Focused Local Learning With Wavelet Neural Networks

Eric A. Rying, *Student Member, IEEE*, Griff L. Bilbro, *Senior Member, IEEE*, and Jye-Chyi Lu

***Abstract*—In this paper, a novel objective function is presented that incorporates both *local* and global error as well as model parsimony in the construction of wavelet neural networks. Two methods are presented to assist in the minimization of this objective function, especially the *local error* term. First, during network initialization, a *locally* adaptive grid is utilized to include candidate wavelet basis functions whose local support addresses the *local* error of the local feature set. This set can be either user-defined or determined using information derived from the wavelet transform modulus maxima (WTMM) representation. Second, during network construction, a new selection procedure based on a subspace projection operator is presented to help *focus* the selection of wavelet basis functions to reduce the *local* error. Simulation results demonstrate the effectiveness of these methodologies in minimizing *local* and global error while maintaining model parsimony and incurring a minimal increase on computational complexity.**

***Index Terms*—Local error, objective function, wavelets.**

I. Introduction

NEURAL networks are used in process modeling, data mining, artificial intelligence, machine learning and many other applications. As noted in [1] and [2], artificial neural networks (ANNs) have limited ability to characterize local features, such as discontinuities in curvature, jumps in value or other edges. These local features, which are located in time and/or frequency, typically embody important process-critical information such as aberrant process modes or faults [1]. Bottou [3] has noted that improved localized modeling can aid both data reduction (or compression) and subsequent classification tasks that rely on an accurate representation of these local features.

Zhang and Benveniste [4] and Bakshi *et al.* [5] improved upon this weakness of ANNs by developing wavelet neural networks (WNNs), which are a type of feedforward neural network. However, because the objective functions used in guiding the network construction in ANNs and WNNs is based on global mean square error (MSE), the modeling quality of such key local features is not emphasized. More importantly, as addressed in Martell [2], existing wavelet-based model selection methods (e.g., Saito [6]; Donoho and Johnstone [7]) focus on data de-noising and use an excessive number of wavelet coefficients/bases in their approximation models. This limits wavelet's applicability to potentially large size data encountered in many recent applications such as intelligent manufacturing, which encounter numerous sources of sensor information and image data. This paper extends the ability of WNNs developed in the literature by improving their ability to model local features, thereby minimizing local error, and by reducing the number of wavelets used. Thus, our new WNN can handle more complicated data patterns from large sample signals. Ultimately, these improvements enable the process engineer in assessing process performance and making process-relevant decisions in a timely and cost-effective manner.

This paper makes significant contributions to the following three problems: 1) identification of specific local features in potentially large nonstationary datasets; 2) compression of entire datasets consisting of smooth and nonsmooth trends; and 3) improvement in the quality of modeling for important local features carrying key information. We present a methodology for solving these problems with our new extended wavelet neural network (EWNN) that combines three approaches.

First, a new objective function is presented that reflects local error as well as standard global error and network size (or model parsimony). By separating the losses from modeling local features and global data patterns in the objective function, our network can focus more on local features.

Second, during network initialization, an adaptive number of wavelet basis functions is presented to *focus* the network modeling effort and improve the quality of fit in local regions of interest by increasing the number of finer-resolution wavelet basis functions within the neighborhood of each element u_j of a local feature set U. This set can be defined prior to network construction by the user, but we will propose an automated edge detection methodology based on the wavelet transform modulus maxima (WTMM) technique, as described in Section IV. This automated technique detects significant edges, segments the signal and prioritizes the set of local features or edges based on their local regularity.

Third, a local projection operator is defined to *focus* the selection of wavelet basis functions within any neighborhood during network construction by including finer resolution wavelet basis

Manuscript received February 7, 2001; revised September 21, 2001. This work was supported by the U.S. Department of Education under Graduate Assistance in Areas of National Need (GAANN) Fellowship P200A50074, as well as the NCSU Engineering Research Center for Advanced Electronic Materials Processing (AEMP).The work of J.-C. Lu was supported in part by NSF DMS—0 072 960.

E. A. Rying was with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695-7911 USA. He is now with PDF Solutions, Inc., San Jose, CA 95110 USA (e-mail: earying@alumni.carnegiemellon.edu).

G. L. Bilbro is with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695-7911 USA (e-mail: glb@eos.ncsu.edu).

J.-C. Lu is with the School of Industrial and Systems Engineering (ISyE), Georgia Institute of Technology, Atlanta, GA 30332–0205 USA (e-mail: JCLU@isye.gatech.edu).

Publisher Item Identifier S 1045-9227(02)01799-X.

functions. By splitting the large size data into less-smooth and more-smooth components, parallel computing techniques can be utilized to process these data segments simultaneously to shorten the network construction duration. However, the topic of network construction involving parallel computing techniques is left for future work. For an account of parallel computing techniques for implementing neural networks, see Sundararajan [8].

The rest of this paper is organized as follows. Section II provides a background to wavelet theory as it pertains to wavelet networks. Section III formulates the problem and discusses the new objective function, the adaptive initialization and the subspace projection method. Section IV discusses the use of the WTMM representation for prioritizing the local features to be minimized. Section V provides simulation results demonstrating the effectiveness of the adaptive initialization and the focused selection schemes. These schemes can be used either individually or jointly to minimize the local error, and therefore the new objective function, in local regions of interest in the signal. Section V-D reports on results demonstrating the automatic determination of the model order using the minimum of the new objective function. Alternatively, a new error space analysis (ESA) technique is proposed in Section V-E to help visualize how the new initialization, selection and automatic model order determination procedures tradeoff local and global error against model parsimony. Next, Section V-F reports on a procedure that uses the WTMM representation to automate construction of the local feature set, U. Finally, Section VI finishes with some concluding remarks.

II. Background

WNNs have recently emerged as a powerful new type of ANN [4], [5]. They resemble radial basis function (RBF) networks because of the localized support of their wavelet basis functions [9]. In contrast to classical sigmoidal-based ANNs, wavelet networks provide efficient network construction techniques, faster training times, and multiresolution analysis capabilities.

Wavelet neural networks (WNNs) were first proposed by Zhang and Benveniste [10] as an alternative to classical feedforward ANNs for approximating nonlinear functions. WNNs are feedforward neural networks with one hidden layer, comprised normally of radial (e.g., Mexican hat) wavelets as activation functions, and a linear output layer [11]. The output layer of the WNN represents the weighted sum of the hidden layer units, i.e., wavelet basis functions. Moreover, similar to other neural networks, Zhang and Benveniste [10] have utilized gradient-based techniques for updating the weights in the WNN to further minimize the standard MSE of the network's approximation after network construction. For an account of neural networks, see Haykin [12]. In addition, Bakshi *et al.* [5] introduced an orthogonal wavelet network, *wave-net*, for approximation and classification based on multiresolution analysis. Bakshi's wave-net learns locally in a hierarchical manner, i.e., it can model a hierarchy of resolutions that range from coarse or general data trends, to highly time-localized events such as discontinuities of curvature or edges. Bakshi's wave-net accomplishes this by using orthonormal wavelets and their associated scaling functions as the network activation functions. However, the localized learning in Bakshi's wave-net is achieved *implicitly*, by utilizing the localized wavelet basis functions to minimize a standard *global* error measure. In contrast, this paper will demonstrate *explicit* localized learning using a new *local* error measure, to be discussed further in Section III. Bakshi used wave-nets to predict chaotic time series and to classify experimental data for process fault diagnosis. He found that the training and adaptation efficiency of wave-nets is at least an order of magnitude better than classical ANNs. The reason for this is that wavelet networks are a linear combination of localized basis functions that offer several advantages including orthogonality, efficient numerical procedures and multiresolution analysis capabilities over an RBF network. Wavelet networks have been applied to a wide variety of applications including: nonlinear functional approximation and nonparametric estimation [4], system identification and control tasks [13], and modeling and classification [9].

Wavelet networks can be viewed as an adaptive discretization of the inverse wavelet transform. Following Mallat [14], the inverse wavelet transform is discretized into the following:

$$\hat{f}^M(w,t) = \sum_{i=1}^{M} w_i \frac{1}{\sqrt{s_i}} \psi\left(\frac{t-u_i}{s_i}\right) \tag{1}$$

where the discrete version shown in (1) must be constructed from a family $\mathbf{W}$ of dilated and translated wavelets

$$\mathbf{W} = \left\{ \frac{1}{\sqrt{s_i}} \psi\left(\frac{t-u_i}{s_i}\right) : i \in \mathbf{Z} \right\} \tag{2}$$

so that $\mathbf{W}$ represents an orthonormal basis of $L_2(\mathbf{R})$. The family consists of scaling and translating a mother wavelet, $\psi_{u,s}(t)$

$$\psi_{u,s} = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right)$$

where u and s are the translation and scale parameters, respectively. The $1/\sqrt{s}$ term normalizes $\|\psi_{u,s}(t)\| = 1$. The family in (2) is taken from a double indexed regular lattice

$$\{(s_m, u_n) = (\alpha^m, n\beta\alpha^m) : m, n \in \mathbf{Z}\}$$

where the parameters α and β denote the step sizes of the dilation and translation parameters, e.g., $\alpha = 2$ and $\beta = 1$ form the standard *dyadic* lattice [14]. Using this lattice or grid, the discretized version of the wavelet, $\psi_{m,n}(t)$, becomes

$$\psi_{m,n}(t) = 2^{-m/2} \psi(2^{-m}t - n)$$

Fig. 1(top) shows an example of such a lattice, where the translation parameter has been normalized to the interval $t \in [-1, 1]$.

