximating any f within any given tolerance ϵ for any (arbitrarily long) period t_f. Later in the paper, a corollary is given which allows a more general form for the memoryless stage. An additional corollary shows that for approximately finite memory f, an approximation can be developed which is accurate for all time. The details of the structure to which Theorem 1 applies are illustrated in Fig. 2.

Theorem 1: Let f be a continuous, causal, time-invariant function from X^n to R. If B is a complete memory, then given any $\epsilon > 0$ and any positive integer t_f there exist real numbers a_j and c_{ijk} elements b_{ijk} of B, and positive integers p and m such that

$$\left| (f \circ x)(t) - \sum_{j=1}^{p} a_j \exp\left(\sum_{i=1}^{n} \sum_{k=1}^{m} c_{ijk} (b_{ijk} \circ x_i)(t) \right) \right| < \epsilon \tag{6}$$

for all $x \in X$ and all t such that $0 \leq t < t_f$.

The proof of this theorem is given in the Appendix.

It is important to notice that the input processing functions b_{ijk} in Theorem 1 depend on j. This means that different hidden units in the feedforward network may have different input values. This dependency is not necessary. One can show that for any approximation sum of the form described in Theorem 1, there is an equivalent network without this dependency. Such a network is illustrated in Fig. 3.

Corollary 1: Let g be an approximation sum of the form $g(x) = \Sigma_{j=1}^{p} a_j \exp(\Sigma_{i=1}^{n} \Sigma_{k=1}^{m} c_{ijk}(b_{ijk} \circ x_i))$. Then there is an h of the form

$$h(x) = \sum_{j=1}^{p} a_j \exp\left(\sum_{i=1}^{n} \sum_{d=1}^{M} w_{ijd} (s_{id} \circ x_i) \right) \tag{7}$$

with real numbers w_{ijd} (weights to the hidden units), a positive integer M, and elements s_{id} of B, such that $g(x) = h(x)$ for all $x \in X^n$.

Proof: The key to the proof is to relabel the collection $\{b_{ijk}\}$ as $\{s_{id}\}$ and use zero weights where necessary. Let $M = pm$. For each value of i, j, and k, let $d = (j-1)*m+k$ and let $s_{id} = b_{ijk}$. Observe that each s_{id} is uniquely defined in this manner. Set w_{ijd} similarly. For each value of i, j, and k, let $d = (j-1)*m+k$ and let $w_{ijd} = c_{ijk}$. This time some w_{ijd} terms will remain undefined. Set those terms to zero. Since k varies between one and m, there are only m nonzero w_{ijd} values for each (i, j) pair. Those m values are the c_{ijk} values in the definition of g. By our choice of w_{ijd} and s_{id} we have

$$h(x) = \sum_{j=1}^{p} a_j \exp\left(\sum_{i=1}^{n} \sum_{k=1}^{m} c_{ijk} (b_{ijk} \circ x_i) \right). \tag{8}$$

Clearly, $h = g$, and the proof is complete. ■

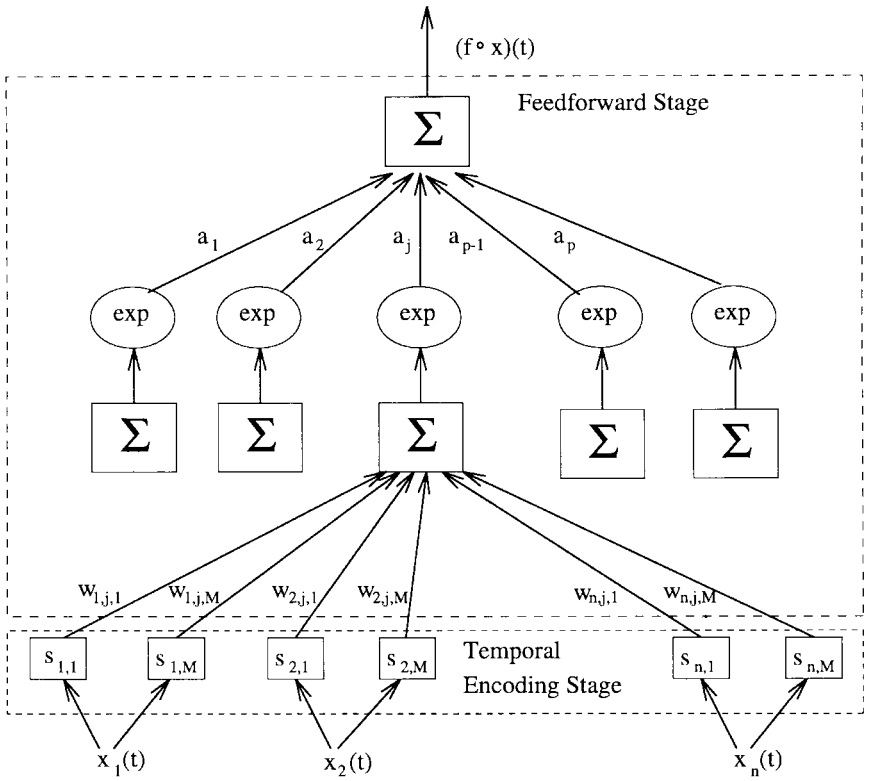

Fig. 3. Approximation structure in corollary 1. The outputs from a single set of temporal encoding functions are presented simultaneously to all the hidden units in the feedforward stage.

Until now we have considered a very specific memoryless stage, a summation of exponential functions. Corollary 2 allows the feedforward stage to be generalized to any structure capable of uniformly approximating real-valued continuous functions defined on compact (closed and bounded) subsets of real finite-dimensional vectors. Examples of such feedforward structures include MLP's, RBF's, and polynomials. This generalized structure is illustrated in Fig. 4.

Before presenting the corollary, it is necessary to explain some of the notation to be used. Let $\{s_{id}\}$ be a set of mappings from Z_+ to $\Re$ for each integer $i, 1 \le i \le n$, and each integer $d, 1 \le d \le m$. Let $R^{n\times m}$ denote the set of $n \times m$ real matrix-valued functions defined on Z_+. Let V_{nm} from $R^{n\times m}$ to R^{nm} be defined by

$$(V_{nm} \circ s)(t) = [s_{11}(t), s_{12}(t), \cdots, s_{21}(t), s_{22}(t), \cdots s_{nm}(t)]. \tag{9}$$

For each positive integer k let E_k be the set of all mappings from $\Re^k$ to $\Re$, and let D_k be a subset of E_k that satisfies the following universal approximation condition. For any continuous function u from $\Re^k$ to $\Re$, any $\delta > 0$, and any compact $U \subset \Re^k$, there exists $\gamma \in D_k$ such that $|\gamma(y) - u(y)| < \delta$ for all $y \in U$. Let E be the union of sets E_k for all positive integers k. Similarly let D be the union of all sets D_k.

For any set of functions $\{s_{id}\}$ from R to R for integers $i \in [1, n]$ and $d \in [1, m]$, let s denote the function from R^n to $R^{n\times m}$ defined by

$$(s \circ x)_{id} = (s_{id} \circ x_i)$$

for $x \in R^n$.

Corollary 2: Let $\epsilon > 0$. Let h be an approximation function of the form

$$h(x) = \sum_{j=1}^{p} a_j \exp\left(\sum_{i=1}^{n}\sum_{d=1}^{m} w_{ijd}(s_{id} \circ x_i)\right). \tag{10}$$

Given any such h there exists $\gamma \in D_{nm}$ such that

$$|(h \circ x)(t) - \gamma(V_{nm}(s \circ x)(t))| < \epsilon \tag{11}$$

for all $t \in Z_+$ and all $x \in X^n$.

