Wavelet-Based Modulation in Control-Relevant Process Identification

John F. Carrier and George Stephanopoulos
Massachusetts Institute of Technology, Cambridge, MA 02139

The identification of models from operating data for process controller design requires that the information from the process be extracted in pieces that are localized in both time and frequency. Such an extraction process would allow the separation of valuable signal information from the effects of nonstationary disturbances and noise. The wavelet transform provides an efficient approach for such a decomposition, which is organized in a multiscale, hierarchical fashion. By using the method of modulating functions in conjunction with the wavelet decomposition, it is demonstrated that recursive state-space models, which are multiscale in character and suitable for the design of model-predictive controllers, may be readily constructed with lower levels of modeling error than yielded by traditional techniques. The method is especially suitable for the identification of time-varying and nonlinear models, where the nonlinear process is represented by a set of linear models. The multiscale character of the wavelet basis makes it particularly suitable for multirate, multivariable processes. A series of examples illustrates various aspects of the proposed approach and its inherent advantages.

Introduction

The modern view of controller design is based on the theoretically sound premise of the internal model principle (Bengtsson, 1977; Desoer and Wang, 1980), which requires that a controller should provide (a) an effective inversion of process dynamics, and (b) a direct generation of the external disturbance's structure. Such a viewpoint has led to fairly successful industrial deployment of large-size multivariable *model-predictive* control schemes (Cutler and Ramaker, 1980; Qin and Badgwell, 1997) and has underlined the fact that generation of the requisite process models and associated model uncertainty is the most important element of controller design. As a consequence, *control-relevant* process model generation has attracted significant attention and is currently one of the most active areas of research (Gaikwad and Rivera, 1995; Ling and Rivera, 1995; Braatz and Mijares, 1995).

The construction of a feedback controller presents the following dilemma: the stability, performance, and noise and disturbance rejection properties of a feedback loop are best understood in the frequency domain, yet time-varying and nonlinear effects render the frequency domain analysis invalid over time scales where these effects become appreciable. The objective of this work is to trade off the benefits of a frequency-domain analysis of the process behavior with the ability to detect time-varying and nonlinear effects, and to incorporate this information into the control scheme.

The realization that the open-loop crossover frequency behavior of the process is most critical in terms of closed-loop performance was recognized very early by Ziegler and Nichols (1942), whose famous tuning rules for PID controllers are based only on information at the open-loop crossover frequency. Åström and Wittenmark (1989) provide an excellent example on the behavior of different processes under unity feedback. The results show that processes that have similar crossover frequency responses give similar closed-loop responses regardless of their open-loop behavior, whereas processes that have different crossover frequency responses display very different behaviors under closed-loop control, regardless of similarities in the open-loop behavior.

Recent work in the development of feedback controllers for LTI systems reconfirms the use of a frequency-domain approach (data prefiltering) to minimize closed-loop error in the presence of modeling errors and disturbances (Rivera, 1991; Muske and Rawlings, 1993; Palavajjhala et al., 1996). For these more advanced schemes, knowledge over a frequency *region* rather than at a single point is required. This can be more clearly understood by interpreting the purpose of the controller in a model-based framework (i.e., IMC); the controller is trying to increase the closed-loop bandwidth as much as possible in the face of physical and modeling constraints. Since the open-loop behavior in the frequency domain is initially unknown, the tasks of system identification and controller design form a process that is "inevitably iterative in nature" (Skelton, 1989). Therefore a system identification technique where rapid and nonredundant manipulation of the data in different frequency regions can be achieved would be of great merit when used in such an iterative framework.

However, real processes are inherently time varying. The required system identification technique must therefore be able to discount or discard data as they become "old" (that is when they no longer reflect the process behavior), or over temporal regions corrupted by nonstationary noise and disturbances. Consequently, any process identification scheme should not only be able to extract information from input–output data at specific frequencies (or, frequency ranges), but it should also be capable of extracting that information from segments of records of input–output data that are localized in time. These two requirements imply that the desired methodology for system identification should be capable of constructing accurate models using input–output information that has been extracted from specific *time* and *frequency* regions. Traditional system identification techniques (least squares, Kalman filters) that utilize input–output information that is localized in time, cannot localize the content of the input–output data in specific frequency ranges. Similarly, frequency-based identification techniques, which can perfectly localize the content of input–output data on the frequency axis, cannot deconvolve the temporal interactions of data and thus are global in time (Ljung, 1987). Consequently, neither of these two classes of system identification techniques can handle satisfactorily the frequency and time-localized needs for controller design. In this article, we propose a system identification approach that is able to construct models by extracting the requisite information from input–output data over selected regions of frequency and time. The proposed methodology is based on the theory of wavelet transform, which allows the decomposition of any square integrable function onto a set of basis functions, local in time and in frequency.