III. Problem Formulation

Given: Assume two random variables t and y satisfy the following standard regression model

$$y = f(t) + e \tag{3}$$

where $f(t) \in \mathbf{R}$ is an unknown nonlinear function, e is the standard zero mean independently identically distributed

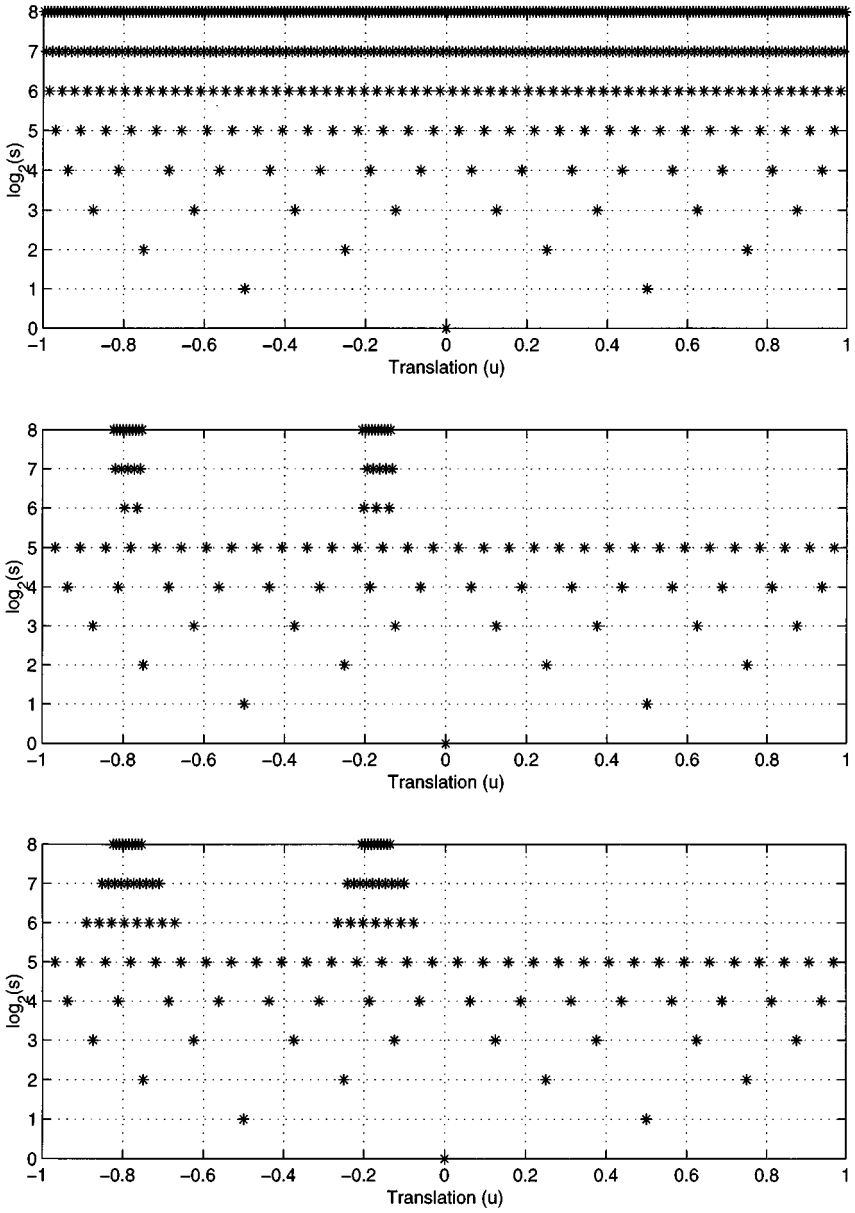

Fig. 1. Examples of the three types of initialization grids. (Top) Standard grid. (Middle) Adaptive grid, with box-like distribution of wavelet centers comprising the local initialization set, $\mathbf{W}_L$. $(G = 5, \rho = 3)$. (Bottom) Adaptive grid, with trapezoidal distribution for $\mathbf{W}_L$. $(G = 5, \rho = 3)$.

(i.i.d.) Gaussian noise error term, i.e., $e \sim N(0, {\sigma_e}^2)$ and $T = \{(t_i, y_i)\}_{i=1}^N$ denotes the set of training data.

Problem Statement: Determine a parsimonious wavelet network approximation, denoted by

$$\hat{f}^M(w,t) = \sum_{j=1}^{M} w_j \psi_j(t)$$

using an M-size subset $(M < L)$ of the wavelet basis functions $\psi_j(t) \in \mathbf{W} = \{\psi_1(t), \ldots, \psi_L(t)\}$ such that $\hat{f}^M(w,t)$ minimizes a cost function denoted by

$$F(M, \lambda, \delta, C) = \sum_{j=1}^{P} \int_{I_j(\delta,C)} \frac{[f(t) - \hat{f}_{jL}(t)]^2}{||f_L||^2}\, dt + \int_{U^c(\delta,C)} \frac{[f(t) - \hat{f}_G(t)]^2}{||f_G||^2}\, dt + \lambda \frac{M}{N} \quad (4)$$

where

- $I_j(\delta, C) = [u_{j,\delta} - C/2, u_{j,\delta} + C/2]$ is the C-size interval for the jth local feature $(u_j \in U = \{u_1, \ldots, u_P\})$ centered at $u_{j,\delta}$ with a priority level defined by the threshold δ;
- $\hat{f}_{jL}(t)$ is the local fitting of the function $f(t)$ for the local interval $I_j(\delta, C)$ and $||f_L||^2$ is the sum of squares of the local window lengths (i.e., $||f_L||^2 = PC^2$);
- $\hat{f}_G(t)$ is the global fitting of the function $f(t)$ on the set $U^c(\delta, C) = (0, \infty) - \cup_{j=1}^{P} I_j(\delta, C)$ and $||f_G||^2$ is the sum of squares of the supports of the global window lengths, or

$$||f_G||^2 = \left(u_1 - \frac{C}{2}\right)^2 + \sum_{i=2}^{P} (u_i - u_{i-1} - C)^2 + \left(N - u_P - \frac{C}{2}\right)^2.$$

Model Parsimony: The third term of (4), $\lambda M/N$, is a monotonically increasing data reduction measure defined as the ratio of the number of model terms, M, to the overall data size, N. The penalty parameter, λ, allows the user to tradeoff network size with local and global error modeling.

Selection of Window Size C: The parameter C_j is the interval width or window size surrounding each local feature, $u_j \in U$. If the local model is a "point-match" function, the window size is one data point. If the local model is larger than point-match, the window size should be larger than the support of the finest resolution wavelet but should not exceed the minimum of all the half-distances between the locations of two adjacent local features, u_j and u_{j+1}. The smaller the window size C is, the larger the region of data that is modeled by the global model, f_G. In this case, the model will be simpler and use fewer coefficients to approximate a given function, $f(t)$. In general, the window size C_j can be an arbitrary function of each local feature u_j of the P critical local features. In this paper, we assume constant window size: $C_j = C$.

Selection of Priority Threshold δ: The parameter δ determines the priority-classes of the local features and determines which local features are included in the wavelet network model. In general, this parameter is determined by the WTMM representation, as described in Section IV.

Explicit Localized Learning: It is important to note that traditional network methodologies achieve localized learning *implicity*, not explicitly as presented in this paper. The reason is that during the construction phase of traditional wavelet and RBF networks, the localized basis functions are chosen for their ability to reduce a *global* error measure, such as the mean-square error. Such a strategy achieves "localized learning" of signal trends only if the local features or regions dominate the global error. In contrast, this paper proposes *explicit* "localized" learning by allocating network resources (i.e., basis functions) in those local regions that are included in the new explicit *local error term* of the objective function [see (4)].

The following sections will propose two novel approaches to minimize $F(M, \lambda, \delta, C)$ [see (4)]. Section III-A will introduce a novel network initialization scheme. Section III-B will discuss the standard selection methods. Next, Section III-C will present a novel approach to selecting basis functions for the wavelet network. Section III-E will demonstrate how to automatically determine the model order, or size, of the wavelet network using the minimum of the objective function. Finally, Section IV will discuss automatic generation of the local feature set, U, using the WTMM representation.

A. Network Initialization: Locally Adaptive Discretization Grid

To help minimize the *local* error term of this new objective function $F(M, \lambda, \delta, C)$ [see (4)], the initialization set of candidate wavelet basis functions, $\mathbf{W}$, is constructed to include those finer-resolution wavelets whose support influences the *local* error within the interval width C of each local feature, $u_j \in U = \{u_1, \ldots, u_P\}$. Denote the initialization set of basis functions as

$$\mathbf{W}^* = \mathbf{W}_G \cup \mathbf{W}_L$$

where $\mathbf{W}_G$ and $\mathbf{W}_L$ denote the global and local sets of wavelet basis functions, respectively. $\mathbf{W}_G$ is constructed using the standard *dyadic* [14] lattice or pyramid of wavelet basis functions up to a given scale, $m = G$.

$$\mathbf{W}_G = \{\psi_{m,n}(t) \colon m \le G\}.$$

To construct the local set $\mathbf{W}_L$, wavelet basis functions at ρ finer resolutions than G are selected. In general, G is chosen based on the degree to which the user wants to model global trends. Similarly, ρ defines the accuracy for modeling local features, but is also limited by the sampling rate of the data. The following two schemes can be used to build the local set, $\mathbf{W}_L$:

Box Scheme: This scheme distributes the centers (located at $u_j = n2^m$) of the basis functions in the time-scale plane according to an abrupt *box-like* distribution, as defined in (5) and illustrated in Fig. 1(middle)

$$\mathbf{W}_L = \{\psi_{m,n}(t) \colon G < m \le (G+\rho) \quad \text{and} \quad |2^{-m}u_j - n| \le C \; \forall u_j \in U\} \quad (5)$$

where $j = 1, \ldots, P$. It should be noted that this scheme can cause Gibbs-like oscillations in the network's approximation outside of a given local feature interval, I_j. To help compensate for these effects, the following alternative is proposed.

Trapezoidal Scheme: This scheme distributes the centers of the basis functions using a more gradual *trapezoidal* profile, as defined by (6) and illustrated in Fig. 1(bottom)

$$\mathbf{W}_L = \{\psi_{m,n}(t) \colon G < m \le (G+\rho) \quad \text{and} \quad |2^{-m}u_j - n| \le C(m) \; \forall u_j \in U\} \quad (6)$$

where

$$C(m) = C(1 - (m - (G+\rho)))$$

so that $C(G+1) = C\rho$ and $C(G+\rho) = C$.

Now that the novel schemes for initializing the set of candidate wavelet basis functions have been discussed, it is important to review network construction, or the selection of basis functions from the initialization set, $\mathbf{W}^*$.

B. Network Construction: Motivation and Background

Once the initialization set $\mathbf{W}^*$ of wavelet basis functions has been constructed, the next step is to select the "best" M-size subset $(M < L)$ of wavelet basis functions in $\mathbf{W}^*$ for estimating $f(t)$. However, in general, a search through all the M-size subsets is a computationally expensive combinatorial optimization problem, i.e., it is NP complete [9]. One nonlinear optimization technique would be to use *genetic algorithms* (GAs), which have been utilized successfully by Echauz [9] for radial wavelet networks and Chen [15], [16] for RBF networks. While GAs can provide optimal or near-optimal network topologies, they do so at the cost of extensive computational requirements [16]. Consequently, this paper will consider the heuristic algorithms proposed by Zhang [4], who viewed this subset selection task within the framework of statistical *regression analysis*.

To select the most significant wavelet *regressors* from the library or set $\mathbf{W}$ of candidate basis functions, Zhang [4] utilized three heuristic algorithms:

1) residual-based selection, which is similar to Mallat's matching pursuit (MP) algorithm [14],
2) stepwise selection by orthogonalization, which is similar to the orthogonal matching pursuit (OMP) of Pati [17], and
3) backward elimination.

The stepwise procedure was also used by Chen [18] for nonlinear modeling with RBF networks, and is summarized by Algorithm 1 in Appendix I.