Proof: Let $\epsilon > 0$. Let q be a function from $\Re^{nm}$ to $\Re$ defined by

$$q(y) = \sum_{j=1}^{p} a_j \exp\left(\sum_{i=1}^{n}\sum_{d=1}^{m} w_{ijd} y_{m(i-1)+d}\right). \tag{12}$$

Observe that q is continuous and that $h(x) = q(V_{nm}(s \circ x))$. Recall that from the first property of a complete memory, there exist real numbers μ and β such that $(s_{id} \circ x_i)(t) \in [\mu, \beta]$ for all $t \in Z_+$ and $x \in X^n$. Therefore $V_{nm}(s \circ x)(t) \in [\mu, \beta]^{nm}$. Since q is continuous and $[\mu, \beta]^{nm}$ is a compact set, there exists $\gamma \in D_{nm}$ such that

$$|\gamma(y) - q(y)| < \epsilon \tag{13}$$

for all $y \in [\mu, \beta]^{nm}$. Since $V_{nm}(s \circ x)(t) \in [\mu, \beta]^{nm}$

$$|(h \circ x)(t) - \gamma(V_{nm}(s \circ x)(t))| < \epsilon \tag{14}$$

for all $t \in Z_+$ and $x \in X^n$. This completes the proof.

We have now shown that any feedforward stage structure which satisfies the universal approximation condition placed

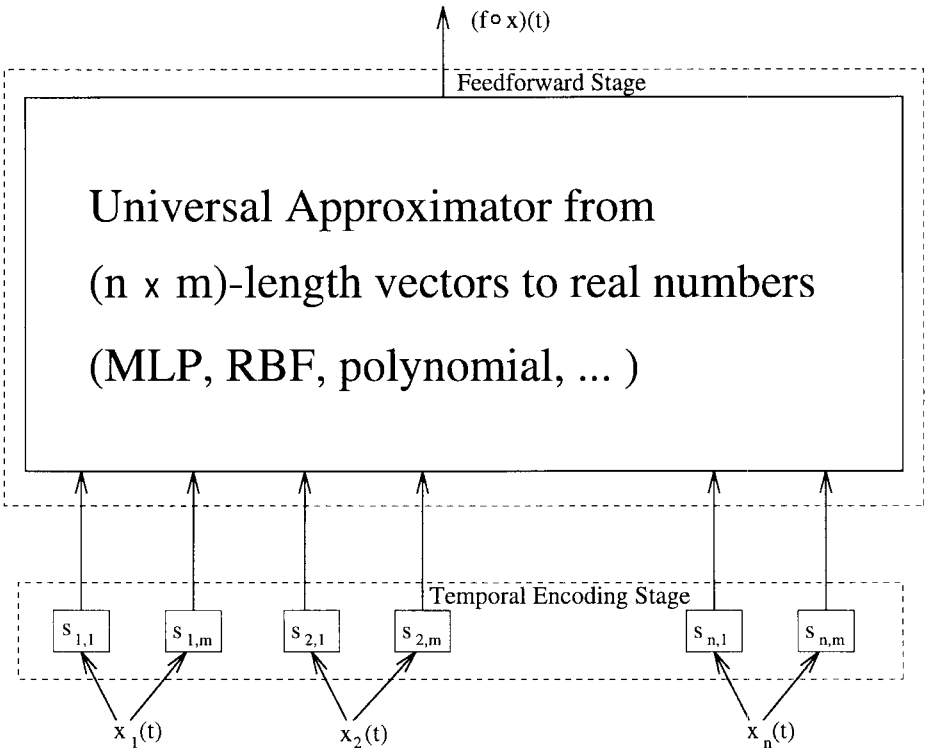

Fig. 4. General approximation structure in Corollary 2.

upon D can be used in obtaining an approximation result similar to Theorem 1. There are a number of structures which have been shown to satisfy this universal approximation condition. One of the most commonly used is an MLP with a single hidden layer [1]. Additional candidates include multivariable polynomials, lattice functions, and RBF networks [28], [2].

So far, we have considered approximations which are valid for some finite but arbitrarily long period of time t_f. If we make the assumption that the function f to be approximated has approximately finite memory, then we can show that f can be approximated arbitrarily well for all $t \in Z_+$. First we will define what is meant by approximately finite memory.

Let $W_{k,p}$ be the mapping from R^n to R^n defined by

$$(W_{k,p} \circ x)(t) = \begin{cases} \theta & \text{if } t < k-p \text{ or } t > k \\ x(t) & \text{otherwise.} \end{cases} \tag{15}$$

We say that a function f from X^n to R has approximately finite memory on X^n if for each $\epsilon > 0$ there exists a positive integer p such that

$$|(f \circ x)(t) - (f \circ W_{t,p} \circ x)(t)| < \epsilon \tag{16}$$

for all $x \in X^n$ and all $t \in Z_+$ [16].

Corollary 3: Let $\epsilon > 0$. Let f be a continuous, causal, time-invariant function from X^n to R that has approximately finite memory on X^n. Let B be a complete memory. Given these conditions there exist positive integers m and $k, \gamma \in D_{nm}$, and elements s_{id} of B for all integers $i \in [1,n]$ and $d \in [1,m]$, such that

$$|(f \circ x)(t) - \gamma(V_{nm}(s \circ W_{t,k} \circ x)(t))| < \epsilon \tag{17}$$

for all $x \in X$ and $t \in Z_+$.

The proof of this corollary is given in the Appendix.

We have now shown that a two-stage network which includes a temporal encoding stage is sufficient for approximating a wide range of discrete-time systems. At this point, it is advantageous to consider which properties of a complete memory are *necessary* to achieve an arbitrarily good approximation. In the following two corollaries we show that the second property of a complete memory is necessarily a property of the temporal encoding stage of any two-stage network which has the approximation power of the structure in Theorem 1 or Corollary 3. A related result is can be found in Theorem 4 of [18].

Corollary 4: Let S be a set of functions from X to R. Let K be a set of functions from X^n to R of the form

$$k(x) = \gamma(V_{nm}(s \circ x)) \tag{18}$$

in which $m \in N, \gamma \in E_{nm}$, and $s_{id} \in S$. The set S *must* satisfy the second property of a complete memory if K has the following property. For any continuous causal time-invariant mapping f from X^n to R, any $t_f \in Z_+$, and any positive ϵ there exists $k \in K$ such that $|(f \circ x)(t) - (k \circ x)(t)| < \epsilon$ for all $t < t_f$, and $x \in X^n$.

Proof: By way of contradiction, assume that S does not satisfy the second property of a complete memory. This means that we can choose y and z in X, and j and c in Z_+ such that $j \leq c, y(j) \neq z(j)$, and $(s \circ z)(c) = (s \circ y)(c)$ for all $s \in S$. Let t_f be greater than c. Let ϵ be positive and less than $\frac{1}{2}|y(j) - z(j)|$ Let the function f be defined by

$$(f \circ x)(t) = \begin{cases} 0 & \text{if } t < c-j \\ x_1(t-c+j) & \text{otherwise.} \end{cases} \tag{19}$$

The function f is causal because $c \geq j$. It is also clearly continuous and time-invariant. Let x and $\tilde{x}$ be elements of X^n

such that $x_1 = y, \tilde{x}_1 = z$, and x_i, and $\tilde{x}_i$ are the zero function for all $i \neq 1$. By the hypothesis of the corollary, we have

$$|(f \circ x)(c) - (k \circ x)(c)| < \epsilon \tag{20}$$

$$|(f \circ \tilde{x})(c) - (k \circ \tilde{x})(c)| < \epsilon \tag{21}$$

for some $m \in N, \gamma \in E_{nm}, s_{id} \in S$ and $k \in K$ such that $k(x) = \gamma(V_{nm}(s \circ x))$.