At this point, the relationship between basis functions and system identification methods may not be clear. In the following section this relationship is clarified by viewing the common system identification techniques *as a series of projection operations onto a set of basis functions*, and the properties of the model constructed by such a technique are determined by the properties of the basis functions in the set. This observation reduces the selection of a proper system identification technique to the proper selection of basis functions. The third section introduces the theory of wavelet decomposition, which provides a set of basis functions local in time and in frequency (Bakshi and Stephanopoulos, 1994; Rioul and Verletti, 1991). The fourth section shows how wavelets can be used for system identification by orthogonal matrix transformation, while the fifth section compares the performance of the wavelet identification techniques with conventional methods for linear time-invariant (LTI) and linear time-varying (LTV) systems and systems corrupted with noise and disturbances. The sixth section extends the methodology developed in the fourth section to nonlinear systems, and illustrates the approach using a nonisothermal exothermic CSTR. Finally, the use of wavelets for the identification of multirate (for which it is most suitable) and multivariable LTI systems is addressed in the seventh section.

The Role of Basis Functions in System Identification

Consider the standard identification problem of finding the parameters of an nth-order strictly causal ARMAX model

$$y(t)+a_1y(t-1)+\cdots+a_ny(t-n)=b_1u(t-1)+\cdots+b_mu(t-m) \quad (1)$$

using available data on the input and output variables, $[y(t), u(t)]$ $t=0, \ldots, N$, where $N \geq m+n-1$. The *objective* is to select the model parameters so as to provide a process representation that can then be used for the design of an "effective" feedback controller, that is, a controller that optimizes the performance of the feedback loop (for the given amount and quality of the input–output information). The set of relationships resulting from Eq. 1, applied over $(N+1)$ time points, represents (usually) an overdetermined system of linear equations, and some form of a least-squares solution will be required. If the data on $y(t)$ and $u(t)$ do not contain any extraneous noise or disturbances, then any identification method, for a given model structure, would yield essentially the same model. Therefore, any differences among the various system identification techniques depend on how these techniques handle the extraneous noise and disturbances contained in $y(t)$ and/or $u(t)$.

It is important to note that any extraneous noise or disturbance is characterized by its content in various frequencies and by the evolution of this content over time. For example, the measurement noise of $y(t)$ may be characterized by high frequencies during the first segment of a data record and gradually shift to lower frequencies over time. An unmeasured disturbance, with slowly varying low-frequency characteristics, may be affecting the values of $y(t)$. In both cases, the process identification methodology should be capable of extracting and rejecting the undesirable effects of the noise and disturbance, effects whose frequency content changes in both the frequency range and over time. Let us now develop the formalism that will allow us to present what any system identification technique attempts to do in handling the effects of extraneous noise and disturbances, and establish the framework in which we will describe the methodologies developed in this article.

The linear equations derived from the ARMAX model of Eq. 1 can be put into the following matrix form

$$\begin{bmatrix} y(n) \\ y(n+1) \\ \cdot \\ \cdot \\ \cdot \\ y(N) \end{bmatrix}$$

$$-\begin{bmatrix} y(n-1) & \cdot & y(0) & u(n-1) & \cdot & u(n-m) \\ y(n) & \cdot & y(1) & u(n) & \cdot & u(n-m+1) \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ y(N-1) & \cdot & y(N-n) & u(N-1) & \cdot & u(N-m) \end{bmatrix} \times \begin{bmatrix} -a_1 \\ \cdot \\ -a_n \\ b_1 \\ \cdot \\ b_m \end{bmatrix} = 0 \quad (2a)$$