The basic idea of the first two algorithms is to select, at each stage i, the wavelet $\psi_i(t) \in \mathbf{W}$ that spans the space "closest" to the output function vector, $f(t)$. Additionally, stepwise selection works together with the wavelets selected from the previous stages to maintain orthogonality, thereby gaining some computational efficiency [4]. The algorithm closely resembles the classical Gram–Schmidt orthogonalization procedure. The following is a brief review of the simpler residual-based selection algorithm.

Residual-Based Selection (Zhang [4]): Define the initial residual vector as $\gamma_0(k) = f(t_k), k = 1, \ldots, N$ and the initial wavelet network approximation as $f_0(t) = 0$. Also, let v_j be the orthogonalized version of the wavelet, $\psi_j(t)$. Then, Zhang's approach is to find the orthogonalized wavelet $v_i \in \mathbf{W}$ at stage $i, i = 1, \ldots, M$, that minimizes the following standard global error criterion:

$$J(v_j) = \frac{1}{N}\sum_{k=1}^{N}(\gamma_{i-1}(t_k) - u_j v_j(t_k))^2 = (\gamma_{i-1} - u_j v_j)^T(\gamma_{i-1} - u_j v_j)$$

with

$$u_j = \left(v_j{}^T v_j\right)^{-1} v_j{}^T \gamma_{i-1} = v_j{}^T \gamma_{i-1}$$

with the last equality a result of orthonormality, i.e., $v_j^T v_j = 1$. Next, by substituting u_j into $J(v_j)$ we obtain

$$J(v_j) = \gamma_{i-1}{}^T \gamma_{i-1} - \left(v_j{}^T \gamma_{i-1}\right)^2$$

where we see that minimizing $J(v_j)$ at stage i is equivalent to maximizing $(v_j{}^T \gamma_{i-1})^2$.

Using this result, one can define a linear subspace projection operator that will help select the "best" orthogonalized wavelet, v_j, that spans the local error subspace and defines the local error term, $e_L = f - \hat{f}_L(t)$ [see (4)]. Subspace projection operators for local error minimization are addressed in the next section.

C. Network Construction: A Novel Technique Using Subspace Projectors

Let us define $\mathbf{\Gamma}^L$ as the *local* linear projection operator that serves to project a given vector $f \in \mathbf{R}^{Nx1}$ onto the subspace $\mathbf{\Gamma}^L f$ spanned by the union of the P local features, $u_j \in U = \{u_1, \ldots, u_P\}$, i.e.,

$$\mathbf{I} = \bigcup_{j=1}^{P} I_j, \quad I_j = \left[u_j - \frac{C}{2}, u_j + \frac{C}{2}\right] \tag{7}$$

so that $\mathbf{\Gamma}^L$ can be written in discrete form as a diagonal matrix

$$\mathbf{\Gamma}^L_{ii} = \begin{cases} 1 & |u_j - i| \le \frac{C}{2} \quad \forall u_j \in U \\ 0 & \text{otherwise} \end{cases}$$

and its corresponding orthogonal complement, $\mathbf{\Gamma}^G = (\mathbf{\Gamma}^L)^{\perp}$, can be defined as

$$\mathbf{\Gamma}^G = I - \mathbf{\Gamma}^L$$

where $I \in \mathbf{R}^{NxN}$ is the identity matrix. Fig. 3 shows an example of these two projection operators for $U = \{28\,107\}$ and $C = 10$. Alternatively, a more gradual or tapered form for the local projection operator $\mathbf{\Gamma}^L$, which is trapezoidal in shape is given by the following:

$$\mathbf{\Gamma}^L_{ii} = \begin{cases} \frac{3}{2} - \frac{1}{C}(u_j - i) & u_j - \frac{3C}{2} \le i < u_j - \frac{C}{2} \quad \forall u_j \in U \\ 1 & |u_j - i| \le \frac{C}{2} \quad \forall u_j \in U \\ \frac{3}{2} + \frac{1}{C}(u_j - i) & u_j + \frac{C}{2} \le i < u_j + \frac{3C}{2} \quad \forall u_j \in U \\ 0 & \text{otherwise} \end{cases}$$

Excluding the trapezoidal projector above, one should note here that $(\mathbf{\Gamma}^L)^T \mathbf{\Gamma}^G = 0$, and Γ^G projects $\mathbf{f}$ onto the subspace spanned by $U^c(\delta, C) = (0, \infty) - \cup_{j=1}^{P} I_j(\delta, C)$ [see (4)].

These linear operators act to project the vector f onto its corresponding local and global subspaces

$$\mathbf{\Gamma}^L f = f_L \quad \mathbf{\Gamma}^G f = f_G$$

so that $f_L + f_G = f$, as shown in Fig. 3. Using these operators, one can rewrite the standard global error minimization criteria as

$$E = \min_{v_i \in W} \left\| \left(f_G - \sum_{i=1}^{M_G} w_i \psi_i \right) + \left(f_L - \sum_{i=1}^{M_L} w_i \psi_i \right) \right\|$$

where $M_L + M_G = M$. One observes that at each stage, the chosen wavelet v_i serves to minimize either the local or global error, or possibly both, depending on the support of the wavelet, v_i. The degree to which the wavelet minimizes the error is dependent on its scale level, and therefore its local support. This is discussed further in the following theorem.

Theorem 1 (Local Error Subspace Projection Theorem): Given a set of L normalized wavelet basis functions $\mathbf{W} \in \mathbf{R}^{NxL}$ and a function $f \in \mathbf{R}^{Nx1}$, then for a given disjoint *local feature set* $U = \{u_1, \ldots, u_P\}$ defined over an interval width $C, \exists$ a projection operator, $\mathbf{\Gamma}^L \subset \mathbf{R}^{(N-CP)xN}$ for a finite interval width $C < N$ such that the wavelet $v_{l_i} \in \mathbf{W}$ chosen at stage $i \in I^* = \{1, \ldots, L\}$ will help minimize the local error defined as

$$e_L = \|f - \hat{f}_L\| = \|f - \mathbf{\Gamma}^L f\|$$

TABLE I
WAVELET BASIS FUNCTION SELECTION METHODS FOR MODIFIED STEPWISE SELECTION BY ORTHOGONALIZATION

Method	Error Target	Error Target Domain	Function Argument (ζ)
I	$\text{Local}_{\text{MLSE}}$	$\cup_{j=1}^{P} \text{I}_j(\delta,\text{C})$	$\mathbf{\Gamma}^L f$
II	$\text{Global}_{\text{MSE}}$	$(0,\infty)$	$\mathbf{I}f$
III	$\text{Global}_{\text{MSE}}$ + $\text{Local}_{\text{MLSE}}$	$(0,\infty)$	$(\mathbf{I}+\mathbf{\Gamma}^L)f$
IV	$\text{Local}_{\text{MLSE}}$ for odd stages i	$\cup_{j=1}^{P} \text{I}_j(\delta,\text{C})$	$\mathbf{\Gamma}^L f$ for odd stages i
	$\text{Global}_{\text{MGSE}}$ for even stages i	$\text{U}^c(\delta,\text{C})$	$\mathbf{\Gamma}^G f$ for even stages i

TABLE II
COMPARISON OF STANDARD AND ADAPTIVE INITIALIZATION GRIDS ($M = 60$)

Grid	# Wavelons		MSE	MGSE	MLSE	# FLOPS	Start
Type	Global	Local	(x 10^{-3})	(x 10^{-3})	(x 10^{-2})	(x 10^{7})	#
Standard	48	12	2.95	2.65	6.17	8	511
Adaptive (Box)	44	16	1.66	1.80	0.10	2.1	98
Adaptive (Trap.)	37	23	2.69	2.92	0.26	2.4	118

for each of the local features $u_j \in U$ if the wavelet chosen at stage i is v_{l_i}, where

$$l_i = \arg\max_{j \in I^*} \left(v_j{}^T \zeta\right)^2$$

and $\zeta = \mathbf{\Gamma}^L f$. The resulting wavelet network can be expressed as

$$\hat{f}^M(w,t) = \sum_{i=1}^{M} w_i v_{l_i}(t) = Aw$$

where $A = [v_{l_1}, \ldots, v_{l_M}]$.

Proof (Sketch): The main idea of the projection operator, $\mathbf{\Gamma}^L$, is to help find and select the orthogonalized wavelet, $v_i \in \mathbf{W}$ that will help minimize the local error or residual, $e_L = \gamma^L$. (Also, note that $e_G = \gamma^G$.) This is equivalent to finding the wavelet v_i which spans the space "closest" to f_L, or whose support width influences the local error. The support of a wavelet ψ_i varies inversely with scale $s = 2^j$, i.e.,

$$\text{support}\,(\psi_{m,n}) = [2^{-m}n, 2^{-m}(n+1)]$$

The error, e_L, will be maximized the greatest for wavelets v_i whose support is less than the interval width by some finite precision error $\epsilon > 0 \in \mathbf{R}$, i.e.,

$$\text{support}\,(v_i) - C \leq \epsilon$$

Define the *local* and *global* residual vectors as $\gamma^L = \mathbf{\Gamma}^L f$ and $\gamma^G = \mathbf{\Gamma}^G f$, respectively. Now, using this result and the fact that the residual can be split up into its global and local error components, i.e., $\gamma_{i-1} = \gamma_{i-1}^G + \gamma_{i-1}^L$ (following the residual-based selection algorithm in Section III-B)

$$J(v_j) = \left(\gamma_{i-1}^G + \gamma_{i-1}^L\right)^T \left(\gamma_{i-1}^G + \gamma_{i-1}^L\right) - v_j{}^T \left(\gamma_{i-1}^G + \gamma_{i-1}^L\right)^2$$

where the first term can be expressed as

$$J(v_j) = \left(\gamma_{i-1}^G\right)^T \gamma_{i-1}^G + \left(\gamma_{i-1}^L\right)^T \gamma_{i-1}^L - v_j{}^T \left(\gamma_{i-1}^G + \gamma_{i-1}^L\right)^2$$

due to orthogonality, i.e., since $(\gamma_{i-1}^G)^T \gamma_{i-1}^L = 0$. Next, we consider the two cases based on the local support of the wavelet v_j at stage i.

Case I: $v_j \in \mathbf{W}^L$

For this case, the local support of the wavelet v_j will contribute maximally to minimizing the local error, $e_L = \gamma^L$, so that $v_j{}^T(\gamma_{i-1}^G) < v_j{}^T(\gamma_{i-1}^L)$. Then, we have that

$$J(v_j) = \left(\gamma_{i-1}^L\right)^T \gamma_{i-1}^L + \left(\gamma_{i-1}^G\right)^T \gamma_{i-1}^G - \left(v_j{}^T \gamma_{i-1}^L\right)^2$$

and we see that minimizing $J(v_j)$ is equivalent to maximizing $(v_j{}^T \gamma_{i-1}^L)^2$. However, this is equivalent to finding v_{l_i} such that

$$l_i = \arg\max_{j \in I^*} \left(v_j{}^T \zeta\right)^2$$

where $\zeta = \gamma_{i-1}^L = \mathbf{\Gamma}^L f$.