However by our choice of x and $\tilde{x}, (s_{id} \circ x)(c) = (s_{id} \circ \tilde{x})(c)$ for all $s_{id} \in S$, and therefore $(k \circ x)(c) = (k \circ \tilde{x})(c)$ for all $k \in K$. Since $|(f \circ x)(c) - (f \circ \tilde{x})(c)| = |y(j) - z(j)| > 2\epsilon$, the assumption that S does not satisfy the second property of a complete memory contradicts the hypothesis of the corollary. Therefore S must satisfy the second property of a complete memory and the proof is complete.

Now we have shown that a temporal encoding stage which satisfies the second property of a complete memory is necessary in any two-stage network with the approximation capability of the structure in Theorem 1. Similarly we can show that this property is necessary for any two-stage network which has the approximation capability of the structure described in Corollary 3.

Corollary 5: Let S be a set of functions from X to R. Let K be a set of functions from X^n to R of the form

$$(k \circ x)(t) = \gamma(V_{nm}(s \circ x)(t)) \tag{22}$$

in which m is a positive integer, $\gamma \in E_{nm}$, and $s_{id} \in S$.

The set S *must* satisfy the second property of a complete memory if K has the following property. For any continuous causal time-invariant approximately finite memory mapping f from X^n to R, and any positive ϵ there exists $k \in K$ and positive integer p such that $|(f \circ x)(t) - (k \circ W_{t,p} \circ x)(t)| < \epsilon$ for all $t \in Z_+$ and $x \in X^n$.

The proof of this corollary is in the Appendix.

To summarize, Theorem 1 shows that a particular structure which consists of elements of a complete memory followed by a feedforward network with exponential activation functions is capable of approximating arbitrarily well for an arbitrarily long period of time any continuous, causal, time-invariant mapping from X^n to R. Corollary 1 shows that a structure in which the same inputs are presented to each of the hidden nodes in the feedforward network is sufficient to achieve the approximation result. Corollary 2, establishes that the structure of the feedforward stage required for the result can be generalized to any set of functions D from real vectors to $\Re$ which is a universal approximator. For example, an MLP, RBF, lattice function, or polynomial feedforward stage would be sufficient. In Corollary 3, we show that a mapping f can be approximated arbitrarily well over all time if in addition to the previous requirements it has approximately finite memory on X^n. The structure used to perform this approximation is identical to the general structure discussed in Corollary 3 with the exception that a windowing function is applied to the inputs. Finally in Corollaries 4 and 5, we show that any two-stage network which has the approximation capability of the structures in Theorem 1 or Corollary 3 must have a temporal encoding stage that satisfies the second property of a complete memory.

IV. Examples of Complete Memory Structures

In this section, examples of complete memories are presented. These complete memories can be used to implement a temporal encoding stage for the structures presented in the previous section.

A. Linear Examples

First we will discuss linear temporal encoding stages that are complete memories. In [18],[2] the concept of a basic set is described. A subset K of X is a basic set on X if given any positive integer $\alpha, \epsilon > 0$, and $x \in X$, there is an h in the set of finite linear combinations of elements of K such that

$$|x(t) - h(t)| < \epsilon \tag{23}$$

for $t = 0, 1, 2, \cdots, \alpha$. For a basic set K, let S_K denote the set of convolutions with each element of K of the form

$$(s \circ x)(t) = (h * x)(t) \tag{24}$$

where $h \in K$.

It is clear that the elements of S_K satisfy the third and fourth properties of a complete memory and are continuous. It is a specific case of Corollary 1 in [18], that a particular two-stage network which uses any such S_K as a temporal encoding stage satisfies the hypothesis of Corollary 5. This implies that such an S_K satisfies the second property of a complete memory. Therefore any S_K generated by a basic set K is a complete memory if it satisfies the first property of a complete memory that there exist real numbers α and β such that $(s \circ x)(t) \in (\alpha, \beta)$ for all $s \in S_K, x \in X$, and $t \in Z_+$.

Using this relationship between basic sets and complete memories, one can show that some commonly used linear temporal encoding stages are in fact complete memories. Let h be the element of X for which $h(0) = 1$ and $h(i) = 0$ for all $i \neq 0$. Let $h_j = (T_j \circ h)$ for each $j \in Z_+$. As a specific case of Example 3 in [18], the set $P = \{h_j : j \in Z_+\}$ is a basic set. One sufficient condition for which a basic set K gives rise to a complete memory S_K is as follows. For all $t \in Z_+$, and all $h \in K$,

$$\sum_{\ell=0}^{t} |h(\ell)| < c \tag{25}$$

for some real number c. Clearly the S_K generated by a basic set K which satisfies this inequality must satisfy the first property of a complete memory, and therefore such an S_K must be a complete memory. Since each $h \in P$ satisfies (25) for $c = 2, S_P$ is a complete memory. The two-stage network structure which uses S_P as a temporal encoding stage and an MLP as a feedforward stage is the familiar time-delay neural network. Similarly, the temporal encoding stage of a focused gamma network [10] has been shown to be the resultant S_K for some basic set K [18]. The set of functions in the temporal encoding stage of a focused gamma network, B_μ, is defined

[2] In [18] the input space X is defined differently (the range of the input values are allowed to be negative) but is still uniformly bounded. This is a minor point because (as discussed in Section III) there is a simple invertible transformation (scaling and adding an offset) between the input space discussed in [18] and X as defined in this paper.

as follows for a particular real number μ. For all $x \in X$ and $t \in Z_+$.

$$(b_0 \circ x)(t) = x(t). \tag{26}$$

For $j > 0$ and $t = 0 (b_j \circ x)(t) = 0$. For $j > 0$ and $t > 0$

$$(b_j \circ x)(t) = (1-\mu)(b_j \circ x)(t-1) + \mu(b_{j-1} \circ x)(t-1) \tag{27}$$

$$B_\mu = \{b_j : j \in Z_+\}. \tag{28}$$

Since there is a basic set, K which generates $S_K = B_\mu$, in order to show that B_μ is a complete memory, it is sufficient to demonstrate that $(b_j \circ x)(t) \in [0,1]$ for all $j \in Z_+, x \in X$, and $t \in Z_+$. This fact is readily shown by mathematical induction for all $\mu \in (0,1)$. So the gamma memory is a complete memory for $\mu \in (0,1)$.[3] For the case when $\mu = 1$, the gamma memory degenerates to the temporal encoding stage of a TDNN.

B. Nonlinear Examples

We now present two examples of nonlinear temporal encoding stages that are complete memories. For other such examples see [25] and [27]. The first example is a set of functions based on the biologically observed habituation mechanism. This mechanism has been suggested to be one method used by biological neural systems, such as the mollusk *Aplysia*, to encode temporal information [30], [31]. In [25], the biological motivation behind this structure is discussed and empirical results on the classification of spatio-temporal signals are presented.

Theorem 2: Let $b_{-1} = 1$. A habituation function b is defined recursively by

$$(b \circ x)(0) = b_{-1} + \alpha\tau(1 - b_{-1}) - \tau b_{-1} x(0) \tag{29}$$

and

$$(b \circ x)(t) = (b \circ x)(t-1) + \alpha\tau(1 - (b \circ x)(t-1)) - \tau(b \circ x)(t-1)x(t) \tag{30}$$

in which α and $\tau \in \Re$ are such that $\alpha > 0, \tau > 0, \tau < 1$, and $\alpha\tau + \tau < 1$. Let B be the set of all such functions. B is a complete memory.

The proof of this theorem is given in the Appendix.