or, equivalently

$$y - Ab = 0. \quad (2b)$$

This set of linear equations may be transformed into an equivalent set via an orthogonal transformation. Let P be an orthogonal operator, normalized to unity. Project Eqs. 2b onto the orthonormal basis, represented by the rows of P, and take

$$P[y - Ab] = 0. \quad (3)$$

Representing the matrix, P, as a column vector of its orthonormal set of rows, and the input–output data matrix, A, as a row vector containing the input–output data, that is,

$$P \equiv \begin{bmatrix} p_1^T \\ p_2^T \\ \cdot \\ \cdot \\ \cdot \\ p_N^T \end{bmatrix}; \qquad A \equiv [y_{n-1} \cdots y_0 u_{n-1} \cdots u_{n-m}]; \qquad y \equiv y_n,$$

then the ith equation of the set of Eqs. 3 is given by:

$$[\langle p_i^T, y_n\rangle + a_1\langle p_i^T, y_{n-1}\rangle + \cdots + a_n\langle p_i^T, y_0\rangle - b_1\langle p_i^T, u_{n-1}\rangle - \cdots - b_m\langle p_i^T, u_{n-m}\rangle] = 0. \quad (4a)$$

From Eq. 4a we can easily see that each transformed equation is associated with a single eigenrow (eigenvector) of the orthonormal projection operator, P, and that the input–output information contained in this equation is determined solely by the properties of its associated eigenrow (eigenvector). In other words, in Eq. 4a all the process data have been projected against a single eigenvector of the orthonormal transformation, and thus it contains only information in the subspace defined by the eigenvector p_i.

The relative effect of the information in the subspace spanned by the eigenvector p_i upon the predictive accuracy of the modeling relationship (Eq. 4) can be adjusted through the value of a weight $c_i^{1/2}$, that is,

$$c_i^{1/2}[\langle p_i^T, y\rangle + a_1\langle p_i^T, y_{n-1}\rangle + \cdots + a_n\langle p_i^T, y_0\rangle - b_1\langle p_i^T, u_{n-1}\rangle - \cdots - b_m\langle p_i^T, u_{n-m}\rangle] = 0 \quad (4b)$$

or, equivalently for the set of Eq. 3, by the entries $C_i^{1/2}$ of the diagonal matrix, C, that is,

$$CP[y - Ab] = 0. \quad (5)$$

The projection of a differential (or difference) equation against a general function (or finite dimension vector) as part of a parameter identification procedure, is referred to as the *Method of Modulating Functions* (Shinbrot, 1957). Some of the suggested modulating functions have been sinusoids (Pearson and Lee, 1985; Co and Ydstie, 1990) and cubic splines (Matelinsky, 1979; Preisig and Rippin, 1993a,b,c). Specific to the present work is the use of orthogonal functions with localized time and frequency characteristics as modulating functions, for the purposes of identifying models for feedback control.

Once the appropriate projection operator, P, and the weighting matrix, C, have been selected, in order to achieve the desired separation and relative weighting of the information contained in the records of input–output data, which is relevant to process dynamics from the irrelevant ones, the least-squares solution to Eq. 5 yields the values of the unknown model parameters, that is,

$$b = [A^T P^T C P A]^{-1} A^T P^T C P y. \quad (6)$$

At this point it is instructive to illustrate the preceding concepts using standard, well-known identification techniques in the frequency (spectral methods) and time (least-squares, Kalman filtering) domains.

Case 1: Frequency-Domain Identification Techniques (Spectral Methods). Let the projection operator, P, in Eq. 3 be the complex exponential matrix, Z, whose rows are composed of the complex sinusoids of the discrete Fourier series. Then

$$Z[y - Ab] = 0$$

and the elements $e^{(2\pi i)/(N+1)t}$, $t = 0,1,2,\ \ldots,\ N$ define the corresponding ith eigenvector (i.e., the equivalent of p_i in Eq. 4). The net result is that each of the original linear equations has been projected against a single frequency.

The frequency-localized structure of the transformed linear equations allows the information at frequencies irrelevant to the modeling task to be discounted or removed in a straightforward manner via the diagonal weighting matrix C. For instance, if C is chosen to be

$$C = \mathrm{diag}([\,|H(0)|^2\ |H(i\omega_0)|^2\ |H(2i\omega_0)|^2 \cdots]),$$

the result is identical to prefiltering the data with a filter, $H(i\omega)$, before performing a common least-squares parameter estimation.