Case II: $v_j \in \mathbf{W}^G$

For this case, the local support of the wavelet v_j will contribute maximally to minimizing the global error, $e_G = \gamma^G$, so that $v_j{}^T(\gamma_{i-1}^L) < v_j{}^T(\gamma_{i-1}^G)$. Then, we have that

$$J(v_j) = \left(\gamma_{i-1}^L\right)^T \gamma_{i-1}^L + \left(\gamma_{i-1}^G\right)^T \gamma_{i-1}^G - \left(v_j{}^T \gamma_{i-1}^G\right)^2$$

and we see that minimizing $J(v_j)$ is equivalent to maximizing $(v_j{}^T \gamma_{i-1}^G)^2$. However, this is equivalent to finding v_{l_i} such that

$$l_i = \arg\max_{j \in I^*} \left(v_j{}^T \zeta\right)^2$$

where $\zeta = \gamma_{i-1}^G = \mathbf{\Gamma}^G f$. ■

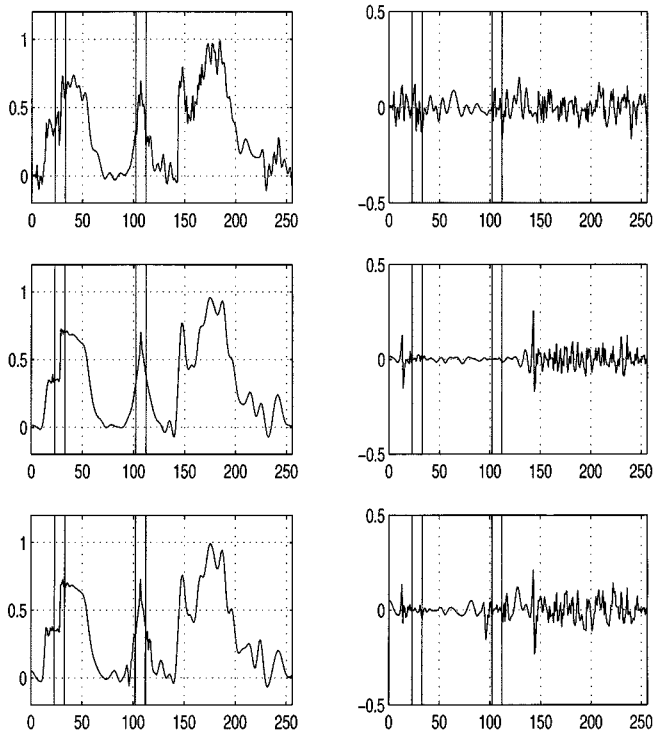

Fig. 2. Comparison of the wavelet network approximations using (top left) standard, (middle left) adaptive (box-profile) and (bottom left) adaptive (trapezoidal profile) initialization grid. Plots (top right), (middle right), and (bottom right) show the corresponding pointwise error for the approximations in (top left), (middle left), and (bottom left), respectively. (Network size: $M = 60$, Interval width: $C = 10$ and local feature set $U = \{28, 107\}$.)

D. Network Construction: Projection Schemes for Basis Selection

Using the proof sketch above, we propose the four basis function selection methods shown in Table I, where ζ is the argument of (8)

$$l_i = \arg\max_{j \in I^*} \left({v_j}^T \zeta \right)^2. \tag{8}$$

Each selection method selects at each stage i of the selection procedure (see Algorithm 1), the wavelet basis function $v_{l_i} \in \mathbf{W}^*$ that is closest in the inner-product sense to the subspace spanned by the error target, ζ, used in (8).

However, it is important to note that selection methods II-IV target different "global" errors. In terms of global error, Methods II and III target the minimization of the standard MSE defined over the whole domain of the function f, i.e., $(0, \infty)$. In contrast, for even selection stages i, method IV targets the minimization of the mean-of-global square error (MGSE), *not* the standard MSE, defined over the domain $U^c(\delta, C) = (0, \infty) - \cup_{j=1}^{P} I_j(\delta, C)$, and given by (9)

$$\text{MGSE} = \frac{1}{N - CP} \sum_{i=1}^{N-CP} (f_{i,G} - \hat{f}_{i,G})^2, \quad \forall i \notin \mathbf{I} \tag{9}$$

TABLE III
COMPARISON OF SELECTION METHODS I–IV USING THE STANDARD INITIALIZATION GRID ($M = 60, C = 10, U = \{28, 107\}$)

Selection Method	MSE (x 10^{-3})	MGSE (x 10^{-3})	MLSE (x 10^{-3})
(I) Local	26.8	29.0	4.09
(II) Global	2.95	2.65	6.17
(III) Local + Global	3.17	3.21	2.68
(IV) Alt. Local + Global	2.81	3.01	0.63

as well as the mean-of-local square error (MLSE) for odd stages i, given by (10)

$$\text{MLSE} = \frac{1}{CP} \sum_{i=1}^{CP} (f_{i,L} - \hat{f}_{i,L})^2, \quad \forall i \in \mathbf{I} \tag{10}$$

where $\mathbf{I}$ is the union of the supports of the local feature set, U [see (7)]. These error targets (i.e., MSE, MLSE, and MGSE) and their respective domains are indicated as subscripts in Table I. The following discusses the selection methods in more detail.

Method I: This method determines the wavelet, at selection stage i, that is closest to the local subspace spanned by the projected version of the signal, $\mathbf{\Gamma}^L f$ [see Fig. 3(middle)]. The purpose is to focus the selection of the wavelet basis functions, for all selection stages $i \in \{1, \ldots, M\}$, solely on the local feature set (thereby minimizing the MLSE) and ignore the global part of the signal, $\mathbf{\Gamma}^G f$ [Fig. 3(bottom)].

Method II: This method is the standard selection algorithm, and therefore chooses the wavelet at stage i which minimizes the standard global error, i.e., MSE. This method does not, in general, place any emphasis during the selection process on the minimization of the error (i.e., the MLSE) for the set of local features, U. Consequently, the local error (i.e., the MLSE) is minimized *implicity*, i.e., if the local features tend to dominate the global error (i.e., the MSE).

Method III: This method compromises between local and global fitting by selecting the wavelet which has the largest inner product with the superposition of the two projected signals, $(\mathbf{I} + \mathbf{\Gamma}^L) f$. In other words, Method III *weights* the signal more in the local regions of interest [see Fig. 3(middle)]. The aim is to choose wavelets which will impact the error in these local regions (i.e., the MLSE) at a higher rate than the error in the *global* parts (i.e., the MGSE) of the signal, $\mathbf{\Gamma}^G f$ [Fig. 3(bottom)].

Method IV: This method attempts to trade-off the minimization of the local (i.e., the MLSE) and global error (i.e., MGSE), *not* MSE) by alternating between the local projected signal, $\mathbf{\Gamma}^L f$ [Fig. 3(middle)], and its complement, $\mathbf{\Gamma}^G f$ [see Fig. 3(bottom)], according to the following scheme:

- at *odd* selection stages i, select the wavelet $v_{l_i} \in \mathbf{W}^*$ which is closest to the *local* subspace, i.e., $\mathbf{\Gamma}^L f$ [see Fig. 3(middle)], and
- at *even* selection stages i, select the wavelet $v_{l_i} \in \mathbf{W}^*$ which is closest to the *global* subspace, i.e., $\mathbf{\Gamma}^G f$ [see Fig. 3(bottom).

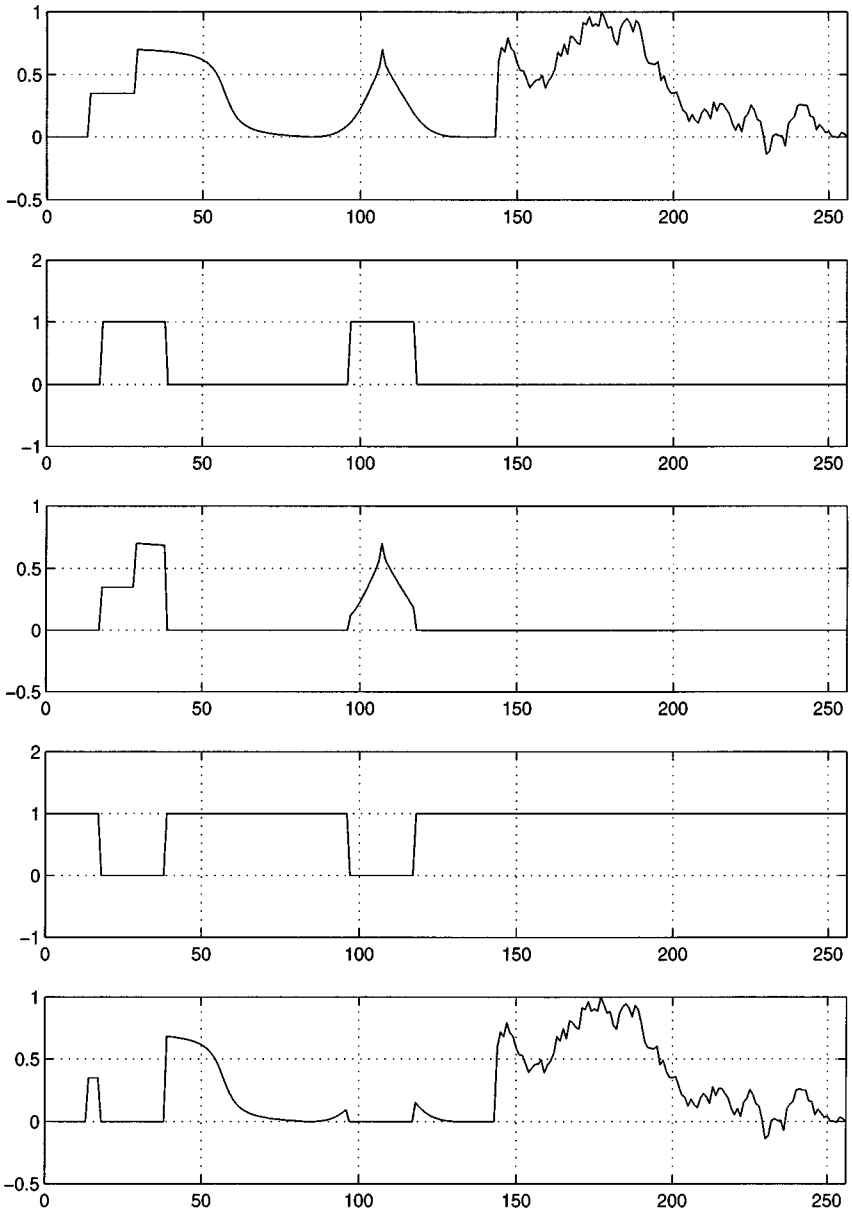

Fig. 3. Plots illustrating the subspace projection operators. Given a signal (top) $\mathbf{f}$, the (second from top) local projection operator $\mathbf{\Gamma}^L$ projects $\mathbf{f}$ onto its local subspace (middle) $\mathbf{\Gamma}^L\mathbf{f}$. The (second from bottom) global projection operator $\mathbf{\Gamma}^G$ projects $\mathbf{f}$ onto (bottom) its global subspace $\mathbf{\Gamma}^G\mathbf{f}$. Note that $\mathbf{\Gamma}^L\mathbf{f} + \mathbf{\Gamma}^G\mathbf{f} = \mathbf{f}$.