The set of habituation functions is a complete memory and therefore by Theorem 1 and Corollary 2, a structure such as that illustrated in Fig. 4 with habituation functions s_{id}, can approximate arbitrarily well any continuous, causal, time-invariant, approximately finite memory mapping from X^n to R.

Another example of a nonlinear complete memory is the set of pattern search memory units, P. The n-tapped delay line d_n is a mapping from X to X^n defined by

$$(d_n x)(t)_i = \begin{cases} x(t-n+i) & \text{if } t \geq n-i \\ 0 & \text{otherwise.} \end{cases}$$

[3] A very closely related result concerning gamma networks is given in [29].

P consists of functions p from X to R of the form

$$(px)(t) = \begin{cases} \max\{\alpha\beta\exp(-\gamma\|v\|^2) \\ \quad \beta\exp(-\gamma\|v - (d_n x)(t)\|^2)\} & \text{if } t = 0 \\ \max\{\alpha(px)(t-1) \\ \quad \beta\exp(-\gamma\|v - d_n x(t)\|^2)\} & \text{otherwise} \end{cases}$$

for all positive β, positive integers $n, v \in [-1,1]^n$, positive γ and α such that $0 < \alpha < 1$. Notation of the form $\{P|n,\beta\}$ is used to mean the subset of P for which the parameters n and β have some given constant value. Let p be an element of P. Whenever an n-length pattern in the input x is seen that closely matches the template v, a Gaussian response is produced which is maximal ($(px)(t) = \beta$) if an exact match is made. At each instant the current response is compared to a decayed (with decay rate α) version of a previous response. The output is chosen to be the maximum of the two. This output then decays over time and is compared with future responses. In this manner, p remembers an old template match until it decays to the point where a newer match supersedes it. The set P is useful for modeling systems in which a particular short-time pattern in the input must be remembered for a long period of time [27]. Examples of such systems include speech recognition and classification of marine biologics [32]. Such systems are often difficult for linear memory structures to model [26]. It is proved in [27] that for any acceptable constant values α, β, and γ, $\{P|\alpha,\beta,\gamma\}$ is a complete memory and therefore by Theorem 1 and Corollary 2, a structure such as that illustrated in Fig. 4 with PSM units, s_{id}, can approximate arbitrarily well any continuous, causal, time-invariant, approximately finite memory mapping from X^n to R even when the parameters α, β, and γ are assigned arbitrarily.

V. Discussion

In this paper, we described a general family of structures based upon the concept of a complete memory. Furthermore we have shown these structures to be quite powerful for approximating a wide class of nonlinear discrete-time systems. In particular, we have discussed two complete memory structures, the habituation based network and the pattern search memory network, which have nonlinear temporal encoding stages. Variants of the habituation based network have been used in a number of studies to classify sets of spatio-temporal signals, [25], [33]. The empirical results found in these studies suggest that habituation based networks compare favorably with TDNN's and focused gamma networks in terms of complexity and classification performance. Similar studies have been performed with pattern search memory networks which have been found to have both general theoretical advantages over linear memory structures [26], and empirical advantages when compared with TDNN's and focused gamma networks on spatiotemporal classification problems [27].

In addition to providing a proof of the approximation power of habituation based networks and pattern search memory networks, the complete memory concept also provides a useful tool for proving the approximation capability of other two stage networks. Since linear memory structures have been shown to be inefficient models for some systems [25], there

is sufficient motivation to study additional nonlinear memory two-stage structures. Such studies are aided by the results in this paper. The theorems presented here are all straightforward and can be applied without any special knowledge of functional analysis or other higher mathematics. (This is perhaps not true for the proofs of these theorems.) Whereas such tool theorems already exist for the case in which temporal encoding is performed by linear functionals [18], the theorems presented in this paper can also be used when the temporal encoding stages considered are nonlinear. In fact, several other nonlinear memory structures have already been found to be complete memories, and thus have the associated approximation power. Among these are cascaded habituation networks [25] and ordered pattern search networks [27]. Both of these structures have been applied to spatio-temporal classification problems in which they compared favorably to other commonly used approximators.

The method used in this paper to prove the approximation capability of two-stage networks is straightforward, as it is sufficient to show that the temporal encoding stage in question satisfies four simple properties. The second of these properties is necessary to yield the approximation results. The other three properties hold for each element of the complete memory. Obviously these three properties are not necessary. Consider the set P of functions consisting of the union of a complete memory and an additional function which does not satisfy the first, third, and fourth properties. Such a P when used as a temporal encoding stage would clearly produce the desired approximation results. However, the first property makes implementation of the resulting two-stage network on a digital computer feasible. Without this property, intermediate values within the network would generate overflow or underflow conditions. The fourth property of a complete memory, causality, is necessary for any physical implementation to be possible. The third property, a mild form of time-invariance, greatly simplifies the mathematical analysis of the networks. Additionally, since the functions which we are trying to approximate are all time-invariant, it seems somewhat quixotic to consider approximating them using time-varying functions. In conclusion, the complete memory temporal encoding stage is sufficient to achieve very powerful approximation results, and is general enough to include most practical two-stage structures that can perform such approximations. Therefore, complete memory theorems can be used as a tool by other researchers to determine the approximation power of novel two-stage network designs.

Much further research is possible in this area. One avenue of research is in the area of finding an approximation of a given function to a particular tolerance. To solve specific problems, it is not enough to state that such an approximation exists; one must also exhibit an algorithm to find it. This problem is difficult and has not even been solved in the general case for the commonly used *memoryless* structures (i.e., MLP, RBF, etc). In the event that it proves intractable, further research in useful heuristics for finding such approximations is also worthwhile. Such heuristics (gradient descent, etc.) have been commonly used previously in both static and dynamic structures [34]. Gradient descent, however, has been found to be problematic for dynamic systems with long term dependencies [21]. For two-stage networks in particular the coupling of the training of the feedforward and temporal encoding stages can lead to problems [25]. A heuristic that separates the training of the memoryless and memory stages has been used effectively for training pattern search networks in [27]. Finally, the difficult problem of analyzing the interaction between the complexity of the feedforward stage versus the temporal encoding stage for specific applications could also be investigated.

APPENDIX

Proof of Theorem 1: In order to prove the theorem we first prove the following Lemma.

Lemma 1: Let t_f be a positive integer. Then under the assumptions of Theorem 1 there exist positive integers p and m, real numbers a_j and c_{ijk}, and elements b_{ijk} of B such that

$$\left|(f \circ x)(t_f-1) - \sum_{j=1}^{p} a_j \cdot \exp\left(\sum_{i=1}^{n}\sum_{k=1}^{m} c_{ijk}(b_{ijk} \circ x_i)(t_f-1)\right)\right| < \epsilon \quad (31)$$

for all $x \in X^n$.

Proof: We first define the set of mappings $\{K_q\}$. For each positive integer q, let K_q be the mapping from $[0,1]^{t_f q}$ to X^q defined by

$$(K_q(u))(t) = \begin{cases} [u_{qt+1}, u_{qt+2}, \cdots, u_{qt+q}] & \text{if } t < t_f \\ [0,0,\cdots,0] & \text{otherwise.} \end{cases} \quad (32)$$

Further we also define a set of mappings $\{H_i\}$ from $\Re^{t_f n}$ to $\Re^{t_f}$. For each integer $i \in [1,n]$, H_i is defined by

$$H_i(u) = [u_i, u_{n+i}, u_{2n+i}, \cdots, u_{t_f n - n + i}]. \quad (33)$$

Observe that if the the components of $u \in \Re^{t_f n}$ are given by

$$u_{kn+i} = x_i(k), \qquad 0 \le k < t_f, \quad 1 \le i \le n$$

then

$$K_n(u) = Q_{t_f}(x), \quad K_1(H_i(u)) = Q_{t_f}(x_i).$$

This observation is important later in the proof. Let g be a mapping from $[0,1]^{t_f n}$ to $\Re$ defined by

$$g(u) = (f \circ K_n(u))(t_f - 1). \quad (34)$$

Observe that g is a continuous function on a compact metric space. (It is continuous because f is continuous.) Similarly, we define B^* to be the set of all functions b^* from $[0,1]^{t_f}$ to $\Re$ of the form $b^*(v) = (b \circ K_1(v))(t_f - 1)$ for each $b \in B$. Each $b^* \in B^*$ is continuous because the corresponding $b \in B$ is continuous.