Case 2: Time-Domain System Identification (Common Least-Squares, Kalman Filtering). It is easy to see that, since the equations of the ARMAX model (see Eq. 2) are already localized in the time domain, the required orthogonal projection operator is simply the identity matrix, that is, $P = I$. Furthermore, as can be seen from Eq. 2a, when the row number of the equation increases, the difference between the temporal behavior represented by that equation and the present time decreases. Therefore, old information may be discounted by decreasing the weight of equations with low row numbers. If C is selected to be

$$C = \mathrm{diag}([\cdots \lambda^3\ \lambda^2\ \lambda\ 1]),$$

where $0 < \lambda < 1$, the resulting procedure is identical to Kalman filtering with exponential forgetting.

Development of a system identification method for non-LTI processes

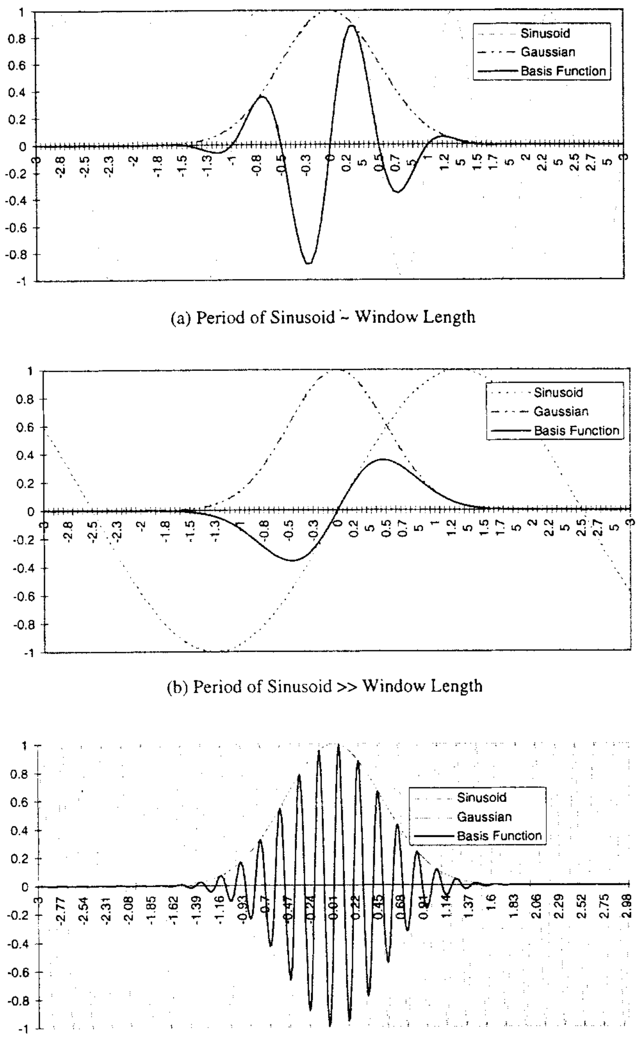
The preceding examples show how a time-localized (or frequency-localized) basis resulted in a system identification method for developing time-localized (or frequency-localized) models. However, the "ideal" methodology for process identification should be capable of generating models that are based on the input-output information that is localized both in frequency and time. Such a system identification technique will be flexible enough to handle extraneous noise and disturbance effects contained in the input-output data that may contain information in various frequency ranges and may vary over time. These requirements suggest that the projection operator, P, in Eq. 3, should be designed in such a way that its eigenrows (eigenvectors) form an orthonormal basis that achieves localization in time and in frequency of the contents of the input-output data records. This type of basis (although not orthonormal) was introduced by Gabor (1946) by modulating each sinusoid with a Gaussian of fixed size. The original Gabor transform has led to a large number of sets of such basis functions, which are collectively known as a short-term Fourier transform (STFT). The STFT of the function $f(t)$ is given by

$$G[(\omega,\tau)] = \int_{-\infty}^{\infty} g(t+\tau) f(t) e^{-2i\pi\omega\tau}\, dt,$$

where $g(t)$ is a Gaussian over a window of fixed length, and is a distinguishing feature of this family of basis sets.