As a result, Method IV divides the allocation of network resources, i.e., wavelet basis functions, equally between local [Fig. 3(middle)] and "global" portions [Fig. 3(bottom)] of the signal. Here, "global" refers to those portions of the signal not included in the local feature set, U. As a result, during even selection stages, this method targets the MGSE, *not* the standard MSE.

Thus, these methods *find* and *select* a wavelet basis function at stage i from the set $\mathbf{W}^*$, but in no way help to provide a local orthonormal basis for each of the local feature intervals, $[u_j - C/2, u_j + C/2]$, $u_j \in U$, such as the local cosine transform (LCT) or similar lapped orthogonal transforms (LOT) [14].

Now that a novel methodology has been proposed for selecting the wavelet basis functions from the initialization set, $\mathbf{W}^*$, it is important to discuss how to determine the size or model order of the wavelet network.

E. Network Size: Model Order Determination

The problem of determining the size M of the wavelet network, i.e., the number of wavelet basis functions in the model, can be viewed as the standard model order determination problem [4]. Some of the standard approaches include Akaike's information-theoretic criteria (AIC), Akaike's final prediction error criterion (FPE), generalized cross-validation (GCV), statistical hypothesis test, and Schwartz and Rissanen's minimum description length (MDL) criterion [4]. Under certain assumptions [4], the GCV method gives an approximate estimate of the MSEs [10], given by (11)

$$\mathrm{GCV}(M) = \frac{1}{N}\sum_{k=1}^{N}(\hat{f}^M(t_k) - f_k)^2 + \frac{2M}{N}{\sigma_e}^2 \qquad (11)$$

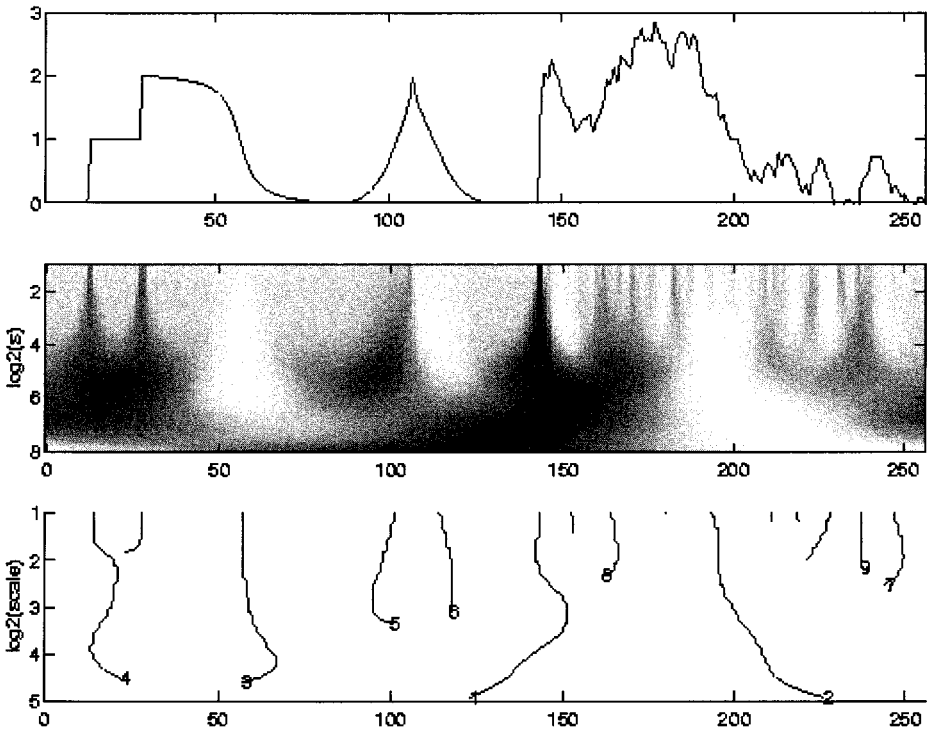

Fig. 4. Plot of (top) a signal with various edge types and (middle) its scale-space representation and (bottom) the modulus maxima extracted from (middle). Black, gray, and white points correspond to positive, zero, and negative wavelet coefficients, respectively.

where $\hat{f}$ is the wavelet network approximation, M is the number of wavelets in the network, N is the sample length of $f(t)$, and ${\sigma_e}^2$ is the variance of the noise e in the regression model [see (3). The model size is determined to minimize the GCV criterion. Zhang [4] also gives an iterative description for incorporating estimates of the noise variance ${\sigma_e}^2$ into the determination of the model order, M.

Similar to the GCV criterion, the novel objective function of (4) can be realized in discrete form by combining the global and *local* error using a diagonal positive definite weighting matrix, $D \in \mathbf{R}^{NxN}$, as shown in (12) below

$$F(M, \lambda, \delta, C) = e^T D e + \lambda \frac{M}{N} = \sum_{k=1}^{N} d_k {e_k}^2 + \lambda \frac{M}{N} \quad (12)$$

where $e = f - \hat{f}^M(w, t)$. The diagonal entries of D are given by

$$D_{ii} = \begin{cases} \dfrac{1}{\|f_L\|^2} & |u_j - i| \le C \quad \forall u_j \in U \\ \dfrac{1}{\|f_G\|^2} & \text{otherwise} \end{cases}$$

where u_j is the jth local feature with window size, C. $\|f_L\|^2$ is the sum of squares of the local window supports and $\|f_G\|^2$ is the sum of squares of the global window supports. These diagonal weighting terms normalize the local and global error with respect to their window support widths, thereby providing a better indication of the global and local error minimization during network construction. Since no diagonal element D_{ii} is zero, then D is positive definite and the quadratic error term $e^T D e$ is convex and has a local minima. Also, the model parsimony term is a monotonically increasing function with model size, M. Finally, it is interesting to note that one can view (12) as a weighted GCV function [see (11)], assuming that $\lambda = 2\sigma_e^2$. As a result, one can determine the λ parameter of (12) by iteratively estimating the noise variance, as noted above and in Zhang [4].

Now that the size of the network can be automatically determined, we discuss the automatic determination of the local feature set, U.

IV. Local Set Prioritization Using the WTMM Representation

To determine both the priority level δ and the abscissa location for each candidate local feature $u_j \in U$ [see (4)] algorithmically, one must first detect and characterize each of the local features. To this end, Mallat [14], [19] showed that the wavelet transform acts as a multiscale differential operator

$$Wf(u, s) = s^n \frac{d^n}{du^n}(f \star \theta_s)(u)$$

where $\psi = (-1)^n \theta^{(n)}$ and θ is typically chosen as a Gaussian. If one uses a wavelet with one vanishing moment, then the WTMM are the maxima of the first-order derivative of f smoothed by θ_s. Using a wavelet with two vanishing moments, the WTMM correspond to high curvature. By definition, the term *modulus maxima* describes any point (u_0, s_0) in scale-space such that $|Wf(u, s_0)|$ is a strict local maximum

in either the right or left neighborhood of $u = v$ [14]. This implies that

$$\frac{\partial Wf(v, s_0)}{\partial u} = 0$$

A *maxima line* is defined as any connected curve $s(u)$ in the scale-space plane (u, s) along which the points are *modulus maxima*, as shown in Fig. 4(c). As a result, the WTMM representation of a signal f can be used for detection and characterization of localized signal features, such as edges and local singularities.

A. Singularity Detection

Mallat and Hwang [14] showed that one is guaranteed to detect local singularities by following the WTMM at fine scales. Once detected, the singularity at point $t = v$ is characterized by the decay of the *modulus maxima* included in the cone $|u-v| \leq Cs$ of the scalogram. For example, Fig. 4(middle) shows the scalogram or time-scale representation for a signal with various edge types, as shown in Fig. 4(top) [14]. One observes that each of the edges and singularities are detected by following the modulus maxima lines to fine scales. Once an interesting feature has been detected, it must be characterized.

B. Singularity Characterization

In order to determine whether a candidate local signal feature (centered at a point $t = v$) is admissible to the feature set, U, it must first be characterized. If a signal $f(t)$ has a nondifferentiable singularity or edge at a point v in time, then the decay across scales of the WTMM can be used to estimate the local regularity of the signal within a neighborhood of v. Moreover, this decay is directly related to the local properties of the signal and can be characterized using Lipschitz, or Hölder, exponents [20], [14].

Definition: A function f is pointwise Lipschitz $\alpha \geq 0$ at v, if there exists $K > 0$, and a polynomial p_v of degree $m = \lfloor \alpha \rfloor$ such that

$$\forall t \in \mathbf{R}, \quad |f(t) - p_v(t)| \leq K|t - v|^{\alpha}. \tag{13}$$

Lipschitz exponents provide uniform and local regularity measures of a signal $f(t)$ within a local neighborhood of a point v in time. If f is uniformly Lipschitz $\alpha > m$ in the neighborhood of v, then one can verify that f is necessarily m times continuously differentiable in this neighborhood [14]. The Lipshitz regularity at v is the maximum slope of $\log_2 |Wf(u, s)|$ as a function of $\log_2(s)$ along a maxima line converging to v. To measure the local regularity of a signal, the wavelet must have $n > \alpha$ *vanishing moments* [see (14)]

$$\int_{-\infty}^{+\infty} t^k \psi(t)\, dt = 0 \quad \text{for } 0 \leq k < n \tag{14}$$

to provide an accurate estimate of the local scaling exponent, α [see (13)]. A wavelet with n vanishing moments is orthogonal to polynomials of degree $n - 1$.

To detect and characterize a local signal edge more accurately at a point v, Mallat [14] has shown that f is uniformly Lipschitz α in the neighborhood of v if and only if there exists $A > 0$ such that each modulus maxima (u, s) in the cone of influence satisfies (15).

$$|Wf(u, s)| \leq As^{\alpha+1}. \tag{15}$$

Thus, the local regularity can be characterized by finding the maximum slope of $\log_2 |Wf(u, s)|$ as a function of $\log_2(s)$ along a maxima line converging to v, using the following:

$$\log_2 |Wf(u, s)| \leq \log_2 A + (\alpha + 1) \log_2(s). \tag{16}$$

Thus, computing the local regularity of a candidate edge localized near a point $u = v$ can be determined using linear regression to estimate α. Once determined, the local feature set U can be constructed by selecting the subset of candidate edges or local features u_j for which $\alpha_j > \delta$. Using this local feature set, one can effectively segment the signal into its local $(\mathbf{\Gamma}^L f)$ and global $(\mathbf{\Gamma}^G f)$ components.