Let S be the set of all functions s from $[0,1]^{t_f n}$ to $\Re$ of the form

$$s(u) = \sum_{j=1}^{p} a_j \exp\left(\sum_{i=1}^{n}\sum_{k=1}^{m} c_{ijk} b^*_{ijk}(H_i(u))\right) \quad (35)$$

with $a_j \in \Re, c_{ijk} \in \Re$, and $b^*_{ijk} \in B^*$. Since g is a continuous real-valued function on a compact metric space and the elements of S are continuous, by the Stone–Weierstrass Theorem [5] we have the following. If S is an algebra, separates the points of $[0,1]^{t_f n}$, and does not vanish on $[0,1]^{t_f n}$, then there exists an $s \in S$ such that $|g(u) - s(u)| < \epsilon$ for all $u \in [0,1]^{t_f n}$. We now show that S has the three required properties. First, clearly S does not vanish because $\exp(w)$ is nonzero for any real value w. Second, it can be readily shown that if a and b are elements of S then the pointwise product $ab \in S, (a+b) \in S$, and $\gamma a \in S$ for any $\gamma \in \Re$. Therefore S is an algebra. All that remains to complete the requirements of the Stone–Weierstrass Theorem is that S separates the points of $[0,1]^{t_f n}$. Let u and v be elements of $[0,1]^{t_f n}$ such that $u \neq v$. S separates the points of $[0,1]^{t_f n}$ if for any such u and v, there exists some $s \in S$ such that $s(u) \neq s(v)$. The fact that u is not equal to v implies that there exists some integers k and i such that $0 \leq k < t_f$ and $K_1(H_i(u))(k) \neq K_1(H_i(v))(k)$. Therefore by the second property of a complete memory, there exists some $b \in B$ such that $(b \circ K_1(H_i(u)))(t_f - 1) \neq (b \circ K_1(H_i(v))(t_f - 1)$. Therefore by the definition of B^* there exists a $b^* \in B^*$ such that $b^*(H_i(u)) \neq b^*(H_i(v))$. Since the exponential function is strictly monotonic, $\exp(b^*(H_i(u))) \neq \exp(b^*(H_i(v)))$. Since the function $u \rightarrow \exp(b^*(H_i(u)))$ belongs to S, S separates the points of $[0,1]^{t_f n}$. By the Stone–Weierstrass Theorem, there exist real numbers a_j and c_{ijk}, natural numbers p and m, and elements b^*_{ijk} of B^* such that

$$\left| g(u) - \sum_{j=1}^{p} a_j \exp\left(\sum_{i=1}^{n} \sum_{k=1}^{m} c_{ijk} b^*_{ijk}(H_i(u)) \right) \right| < \epsilon \quad (36)$$

for all $u \in [0,1]^{t_f n}$. We now make a couple of final observations to complete the proof. First recall that $g(u) = (f \circ K_n(u))(t_f - 1)$ and $b^*_{ijk}(H_i(u)) = (b_{ijk} \circ K_1(H_i(u)))(t_f - 1)$. Finally observe that because of the causality of f and b_{ijk}, for each $x \in X^n$ there is an $u \in [0,1]^{t_f n}$ such that $(f \circ K_n(u))(t_f - 1) = (f \circ x)(t_f - 1)$ and $(b_{ijk} \circ K_1(H_i(u)))(t_f - 1) = (b_{ijk} \circ x_i)(t_f - 1)$. This completes the proof of the lemma.

From Lemma 1 and the fact that $(T_\beta \circ x) \in X^n$ for all positive values of β

$$\left| (f \circ T_\beta \circ x)(t_f - 1) - \sum_{j=1}^{p} a_j \cdot \exp\left(\sum_{i=1}^{n} \sum_{k=1}^{m} c_{ijk} (b_{ijk} \circ T_\beta \circ x_i)(t_f - 1) \right) \right| < \epsilon \quad (37)$$

for all $\beta > 0$ and for all $x \in X^n$. Due to the time invariance of $f, (f \circ T_\beta \circ x)(t_f - 1) = (f \circ x)(t_f - 1 - \beta)$ for $\beta < t_f$. By the third property of a complete memory, $(b_{ijk} \circ T_\beta \circ x_i)(t_f - 1) = (b_{ijk} \circ x_i)(t_f - 1 - \beta)$ for $\beta < t_f$. From these two observations it is apparent that for all β such that $0 \leq \beta < t_f$ and for all $x \in X^n$

$$\left| (f \circ x)(t_f - 1 - \beta) - \sum_{j=1}^{p} a_j \cdot \exp\left(\sum_{i=1}^{n} \sum_{k=1}^{m} c_{ijk} (b_{jik} \circ x_i)(t_f - 1 - \beta) \right) \right| < \epsilon. \quad (38)$$

Now, let $t = t_f - 1 - \beta$, and observe that the proof of the theorem is complete.

Proof of Corollary 3: Let $\epsilon > 0$. By the assumption that f has approximately finite memory on X^n, choose an integer k such that

$$|(f \circ x)(t) - (f \circ W_{t,k} \circ x)(t)| < \epsilon/3 \quad (39)$$

for all $t \in Z_+$ and $x \in X^n$. Let $t_f = 2k$. By Theorem 1, we choose a g of a certain form so that $|(f \circ x)(t) - (g \circ x)(t)| < \epsilon/3$ for all $t < t_f$ and $x \in X^n$. Using Corollary 1 we choose an h of the form

$$h(x) = \sum_{j=1}^{p} a_j \exp\left(\sum_{i=1}^{n} \sum_{d=1}^{m} w_{ijd} (s_{id} \circ x_i) \right) \quad (40)$$

so that $h = g$. By Corollary 2, choose γ, an element of D_{nm}, such that

$$|(h \circ x)(t) - \gamma(V_{nm}(s \circ x)(t))| < \epsilon/3. \quad (41)$$

By the triangle inequality, $|(f \circ x)(t) - \gamma(V_{nm}(s \circ x)(t))| < 2\epsilon/3$ for all $t < t_f$ and $x \in X^n$. Since $(W_{t,k} \circ x) \in X^n$

$$|(f \circ W_{t,k} \circ x)(t) - \gamma(V_{nm}(s \circ W_{t,k} \circ x)(t))| < 2\epsilon/3 \quad (42)$$

for the same values of t and x. Let t_1 be an arbitrary integer greater than $t_f - 1$. Let $\beta = t_1 - (t_f - 1)$. Let P_β be the advance operator defined by

$$(P_\beta \circ x)(t) = x(t + \beta). \quad (43)$$

Observe that by the definitions of $T_\beta, W_{t,k}$, and P_β, and by our choice of t_f

$$(T_\beta \circ W_{t_f - 1,k} \circ P_\beta \circ x_i) = (W_{t_1,k} \circ x_i). \quad (44)$$

By the third property of a complete memory

$$(s_{id} \circ W_{t_1,k} \circ x_i)(t_1) = (s_{id} \circ W_{t_f - 1,k} \circ P_\beta \circ x_i)(t_f - 1). \quad (45)$$