The STFT is traditionally used in signal processing for the analysis of nonstationary signals and has been applied to system identification techniques as the empirical transfer function estimate (ETFE) (Ljung, 1987). Although the fixed window allows resolution in both time and frequency, it is also responsible for the limitations of these methods. First, for a given window of fixed length, the resolution of frequencies whose periods are much longer than the length of the window will be poor, whereas the finest temporal resolution possible is limited to the length of the window despite the fact that these events may occur over much shorter time scales. Second, it is not possible to form an orthogonal basis with a fixed window, which is necessary to obtain unbiased estimates of the parameters in the mathematical description of the process dynamics.

The first point is most clearly understood in geometrical terms. In Figure 1a the sinusoid is modulated with a Gaussian window with standard deviation equal to 1/2 of the sinusoid's period. The resulting basis function still retains a sinusoidal character while also having limited temporal extent, thus allowing good resolution in both time and frequency.



(a) Period of Sinusoid ~ Window Length

(b) Period of Sinusoid >> Window Length

(c) Period of Sinusoid << Window Length

Figure 1. Modulation of sinusoid with fixed (Gaussian) window.

This is not the case when the same window is used in conjunction with a sinusoid of frequency that is 1/10 of that used in Figure 1a. As shown in Figure 1b, when the window length decreases below the period of the sinusoid, the resulting basis function is deficient in any characteristics of this frequency. Thus, it becomes difficult to interpret the decomposition of the signal as the period of the sinusoid increases to lengths that are longer than the length of the window. A similar problem occurs at frequencies that have a short period corresponding to the length of the window, as is shown in Figure 1c for a sinusoid with a frequency that is 10 times the frequency of that used in Figure 1a. Clearly, the resulting basis function retains the characteristics of this sinusoid. However, it is impossible to achieve temporal resolution of any short-term event whose temporal behavior is characterized by periods smaller than the length of the window. Thus, any short-period (high-frequency) event (such as a disturbance) occurring within this window would lose its temporal characteristics contained in the length of the window, resulting in the distortion of pertinent information.

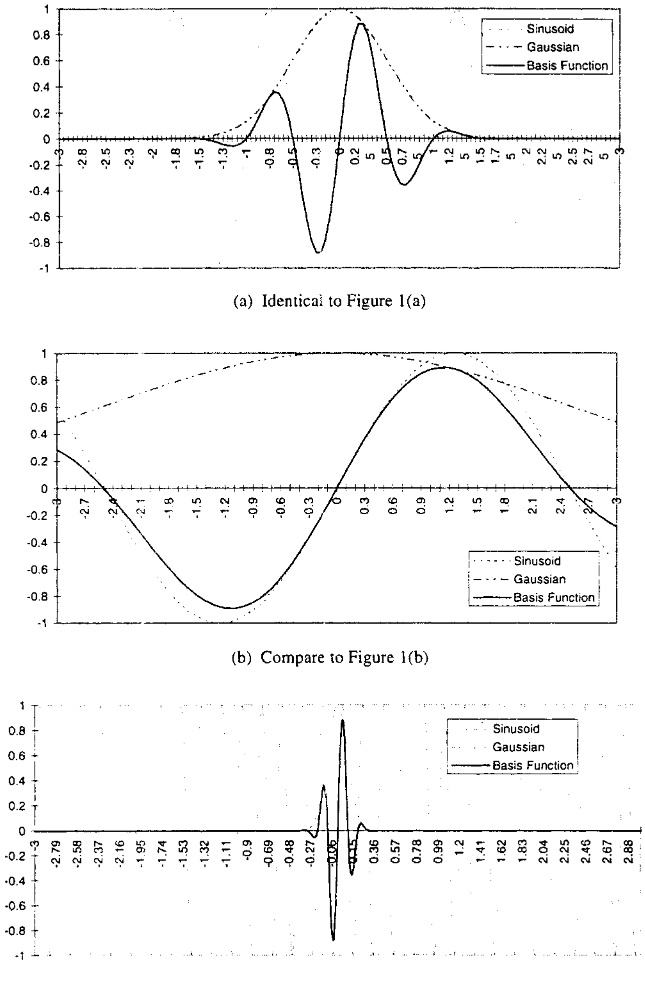
These observations of the behavior of the STFT clearly point out that the most appropriate window length cannot be selected independently of the frequency of the sinusoid. The window size must be varied as a function of frequency in order to obtain a consistent decomposition of trends over the entire time-frequency space. A linear scaling of the temporal length of the window to the period of the sinusoid or