C. Signal Segmentation

The segmentation process can be summarized by the following steps:

1) transformation of the signal f into the time-scale domain using the nonorthogonal continuous wavelet transform (CWT) as shown in Fig. 4(middle);
2) extraction of the *ridges* or *modulus maxima* from this time-scale plot as shown in Fig. 4(bottom);
3) "pruning" of the weaker *maxima* or *ridges* from the initial set, by excluding those maxima which do not traverse a user-specified number of scale levels;
4) use of the dominant *maxima* to accurately determine the localization in time or abscissa of important signal events, u_j;
5) characterization of signal edges by estimating their local Lipschitz exponents, α_j;
6) construction of the local feature set $U = \{u_1, \ldots, u_P\}$ using the priority threshold, δ, such that $\alpha_j \geq \delta$.

Next, it is important to review some simulation studies that demonstrate the effectiveness of the aforementioned methodologies.

V. Simulation Results

We now demonstrate the effectiveness of the two novel approaches to focus the learning of the wavelet network on critical *local* features $u_j \in U$ of interest, thereby enabling the efficient minimization of the novel objective function [see (4)]. First, Section V-A compares the effectiveness of the adaptive initialization grid (Section III-A) to the standard initialization grid using the standard basis function selection scheme [see Table I, Method II]. Second, Section V-B compares the effectiveness of the new subset selection schemes (see Section III-B) to the standard global selection scheme in their respective abilities to minimize the error in the *local* regions of interest. Section V-C demonstrates the effectiveness of a combined strategy, involving both *focused* subset selection and the use of the adaptive initialization grids, to reduce the error in local regions of interest. Section V-D demonstrates automatic model order determination using the minimum of the new objective function. Alternatively,

Section V-E proposes a new ESA strategy to help visualize the minimization of both the global and *local* error as a function of model parsimony. This ESA technique serves to supplement the visualization of error minimization using the new objective function, which is cast in the form of a weighted quadratic as described in Section III-E. Finally, Section V-F shows results of the WTMM-based preprocessing procedure (see Section IV), which is used to construct the local feature set U in a more automated fashion.

The following simulations were conducted in MATLAB using Mallat's 1-D signal [14], which is shown in Fig. 4(top). This signal was chosen because of the various edge types that it exhibits. The wavelet networks were constructed using a radial form of the Sombrero or *Mexican-hat* wavelet [9], whose functional form is given by $\psi(t) = (1 - ||t||^2)e^{(-||t||^2/2)}$. For this simulation work, the MLSE [see (10)] and MGSE [see (9)] are calculated to provide comparisons with the standard MSE. To simplify matters, the subscripts (i.e., MLSE, MGSE and MSE) introduced in Table I to differentiate and to help clarify the different selection methods will be assumed, and therefore dropped, for subsequent tables in this section.

A. Network Initialization: A Comparison of Standard vs. Locally Adaptive Grids

A simulation study was conducted to compare the effectiveness of the adaptive and standard initialization grids in minimizing the local error, as defined by the local feature set, U. First, a wavelet network was constructed using a standard initialization grid (Fig. 1) using nine discrete scale or resolution levels. Next, a wavelet network was constructed using an adaptive initialization grid (Section III-A) with the following attributes:

- local feature set: $U = \{28, 107\}$;
- interval width: $C = 10$;
- resolution levels $(G = 5)$ for $\mathbf{W}_G$: $0 \leq m < 5$;
- resolution levels $(\rho = 3)$ for $\mathbf{W}_L$: $6 \leq m \leq 8$.

For comparison purposes, the total number of wavelet basis functions in each model was arbitrarily set at $M = 60$. Both networks were constructed using the standard global stepwise selection method for choosing wavelet basis functions (Method II $(\zeta = I)$ in Table I).

Table II shows the results of these three simulations, with the approximation and error plots shown in Fig. 2. Both the box and trapezoidal adaptive grids outperformed the standard grid by:

1) reducing computational complexity (i.e., # of FLOPS) of network initialization by a factor of $\sim$3–4 due to the reduction in the starting number of wavelet bases by a factor of five;
2) reducing *local* error by a factor of 30 to 60, or more than an order of magnitude;
3) allocating more wavelet basis functions (or wavelons) from the local set $\mathbf{W}_L$ to facilitate local error minimization.

However, the trapezoidal adaptive grid may overfit near the local regions, $u_j \in U$, as seen in Fig. 2(bottom left and right). Future studies are needed to optimize this trapezoidal profile for reduced overfitting near the local regions. Finally, it should be noted that the reduced computational burden for network initialization, i.e., using the adaptive grids, should prove advantageous for large-sample or multidimensional data sets. Moreover, basis functions can be initialized in regions where they will be most needed, i.e., as defined by the local feature set, U.

Now that the network initialization strategies have been compared, the next section will compare the different selection schemes used for network construction, as described in Section III-D

B. Network Construction: A Comparison of Basis Function Selection Methods I-IV

A series of wavelet networks was constructed to compare and contrast the different basis function selection methods, as described in Section III-D. This set of simulations compares the effectiveness of the various projection operators in focusing the network construction effort in the regions defined by the local feature set, U. Each wavelet network was constructed to a size of $M = 60$ wavelet basis functions. All four networks utilized nine resolution levels $(j = 9)$ of the standard dyadic initialization grid, and therefore started with $L = 98$ basis functions (Section III). The local feature set was chosen as $U = \{28, 107\}$ with an interval width of $C = 10$. The results of these simulations are shown in Table III. While the adaptive grids were not used in these simulations, the results of a combined simulation study, using both standard and adaptive grid techniques and the subset selection strategies, is detailed in the next section [Section V-C].

Table III shows that the localized projector (Method I) does not do a great job of minimizing the local error (MLSE). However, it does beat the standard global method (Method II). This is most likely due to an overemphasis on local error minimization, thereby causing overfitting. The combined local and global projectors (Methods III and IV) provide a good tradeoff between minimization of the local and global error, with the alternating projector (Method IV) performing the best of all four methods. Moreover, its local error performance (MLSE) is an order of magnitude better than the standard method (Method II). Finally, it should be noted that all three of the nonstandard methods do a better job of minimizing the local error (MLSE) than the standard method.

Now that the strategies for network initialization and construction have been compared independently of one another, the next section will investigate the strategies in combination.

C. Combined Strategy: Adaptive Initialization Grid and Localized Selection Methods

A series of wavelet networks was constructed to investigate the effect of using a combined approach, i.e., using both the adaptive initialization grids and the four basis function selection methods. The results are shown for the adaptive box-like and trapezoidal grids in Tables IV and V, respectively.

Table IV shows that local selection method I has the worst global error performance (in terms of MSE and MGSE), but the best local error performance (in terms of MLSE), when using the adaptive box-like grid (Fig. 1). This is to be expected, due to the greater emphasis on the local error, and due to the greater

TABLE IV
COMBINED STRATEGY: COMPARISON OF SELECTION METHODS I–IV USING AN ADAPTIVE BOX-LIKE INITIALIZATION GRID ($M = 60, C = 10, U = \{28, 107\}$)

Selection Method	# of Local Bases ($v_{l_i} \in \mathbf{W}_L$)	MSE (x 10^{-3})	MGSE (x 10^{-3})	MLSE (x 10^{-4})
(I) Local	25	8.02	8.77	0.39
(II) Global	16	1.66	1.80	1.03
(III) Local + Global	20	3.30	3.58	2.84
(IV) Alt. Local + Global	22	2.96	3.23	0.58

TABLE V
COMBINED STRATEGY: COMPARISON OF SELECTION METHODS I–IV USING AN ADAPTIVE TRAPEZOIDAL INITIALIZATION GRID ($M = 60, C = 10, U = \{28, 107\}$)

Selection Method	# of Local Bases ($v_{l_i} \in \mathbf{W}_L$)	MSE (x 10^{-3})	MGSE (x 10^{-3})	MLSE (x 10^{-4})
(I) Local	36	19.2	21.0	1.63
(II) Global	23	2.69	2.92	2.64
(III) Local + Global	24	2.72	2.86	12.6
(IV) Alt. Local + Global	27	3.30	3.51	10.7

number of wavelet basis functions that are utilized from the local set of wavelet basis function, $\mathbf{W}_L$. Conversely, global selection method II has the best global error performance (in MSE), and is next to worst in terms of local error performance (in MLSE). This lower performance in terms of local error directly correlates with the reduced number of wavelets chosen from the local set, $\mathbf{W}_L$. The alternating selection method IV provides the best tradeoff between local and global error performance.

Table V shows that we obtain similar results when using the adaptive trapezoidal grid, with methods I and II efficiently minimizing their respective local and global error targets. However, selection methods III and IV demonstrate a degradation in local error performance (in MLSE) as compared to the adaptive box-like initialization grid, due in part to localized overfitting and in part to interactions between the box-like projection operator (Fig. 3) and the trapezoidal adaptive initialization grid. To investigate this potential cause and mitigate the performance degradation, the trapezoidal projector described in Section III-B was used along with the adaptive trapezoidal initialization grid. Simulation results are shown in Table VI.

Upon comparing Tables IV and V, we observe that the degradation in local error performance is reduced by using a more tapered trapezoidal projection operator. However, further work is needed to study its exact shape and performance with regard to: 1) the number of local features P and 2) the proximity of local features u_j and u_{j+1} in the local feature set U.

Now that the initialization and selection strategies have been discussed, it is important to study the automatic determination of the network size, M.

TABLE VI
COMBINED STRATEGY: COMPARISON OF SELECTION METHODS I–IV USING AN ADAPTIVE TRAPEZOIDAL INITIALIZATION GRID AND TRAPEZOIDAL PROJECTION OPERATOR FOR SELECTION ($M = 60, C = 10, U = \{28, 107\}$)

Selection Method	# of Local Bases ($v_{l_i} \in \mathbf{W}_L$)	MSE (x 10^{-3})	MGSE (x 10^{-3})	MLSE (x 10^{-4})
(I) Local	36	12.1	13.2	2.22
(II) Global	23	2.69	2.92	2.64
(III) Local + Global	22	3.10	3.32	7.95
(IV) Alt. Local + Global	28	2.86	3.11	2.54

D. Automated Model Order Determination

To automatically determine the number of wavelet basis functions in the wavelet network, the new objective function (Section III-E) was plotted as a function of the number of wavelets in the model. A series of four wavelet networks were initialized, using both standard and adaptive network initialization strategies, and constructed using the four network construction strategies (Methods I-IV). Prior to this study, the variance of the noise was estimated and the parameter $\lambda = 2{\sigma_e}^2$ was determined to be approximately 2.8×10^{-3}. In each case, the minimum of the new objective function was used to specify the model order, with results shown in Table VII and Fig. 6.