Similarly, by the time-invariance property of f

$$(f \circ W_{t_1,k} \circ x)(t_1) = (f \circ W_{t_f - 1,k} \circ P_\beta \circ x)(t_f - 1). \quad (46)$$

Since $(P_\beta \circ x) \in X^n$, by a special case of (42)

$$|(f \circ W_{t_f - 1,k} \circ P_\beta \circ x)(t_f - 1) - \gamma(V_{nm}(s \circ W_{t_f - 1,k} \circ P_\beta \circ x)(t_f - 1))| < 2\epsilon/3. \quad (47)$$

Substituting (45) and (46) into this inequality yields

$$|(f \circ W_{t_1,k} \circ x)(t_1) - \gamma(V_{nm}(s \circ W_{t_1,k} \circ x)(t_1))| < 2\epsilon/3. \tag{48}$$

Since t_1 is arbitrary and greater than $t_f - 1$, (42) is true not only for $t < t_f$, but for all $t \in Z_+$. By the triangle inequality and (39), we have

$$|(f \circ x)(t) - \gamma(V_{nm}(s \circ W_{t,k} \circ x)(t))| < \epsilon \tag{49}$$

for all $t \in Z_+$ and $x \in X^n$. This completes the proof.

Proof of Corollary 5: As in the proof of Corollary 4, we give a proof by contradiction: Assuming that S does not satisfy the second property of a complete memory, by Corollary 4 we know there must exist a continuous, time-invariant, causal function f from X^n to R, positive ϵ, and $t_f \in Z_+$ such that for each $k \in K$

$$|(f \circ x)(c) - (k \circ x)(c)| > \epsilon \tag{50}$$

for some positive integer $c < t_f$ and some $x \in X^n$.

Let $\tilde{f}$ be an approximately finite memory function defined by

$$(\tilde{f} \circ x)(t) = (f \circ W_{t,t_f} \circ x)(t) + (T_{t_f+1} \circ x_1)(t). \tag{51}$$

Clearly $\tilde{f}$ is also continuous, causal, and time-invariant. For the case in which $p \geq t_f$, for all $\tau \leq t_f (W_{t,p} \circ x)(\tau) = x(\tau)$. Therefore for $t < t_f, (\tilde{f} \circ x)(t) = (f \circ x)(t)$, and $(k \circ x)(t) = (k \circ W_{t,p} \circ x)(t)$. By (50), for each $k \in K$ and all $p \geq t_f$ there exists some positive integer $c < t_f$ and some $x \in X^n$ such that $|(\tilde{f} \circ x)(c) - (k \circ W_{t,p} \circ x)(c)| > \epsilon$. So, for the hypothesis of the corollary to be true, there must be some $p < t_f$ and some $k \in K$ such that

$$|(\tilde{f} \circ x)(t) - (k \circ W_{t,p} \circ x)(t)| < \delta \tag{52}$$

for all $t \in Z_+$ and all $x \in X$ and any $\delta > 0$. Let x be the zero element of X^n. Let $\tilde{x}$ be the element of X^n such that $\tilde{x}_1(1) = 1$ and $\tilde{x}_i(t) = 0$ for $(i,t) \neq (1,1)$. By the definition of $\tilde{f}, |(\tilde{f} \circ x)(t_f + 2) - (\tilde{f} \circ \tilde{x})(t_f + 2)| = 1$. However, since $p < t_f, (W_{t_f+2,p} \circ x)(t) = (W_{t_f+2,p} \circ \tilde{x})(t)$ for all $t \in Z_+$. Because $k \in K$ implies k is a function,

$$(k \circ W_{t_f+2,p} \circ x)(t_f+2) = (k \circ W_{t_f+2,p} \circ \tilde{x})(t_f+2) \tag{53}$$

for all $k \in K$. Therefore application of the triangle inequality leads to a contradiction of (52). So, the assumption that S does not have the second property of a complete memory leads to a contradiction with the hypothesis of the corollary. Therefore, S must have the second property of a complete memory and the proof is complete.

Proof of Theorem 2: In order to show that the set B of all habituation functions is a complete memory, it is necessary to show that it meets the four required properties. (The elements of B are clearly continuous.) First we will establish the first property, that there exists real numbers a and c such that $(b \circ x)(t) \in (a,c)$ for all $b \in B, x \in X$, and $t \in Z_+$. It is sufficient to show that $(b \circ x)(t) \in [0,1]$. This is proven as follows by using mathematical induction and recalling the range of values α and τ can take. Since $b_{-1} = 1, (b \circ x)(0) = 1 - \tau x(0)$. Because τ and $x(0)$ are elements of $[0,1], (b \circ x)(0) \in [0,1]$. All that remains to be shown is that $(b \circ x)(k) \in [0,1]$ implies $(b \circ x)(k+1) \in [0,1]$. Because of the range of τ and α values

$$0 \leq (1 - \tau x(k+1))(b \circ x)(k) \leq (b \circ x)(k) \tag{54}$$

and

$$0 \leq \alpha\tau(1 - (b \circ x)(k)) \leq 1 - (b \circ x)(k). \tag{55}$$

Since $(b \circ x)(k+1) = (1 - \tau x(k+1))(b \circ x)(k) + \alpha\tau(1 - (b \circ x)(k))$

$$0 \leq (b \circ x)(k+1) \leq (b \circ x)(k) + (1 - (b \circ x)(k)) = 1. \tag{56}$$

So B satisfies the first property of a complete memory.

Next we show that B satisfies the second property: for any $t \in Z_+$ and any t_f such that $0 < t_f \leq t$ the following is true. If x and $y \in X$ and $x(t_f) \neq y(t_f)$, then there exists $b \in B$ such that $(b \circ x)(t) \neq (b \circ y)(t)$.

We first prove the following lemma.

Lemma 2: If $b \in B$ is a habituation function with habituation parameters α and τ as defined in Theorem 2, an equivalent definition for b is the following:

$$(b \circ x)(t) = \alpha\tau + \alpha\tau \sum_{j=1}^{t} \prod_{h=j}^{t} (1 - \alpha\tau - \tau x(h)) + b_{-1} \prod_{i=0}^{t} (1 - \alpha\tau - \tau x(i)). \tag{57}$$

This is readily proven using mathematical induction.

Let x and y be elements of X such that $x(t_f) \neq y(t_f)$ for some $t_f \leq t$. This implies that there exists a natural number i with the following three properties. First, $0 \leq i \leq t$. Second, there exists a $\delta > 0$ such that $|x(i) - y(i)| > \delta$. Third $i < j \leq t$ implies $x(j) = y(j)$. The number i represents the latest time prior to t at which x and y differ. Now we use i to define a value β. If $i < t, \beta$ is defined by

$$\beta = \alpha\tau \prod_{h=i+1}^{t} (1 - \alpha\tau - \tau x(h)) = \alpha\tau \prod_{h=i+1}^{t} (1 - \alpha\tau - \tau y(h)). \tag{58}$$

If $i = t$, then $\beta = \alpha\tau$. Observe $\beta > 0$ because $\alpha\tau + \tau < 1$. Using Lemma 2 and some algebraic manipulation we derive the following:

$$\begin{aligned}(b \circ x)(t) - (b \circ y)(t) &= \beta \sum_{j=1}^{i} \left[\prod_{h=j}^{i} (1 - \alpha\tau - \tau x(h)) - \prod_{h=j}^{i} (1 - \alpha\tau - \tau y(h)) \right] \\ &\quad + \frac{\beta}{\alpha\tau} \left[\prod_{j=0}^{i} (1 - \alpha\tau - \tau x(j)) - \prod_{j=0}^{i} (1 - \alpha\tau - \tau y(j)) \right].\end{aligned} \tag{59}$$

Since we have restricted τ and α to have positive values such that $\alpha\tau+\tau<1$ we can make an important observation that

$$0<(1-\alpha\tau-\tau)\le(1-\alpha\tau-\tau x(t))\le(1-\alpha\tau)<1 \quad (60)$$

for all $x\in X$ and $t\in Z_+$.

Consider the special case in which $i=0$. In this case

$$(b\circ x)(t)-(b\circ y)(t)=\frac{\beta}{\alpha}(y(0)-x(0)). \quad (61)$$

Since $i=0, x(0)\neq y(0)$, and therefore $(b\circ x)(t)\neq(b\circ y)(t)$. Now consider the case in which $i\ge 1$. From (59) and (60), we derive a lower bound on $|(b\circ x)(t)-(b\circ y)(t)|$ in terms of i,δ, and β

$$\begin{aligned}&|(b\circ x)(t)-(b\circ y)(t)|\\&\quad\ge\beta\tau\delta-\beta\sum_{j=2}^{i}(1-\alpha\tau)^j-(1-\alpha\tau-\tau)^j\\&\qquad-\frac{\beta}{\alpha\tau}\big((1-\alpha\tau)^{i+1}-(1-\alpha\tau-\tau)^{i+1}\big). \quad (62)\end{aligned}$$

Because $(1-\alpha\tau)>(1-\alpha\tau-\tau)$ and $i>0$ the following set of inequalities hold.

Let $\phi=1/(\alpha\tau)(\alpha\tau+\tau)-1$. Let $\gamma=(1-\alpha\tau)^2/\alpha\tau^2$

$$\begin{aligned}&\sum_{j=2}^{i+1}[(1-\alpha\tau)^j-(1-\alpha\tau-\tau)^j]\\&\qquad\le\sum_{j=2}^{\infty}[(1-\alpha\tau)^j-(1-\alpha\tau-\tau)^j]=\tau\phi \quad (63)\\&(1-\alpha\tau)^{i+1}-(1-\alpha\tau-\tau)^{i+1}\\&\qquad\le(1-\alpha\tau)^{i+1}\le(1-\alpha\tau)^2 \quad (64)\\&\left(\frac{1}{\alpha\tau}\right)((1-\alpha\tau)^{i+1}-(1-\alpha\tau-\tau)^{i+1})\\&\qquad\le\tau\gamma \quad (65)\\&|bx(t)-by(t)|\\&\qquad\ge\beta\tau(\delta-\phi-\gamma). \quad (66)\end{aligned}$$

Since β,τ, and δ are positive values, it is sufficient to show that the quantity $\phi+\gamma$ can be made arbitrarily close to zero by selecting appropriate values for α and τ. The upper bound on the range of α is given by the inequality $\alpha\tau+\tau<1$. For any c such that $(1-\tau)>c>0$ the following is an acceptable value for α:

$$\alpha=\frac{1-(\tau+c)}{\tau}. \quad (67)$$

Since τ can take values arbitrarily close to zero, we can complete the proof of the second property by demonstrating that we can choose appropriate values τ and c so that $\phi+\gamma<\delta$. Let $\epsilon>0$. For any $\epsilon<0.5$ we can choose $\tau=\epsilon$ and $c=\epsilon$

$$\begin{aligned}\phi+\gamma&=\frac{1+(1-\alpha\tau)^2(1+\alpha)}{\alpha\tau(\alpha\tau+\tau)}-1\\&=\frac{1+(\tau+c)^2\left(1+\dfrac{1-(\tau+c)}{\tau}\right)}{(1-(\tau+c))(1-c)}-1. \quad (68)\end{aligned}$$

If we plug in our assigned values for τ and c we get

$$\phi+\gamma=\frac{1}{(1-2\epsilon)(1-\epsilon)}(1+4\epsilon^2+4\epsilon-8\epsilon^2)-1. \quad (69)$$

Taking the limit as ϵ approaches zero we get

$$\phi+\gamma=\lim_{\epsilon\to 0}(1)(1+0+0-0)-1=0. \quad (70)$$

Since choosing any arbitrarily small ϵ yields values of α and τ in the proper range, there must be acceptable values of α and τ for which $\phi+\gamma<\delta$. Thus B satisfies the second property of a complete memory.

Next we show that the third property holds. If $b\in B$ then $(b\circ T_\beta\circ x)(t)=(b\circ x)(t-\beta)$ for all $t\in Z_+$, all $x\in X$ and any β such that $0\le\beta\le t$. By using mathematical induction it can be readily shown that $(b\circ T_\beta\circ x)(t)=1$ for all $t<\beta$. Using this fact, it is then easy to show that by the recursive definition of habituation given in Theorem 2 the third property is satisfied. Once again we use mathematical induction: Because $b_{-1}=(b\circ T_\beta\circ x)(\beta-1)=1$ and $x(0)=(T_\beta\circ x)(\beta), (b\circ T_\beta\circ x)(\beta)=(b\circ x)(0)$; and since $(T_\beta\circ x)(t)=x(t-\beta)$ for all $t>\beta$, it follows directly that the assumption, $(b\circ T_\beta\circ x)(\beta+k)=(b\circ x)(k)$ implies $b\circ T_\beta\circ x(\beta+k+1)=(b\circ x)(k+1)$ for any $k\in Z_+$. Therefore B satisfies the third property of a complete memory. The fourth requirement for B to be a complete memory is that the elements of B are causal. Causality is readily apparent from the definition of B given in Theorem 2. Thus, B is a complete memory and the proof of Theorem 2 is complete.

REFERENCES

[1] G. Cybenko, "Approximations by superpositions of a sigmoidal function," *Math. Contr., Signals, Syst.*, vol. 2, pp. 303–314, 1989.

[2] J. Park and I. W. Sandberg, "Universal approximation using radial basis function networks," *Neural Computa.*, vol. 3, no. 2, pp. 246–257, Summer 1991.

[3] Y. Shin and J. Ghosh, "Ridge polynomial networks," *IEEE Trans. Neural Networks*, vol. 6, pp. 610–622, May 1995.

[4] ——, "Function approximation using higher-order connectionist networks," Computer and Vision Res. Center, Univ. Texas, Austin, Tech. Rep. TR-92-12-87, May 1992.

[5] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York: McGraw-Hill, 1976.

[6] A. N. Kolmogorov, "On the representations of continuous functions of many variables by superpositions of continuous functions of one variable and addition," *Dokl. Akade. Nauk USSR*, vol. 114, no. 5, pp. 953–956, 1957.

[7] D. A. Sprecher, "A numerical implementation of Kolmogorov's theorems," *Neural Networks*, vol. 9, no. 5, pp. 765–771, 1996.

[8] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function theory," *IEEE Trans. Inform. Theory*, vol. 39, pp. 930–945, May 1993.

[9] K. S. Narendra and K. Parthasarathy, "Identification and control of dynamical systems using neural networks," *IEEE Trans. Neural Networks*, vol. 1, pp. 4–27, Mar. 1990.

[10] B. de Vries and J. C. Principe, "The gamma model—A new neural-net model for temporal processing," *Neural Networks*, vol. 5, pp. 565–576, 1992.

[11] A. Waibel, "Modular construction of time-delay neural networks for speech recognition," *Neural Computa.*, vol. 1, no. 1, pp. 39–46, 1989.

[12] A. D. Back and A. C. Tsoi, "A comparison of discrete-time operator models for nonlinear system identification," in *Advances in Neural Information Processing Systems: Proc. 1994 Conf.*, vol. 7, pp. 883–890.

[13] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 9, pp. 115–133, 1943.

[14] H. T. Siegelmann, B. G. Horne, and C. L. Giles, "Computational capabiliites of recurrent NARX neural networks," to be published in *IEEE Trans. Syst., Man, Cybern.* Also, Univ. Maryland, College Park, MD, Tech. Reps. UMIACS-TR-95-78 and CS-TR-3500.