Fig. 6 shows the corresponding plot of the new objective function [see (4)] as a function of the number of wavelets in the model, for the results in Table VII. The three new selection methods (Methods I, III, and IV) each outperform the standard global selection method (Method II) in terms of:

1) local error minimization, as defined by the local feature set, U;
2) speed at which the minimum of the objective function is achieved;
3) model parsimony, using the minimum of the objective function.

To better visualize the performance of the automatic model order determination procedure, the authors propose the use of the following novel ESA technique.

E. ESA: A New Technique for Visualizing Local and Global Error Minimization and Model Parsimony

To assist in the evaluation of different wavelet network construction or initialization strategies, we propose a new ESA technique. The basic idea behind ESA is to view the wavelet network error by plotting the local network error, i.e., $\|f - \hat{f}_L\|^2$, as a function of the global network error, i.e., $\|f - \hat{f}_G\|^2$. During network construction, the network error as a function of model terms, M, corresponds to a curve in this *error space*. At the start of network construction, the error curve starts in the upper right corner and progresses down and to the left as the model size, M, increases. By observing the vertical and horizontal components of the curve trajectory, one can ascertain the degree to which the local and global

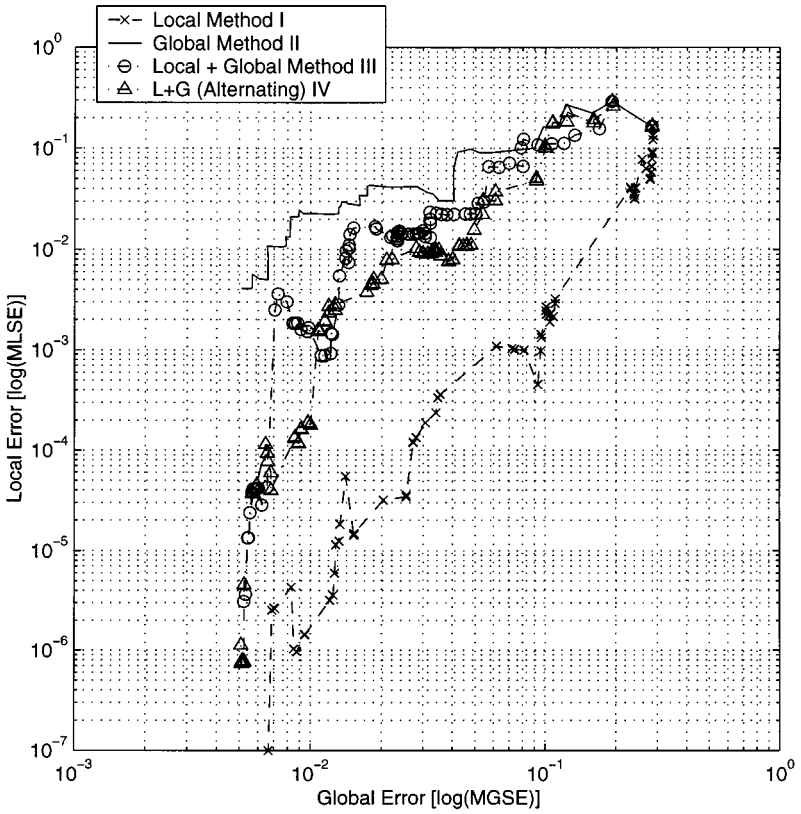

Fig. 5. ESA showing the MLSE as a function of the MGSE for the four basis function selection methods (I–IV) during automatic model order determination.

error are reduced with increasing model terms or wavelet basis function, M. In addition, the ESA technique can provide the user with valuable information regarding the efficiency of a given initialization or construction strategy, especially in terms of model parsimony and error minimization.

For example, Fig. 5 shows an ESA plot for the wavelet network approximations from the automatic model order determination study (Section V-D, Fig. 6). One observes that all three of the nonstandard selection methods (I, III, and IV) outperform the standard global selection method (II) in terms of local error minimization by several orders of magnitude. In addition, the global error (MSE and MGSE) is comparable to that attained by the global selection method (II). Finally, it should be noted that the ESA technique is also useful for assessing the efficiency with which the desired error target is being addressed as each wavelet term is added to the model.

Now it is important to discuss a methodology for automatically determining the local feature set, U.

F. Automated Local Feature Set Generation Using the WTMM Representation

The simulation work in the previous sections was conducted using a local feature set, U, that was specified *a priori* by the user, i.e., it assumed some prior knowledge on the number and significance of edge types in the signal. For large or multidimensional signal databases, such a strategy may prove intractable.

Therefore, to help determine the local feature set, U, in more an automated fashion, the signal shown in Fig. 4(a) was segmented according to the steps shown in Section IV-C. Once the modulus maxima were determined, as shown in Fig. 4(b), the first six dominant maxima lines were chosen. (This number is problem dependent.) Subsequently, the local Lipschitz exponent, α_j, for each maxima line was estimated using linear regression according to (16). The regression was performed within the linear portion ($s = 2$ to 4) of the amplitude of the maxima line $\log_2(|Wf(u,s)|)$ versus $\log_2(s)$. The results are shown in Table VIII.

Next, the maxima were pruned based on a priority threshold of $\delta = \bar{\alpha}$. In other words, maxima whose local exponent was greater than the average of the top six positive exponents were chosen

$$U = \{u_{j,\delta}\colon \alpha_j > 0 \text{ and } \alpha_j \geq \bar{\alpha} = 0.43\}.$$

Using the remaining modulus maxima, the maxima were followed to small scales ($s = 2$), and the abscissa locations were then estimated to be $U = \{57, 104, 111, 190\}$. The window size, C, was set to be the minimum of all half-distances between potential abscissa locations

$$C = \left\lceil \min_{j \in \{1,\ldots,P-1\}} \left(\frac{u_{j+1} - u_j}{2} \right) \right\rceil = 4.$$

Table IX shows the results of using the combined strategy (Section V–C) to minimize the local error for this new feature set and window size, along with the automatic model order generation procedure (Section V-D). One observes that the three

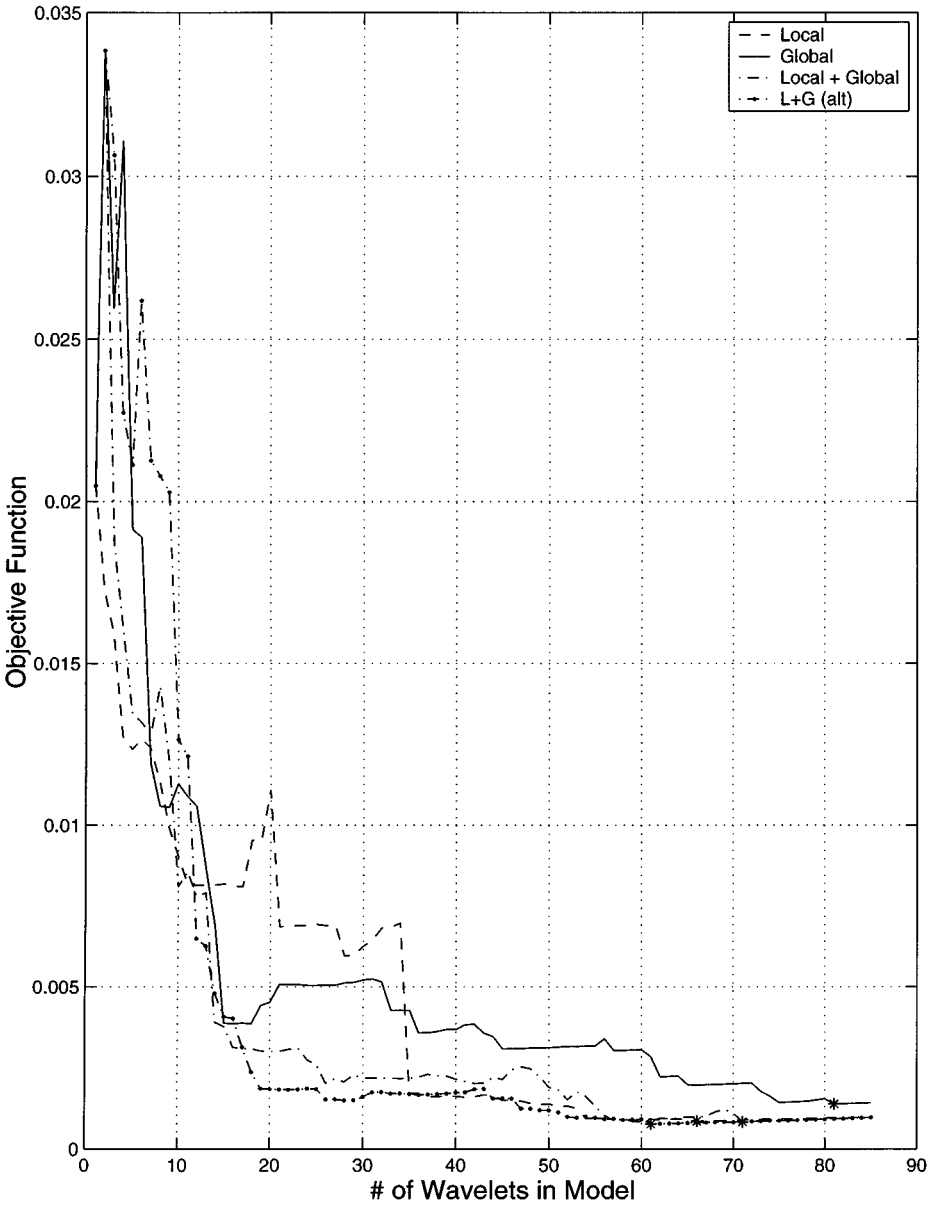

Fig. 6. Plot of the new objective function $F(M, \lambda, \delta, C)$ as a function of the number of wavelet terms in the model. It is important to note that all three of the new selection methods outperform the standard methodology in terms of model parsimony and combined local and global error. The asterisks ('$*$') denote the minimum of the objective function and therefore the number of wavelets in the network.

TABLE VII
AUTOMATIC MODEL ORDER DETERMINATION USING THE MINIMUM OF THE NEW OBJECTIVE FUNCTION ($C = 10, U = \{28, 107\}$)

Grid Type	Selection Method	Start #	Model Order	MSE (x 10^{-3})	MGSE (x 10^{-3})	MLSE (x 10^{-3})
Adaptive	(I) Local	98	66	4.50	4.92	0.0046
Standard	(II) Global	511	81	1.75	1.79	1.32
Adaptive	(III) Local + Global	98	71	1.84	2.01	0.0092
Adaptive	(IV) Alt. Local + Global	98	61	2.88	3.15	0.061

new nonstandard selection methods (I, III, and IV) improve the local error for the automatically generated local feature set, U. In addition, they afford comparable performance in terms of the global error, as measured by both the MSE and MGSE.