[15] H. T. Siegelmann and E. D. Sontag, "On the computational power of neural networks," *J. Comput. Syst. Sci.*, vol. 50, no. 1, pp. 132–150, 1995.
[16] I. W. Sandberg, "Structure theorems for nonlinear systems," *Multidimensional Syst. Signal Processing*, vol. 2, pp. 267–286, 1991. (See also the errata in vol. 3, p. 101, 1992.)
[17] ——, "Multidimensional nonlinear systems and structure theorems," *J. Circuits, Syst., and Computers*, vol. 2, no. 4, pp. 383–388, 1992.
[18] I. W. Sandberg and L. Xu, "Network approximation of input–output maps and functionals," *J. Circuits, Syst., Signal Processing*, vol. 15, no. 6, pp. 711–725, 1996.
[19] T. Chen and H. Chen, "Approximation of continuous functionals by neural networks with application to dynamical systems," *IEEE Trans. Neural Networks*, vol. 4, pp. 910–918, Nov. 1993.
[20] ——, "Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems," *IEEE Trans. Neural Networks*, vol. 6, pp. 918–928, July 1995.
[21] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Networks*, vol. 5, pp. 157–166, Mar. 1994.
[22] T. Lin, B. G. Horne, P. Tiňo, and C. L. Giles, "Learning long-term dependencies in NARX recurrent neural networks," *IEEE Trans. Neural Networks*, vol. 7, pp. 1329–1338, Nov. 1996.
[23] Y. Bengio and P. Frasconi, "Input–output HMM's for sequence processing," *IEEE Trans. Neural Networks*, vol. 7, pp. 1231–1248, Sept. 1996.
[24] S. Grossberg, *Studies of Mind and Brain*. Dordrecht, The Netherlands: Reidel, 1982.
[25] B. W. Stiles and J. Ghosh, "Habituation based neural networks for spatio-temporal classification," *Neurocomputing*, vol. 15, no. 3/4, ppl 273–307, 1997.
[26] ——, "Some limitations of linear memory architectures for signal processing," in *Proc. 1996 Int. Workshop on Neural Networks for Identification, Contr., Robot., Signal/Image Processing*, Venice, Italy, 1996, pp. 102–110.
[27] ——, "Nonlinear memory functions for modeling discrete-time systems," Center for Vision and Image Sci., Univ. Texas Austin, Tech. Rep. UT-CVIS-TR-96-004, available http://www.lans.ece.utexas.edu under "technical reports."
[28] M. H. Stone, "A generalized Weierstrass approximation theorem," in R. C. Buck, Ed., *Studies in Modern Analysis*. The Math. Assoc. Amer., 1962.
[29] I. W. Sandberg and L. Xu, "Uniform approximation and gamma networks," *Neural Networks*, vol. 10, no. 5, pp. 781–784, 1997.
[30] D. Robin, P. Abbas, and L. Hug, "Neural response to auditory patterns," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1673–1682, 1990.
[31] J. H. Byrne and K. J. Gingrich, "Mathematical model of cellular and molecular processes contributing to associative and nonassociative learning in Aplysia," in *Neural Models of Plasticity*, J. H. Byrne and W. O. Berry, Eds. San Diego, CA: Academic, 1989, pp. 58–70.
[32] J. Ghosh, L. Deuser, and S. Beck, "A neural-network-based hybrid system for detection, characterization, and classification of short-duration oceanic signals," *IEEE J. Oceanic Eng.*, vol. 17, pp. 351–363, Oct. 1992.
[33] B. Stiles and J. Ghosh, "A habituation based neural network for spatio-temporal classification," in *Neural Networks for Signal Processing V, Proc. 1995 IEEE Workshop*, Cambridge, MA, Sept. 1995, pp. 135–144.
[34] F. J. Pineda, "Recurrent backpropagation and the dynamical approach to adaptive neural computation," *Neural Computa.*, vol. 1, no. 2, pp. 161–172, 1989.

Bryan Waitsel Stiles (M'91) was born in Portsmouth, VA, on September 8, 1970. He received the degree of Bachelor of Science in electrical engineering from the University of Tennessee at Knoxville in 1992. He worked as a Research Assistant in the Laboratory for Artificial Neural Systems at the University of Texas at Austin where received the master's degree in May 1997.

He joined the Jet Propulsion Laboratory in Pasadena, CA, in 1997.

Mr. Stiles is a member of Eta Kappa Nu. While an undergraduate, he received a number of awards including the Tennessee scholarship, the Andy Holt Scholarship, and the S. T. Harris scholarship. During graduate school, he received the Du Pont Graduate Fellowship in Electrical Engineering and the Microelectronics and Computer Development Fellowship.

Irwin W. Sandberg (S'54–M'58–SM'73–F'73–LF'97) received the B.E.E., M.E.E., and D.E.E. degrees from the Polytechnic Institute of Brooklyn (now the Polytechnic University) in 1955, 1956, and 1958, respectively.

From 1958 to 1986, he was with Bell Laboratories, Murray Hill, New Jersey, as a Member of Technical Staff in the Communication Sciences Research Division and, from 1967 to 1972, as Head of the Systems Theory Research Department. He is presently a Professor of Electrical and Computer Engineering at the University of Texas at Austin, where he holds the Cockrell Family Regents Chair in Engineering. He holds nine patents. He has been concerned with the analysis of radar systems for military defense, synthesis and analysis of linear networks, several studies of qualitative properties of nonlinear systems (with emphasis on the theory of nonlinear networks as well as on the development of input–output stability theory), and with some problems in communication theory and numerical analysis. His more recent interests include studies of the approximation and signal-processing capabilities of dynamic nonlinear networks.

Dr. Sandberg received the first Technical Achievement Award of the IEEE Circuits and Systems Society. He was a Westinghouse Fellow in 1956 and a Bell Laboratories Fellow from 1957 to 1958. He is a Fellow of the American Association for the Advancement of Science, an IEEE Centennial Medalist, an Outstanding Alumnus of Polytechnic University, a former Vice Chairman of the IEEE Group on Circuit Theory, and a former Guest Editor of the IEEE TRANSACTIONS ON CIRCUIT THEORY Special Issue on Active and Digital Networks. He has published extensively and has been an advisor to *American Men and Women of Science.* He is listed in *Who's Who in America*. He has received outstanding paper awards, an ISI Press Classic Paper Citation, and a Bell Laboratories Distinguished Staff Award. He is a member of SIAM, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and the National Academy of Engineering.

Joydeep Ghosh received the B. Tech. degree from the Indian Institute of Technology, Kanpur, in 1983, and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1988.

He is currently an Associate Professor with the Department of Electrical and Computer Engineering at the University of Texas, Austin, where he holds the Endowed Engineering Foundation Fellowship. He directs the Laboratory for Artificial Neural Systems (LANS), where his research group is studying neural-network models inspired by the cerebellar and visual cortex, and investigating their signal and image processing applications. He has published more than 100 refereed papers and edited six books.

Dr. Ghosh has served as the general chairman for the SPIE/SPSE Conference on Image Processing Architectures, Santa Clara, in 1990, and as Conference Cochair of Artificial Neural Networks in Engineering (ANNIE)'93 through ANNIE'96, and in the program committee of several conferences on neural networks and parallel processing. He received the 1992 Darlington Award for the Best Paper in the areas of CAS/CAD, and also "best conference paper" citations for three neural network papers. He is an Associate Editor of *Pattern Recognition*, IEEE TRANSACTIONS ON NEURAL NETWORKS, and *Neural Computing Surveys.*