While this strategy succeeded to provide a local feature set, U, there are several problems which necessitate further work. First, the calculation of the nonorthogonal CWT is computational intensive, especially for large or multidimensional sig-

TABLE VIII
ESTIMATES FOR THE LOCAL LIPSCHITZ EXPONENTS (α) IN THE MALLAT SIGNAL

Maxima Line (j)	Estimated Abscissa Location (u_j)	Local Exponent (α_j)
1	144	-0.09
2	190	0.43
3	57	0.60
4	13	0.09
5	104	0.44
6	111	0.57

nals. Second, the proximity of signal edges to one another can complicate the extraction of their respective abscissa locations, especially causing the extinction of important edge types. For example, the step edge at abscissa location $u = 28$ cannot be successfully detected due to the termination of its modulus maxima at small scales. This is most probably due to its proximity to the edge at $u = 14$, whose cone of influence in scale space can disrupt the propagation of maxima from the edge at $u = 28$ to larger scales in scale space. As a result, further work is needed to improve this edge detection strategy and the determination of the priority threshold, δ.

VI. CONCLUSION

This paper presented a novel objective function that incorporates both global and *local* error as well as model parsimony in the construction of wavelet neural networks. During network initialization, an adaptive dyadic discretization grid was utilized to help reduce the *local* error within the vicinity of a finite set of local regions, as defined by the local feature set, U. This set of local regions can be either user-defined or determined using regularity measurements derived from the WTMM representation to prioritize local features according to a priority threshold, δ. In addition, a modified stepwise selection procedure was presented to help *focus* the selection of wavelet basis functions during network construction to reduce *local* error. Simulation results demonstrated the effectiveness of these new methodologies in minimizing the *local* and global error with fewer wavelet basis functions. In addition, these methodologies provide a net reduction in computational effort or complexity, especially when using the adaptive initialization grids. Moreover, these computational savings motivate the direct transfer of these methodologies to large-signal or multidimensional data sets, especially for *focused* learning within local hyper-regions of the data space. However, further work is needed to optimize the local and global projection operators, especially with regard to localized Gibbs-like oscillations and overfitting [4]. This is critical to modeling very large size datasets. Further work is also needed to explore the use of parallel computing techniques for constructing wavelet neural networks to model such large datasets. Finally, the results and methodologies presented in this paper are viewed by the authors to be directly applicable to the other applications in which WNNs are currently being employed (as reported in Section II), including modeling and classification as well as system identification and related control tasks.

APPENDIX I

The following algorithm for *stepwise selection by orthogonalization* can be found in [4]. It provides the groundwork and motivation for the modified selection algorithm developed in Section III-B.

Algorithm 1: **Step** 1) Set

$$I_1 = \{1, 2, \ldots, L\}$$

find

$$l_i = \arg\max_{j \in I_1} \left(v_j{}^T f\right)^2$$

and set

$$\begin{aligned} \tilde{u}_{l_i} &= v_{l_i}^T f \\ w_{l_i} &= v_{l_i} \\ \alpha_{11} &= 1 \\ p_j{}^{(1)} &= v_j, \quad j = 1, \ldots, L, \quad j \neq l_i. \end{aligned}$$

Step i $(i = 2, \ldots, M)$

$$I_i = I_{i-1} - \{l_{i-1}\}$$

and for each $j \in I_i$, compute

$$\begin{aligned} p_j{}^{(i)} &= p_j{}^{(i-1)} - \left(v_j{}^T w_{l_{i-1}}\right) w_{l_{i-1}} \\ I_i &= I_i - \left\{j : p_j{}^{(i)} = 0\right\} \end{aligned}$$

find

$$l_i = \arg\max_{j \in I_i} \frac{\left(\left(p_j{}^{(i)}\right)^T y\right)^2}{\left(p_j{}^{(i)}\right)^T p_j{}^{(i)}}$$

and set

$$\begin{aligned} w_{l_i} &= \left(\left(p_{l_i}{}^{(i)}\right)^T p_{l_i}^{(i)}\right)^{-\frac{1}{2}} p_{l_i}^{(i)} \\ \tilde{u}_{l_i} &= w_{l_i}^T y \\ \alpha_{ki} &= v_{l_i}^T w_{l_k}, \quad k = 1, \ldots, i-1 \\ \alpha_{ii} &= \left(\left(p_{l_i}{}^{(i)}\right)^T p_{l_i}^{(i)}\right)^{-\frac{1}{2}}. \end{aligned}$$

ACKNOWLEDGMENT

Software used for wavelet analysis includes Dr. Q. Zhang's wavelet network toolbox (v2.1), and Stanford University's wavelet toobox (WAVELAB v802), which is available at www-stat.stanford.edu/~wavelab.

TABLE IX
AUTOMATIC MODEL ORDER DETERMINATION FOR AUTOMATICALLY GENERATED FEATURE SET ($C = 4, U = \{57, 104, 111, 190\}$)

Grid Type	Selection Method	Start #	Model Order	MSE (x 10^{-3})	MGSE (x 10^{-3})	MLSE (x 10^{-6})
Adaptive	(I) Local	91	83	1.65	1.79	1.85
Standard	(II) Global	511	81	1.75	1.83	779
Adaptive	(III) Local + Global	91	57	1.95	2.12	31.7
Adaptive	(IV) Alt. Local + Global	91	78	1.71	1.85	16.9

REFERENCES

[1] J. Jin and J. Shi, "Feature-preserving data compression of stamping tonnage information using wavelets," *Technometrics*, vol. 41, no. 4, pp. 327–339, 1999.

[2] L. Martell, "Wavelet-Based Data Reduction and De-Noising Procedures," Ph.D. dissertation, North Carolina State Univ., Raleigh, 2000.

[3] L. Bottou and V. Vapnik, "Local learning algorithms," *Neural Comput.*, vol. 4, pp. 888–900, 1992.

[4] Q. Zhang, "Using wavelet network in nonparametric estimation," *IEEE Trans. Neural Networks*, vol. 8, pp. 227–236, Mar. 1997.

[5] B. Bakshi and G. Stephanopoulos, "Wave-net: A multiresolution, hierarchial neural network with localized learning," *AIChE J.*, vol. 39, no. 1, pp. 57–81, 1993.

[6] N. Saito, "Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion," in *Wavelets in Geophysics*, E. Foufoula-Georgiou and P. Kumar, Eds. New York: Academic, 1994, pp. 299–324.

[7] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shirinkage," *J. Amer. Statist. Assoc.*, vol. 90, pp. 1200–1224, 1995.

[8] N. Sundarajan and P. Saratchandran, Eds., *Parallel Architectures for Artificial Neural Networks: Paradigms and Implementations*. Los Alamitos, CA: IEEE Comput. Soc. Press, 1998.

[9] J. Echauz, "Wavelet Neural Networks for EEG Modeling and Classification," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, 1995.

[10] Q. Zhang and A. Benveniste, "Wavelet networks," *IEEE Trans. Neural Networks*, vol. 3, pp. 889–898, Nov. 1992.

[11] Y. Harkouss, E. Ngoya, J. Rousset, and D. Argollo, "Accurate radial wavelet neural-network model for efficient CAD modeling of microstrip discontinuities," *Proc. Inst. Elect. Eng. Microwaves, Antennas, Propagation*, vol. 147, no. 4, pp. 277–283, 2000.

[12] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, NJ: Prentice Hall, 1999.

[13] R. M. Sanner and J. E. Slotine, "Structurally dynamic wavelet networks for adaptive control of robotic systems," *Int. J. Contr.*, vol. 70, no. 3, pp. 405–421, 1998.

[14] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. New York: Academic, 1999.

[15] S. Chen, E. S. Chng, and K. Alkadhimi, "Regularized orthogonal least squares algorithm for constructing radial basis function networks," *Int. J. Contr.*, vol. 64, no. 5, pp. 829–837, 1996.

[16] S. Chen, Y. Yu, and B. L. Luk, "Combined genetic algorithm optimization and regularized orthonormal least squares learning for radial basis function networks," *IEEE Trans. Neural Networks*, vol. 10, pp. 1239–1243, Sept. 1999.

[17] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. Signals, Syst., Comput.*, vol. 1, 1993, pp. 40–44.

[18] S. Chen, S. Billings, and W. Luo, "Orthogonal least squares methods and their application to nonlinear system identification," *Int. J. Contr.*, vol. 50, no. 5, pp. 1873–1896, 1989.

[19] A. P. Witkin, "Scale-space filtering," in *Proc. 8th Int. Joint Conf. Artificial Intell.*, vol. 2, 1983, pp. 1019–1022.

[20] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

Eric A. Rying (S'00) received both the B.S. and M.S. degrees in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA in 1995. He received the Ph.D. degree in electrical engineering from North Carolina State University (NCSU), Raleigh, in 2001.

In 1996, he was a Team Member and subsequently the Team Leader in 1997 for the Semiconductor Research Corporation Continuous Quality Improvement (CQI) Programs at NCSU. He is currently an Engineer for PDF Solutions, Inc, San Jose, CA. His interests include yield modeling, wavelets, and their applications, wavelet neural networks, advanced equipment, and process control, run-to-run control and fault detection and classification in semiconductor manufacturing, especially for selective silicon epitaxy and rapid thermal chemical vapor deposition (RTCVD) of ultrathin films. He holds one U.S. patent through NCSU.

Dr. Rying is a member of Eta Kappa Nu. In 1998, he received a U.S. Department of Education Graduate Assistance in Areas of National Need (GAANN) Fellowship through the NCSU National Science Foundation Engineering Research Center for Advanced Electronic Materials Processing (AEMP).

Griff L. Bilbro (M'85–SM'94) received the B.S. degree in physics from Case Western Reserve University, Cleveland, OH, and the Ph.D. degree in 1977 from the University of Illinois, Urbana-Champaign, where he was a National Science Foundation Graduate Fellow in Physics.

He designed computer models of complex systems in industry until 1984, when he accepted a research position at North Carolina State University (NCSU), Raleigh. He is now a Professor of Electrical and Computer Engineering. He has published in image analysis, global optimization, neural networks, microwave circuits, and device physics. His current interests include analog circuits and cathode physics.

Jye-Chyi Lu received the Ph.D. in statistics from University of Wisconsin, Madison, in 1988

He joined the faculty of North Carolina State University (NCSU), Raleigh, where he remained until 1999 when he joined ISyE. He is a professor in the School of Industrial and Systems Engineering (ISyE). He is very active in promoting research, education and extension-service programs with focus on information systems engineering, e-business, e-design, and industrial statistics areas. He has published about 40 journal papers in these areas.

Dr. Lu serves as an Associate Editor of the IEEE TRANSACTIONS ON RELIABILITY.