$$\frac{T_{\text{window}}}{T_{\text{sinusoid}}} = k,$$

where

$$k = O(1)$$

is dimensionally consistent, and therefore guarantees geometric similarity of the basis functions. Assuming that the window length and sinusoidal period shown in Figure 1a are properly matched, the proper window sizes for the sinusoids of Figures 1b and c may be selected based on the scaling equation just shown. The new basis functions resulting from the variable window are shown in Figures 2a–2c. It is evident from Figure 2 (and easily verified algebraically) that the usage of a variable-length window to create the basis functions shown in Figures 2b and c is equivalent to a *dilation* (and, *contraction*) of the basis function in Figure 2a.



(a) Identical to Figure 1(a)

(b) Compare to Figure 1(b)

(c) Compare with Figure 1(c)

Figure 2. Modulation of sinusoid with variable (Gaussian) window.

The advent of *wavelets* has offered a natural framework for the specification of sets of basis functions that satisfy the preceding requirements. The dilations and translations of a single function constitute a complete set of localized basis functions capable of representing any square-integrable function. Remarkably, it has also been shown that in addition to their good localization (in time and in frequency), some wavelets generate orthogonal basis sets (Mallat, 1989). The next section introduces the wavelet transform, placing emphasis on the properties that are relevant for the decomposition of process signals and subsequent system identification.

Review of Wavelet Transform and Discussion of Relevance to System Identification

Like some of its more familiar counterparts, such as the discrete Fourier transform, the wavelet transform is based on the principle that any square-integrable function may be represented (up to a desired accuracy) as a linear combination of a specific subset of discretely sampled functions that are square integrable. This subset is referred to as a *basis*, and the functions that belong to this basis are referred to as *basis functions*. The advantage of representing a function in terms of one of these basis sets is that the desired information contained in a complex function can be detected, extracted, or manipulated more easily after decomposition onto one of these bases. For instance, if the function is an input to a linear differential equation, the decomposition of this function onto a sinusoidal basis (the Fourier and Laplace transforms) is chosen because the sinusoids are the *eigenfunctions* of the differential operator (Strang, 1986). This property, in conjunction with linearity, allows the differential equation to be solved for each basis function (sinusoid) independently of the others, finally obtaining the complete solution by summing together the individual solutions for each basis function.

The members of a wavelet-based basis are generated by dilating and translating a single function, $\Psi(t)$, known as the mother wavelet, or simply wavelet. Any square-integrable function, $\Psi(t)$, that satisfies the so-called *admissibility condition*,

$$\int_0^{\infty} \frac{|\hat{\Psi}(\omega)|^2}{\omega}\, d\omega < \infty$$

can be used as a wavelet (Daubechies, 1988). The admissibility condition guarantees that the Fourier transform of the wavelet function resembles that of a band-pass filter.

An arbitrary square-integrable function, designated by $f(t)$, may be represented as a linear combination of these dilations and translations of $\Psi(t)$, that is,

$$f(t) = \sum_{j=0}^{\infty} \sum_{k=-\infty}^{\infty} d_{jk} \Psi_{jk}(t),$$

where

$$\Psi_{jk}(t) = \frac{1}{\sqrt{2^j}} \Psi(2^{-j}t - 2^j k).$$

The index, j, is the dilation parameter, also referred to as the *level* or *scale*, and the index, k, is the translation parameter.

The effects of varying the parameters j (dilation) and k (translation) can best be understood in conjunction with the time–frequency plane, which is shown in Figure 3. Consider a wavelet at $j=1$, $k=0$. This wavelet's energy in time and frequency is concentrated in the appropriate box. As k is changed, the wavelet is shifted in time, while energy in frequency remains unchanged (recall from Fourier theory that a time shift only affects the phase). In this manner the subspace of the time–frequency plot corresponding to $j=1$ can be completely spanned. Similarly, an increase in j by one results in a dilation of the function Ψ, which doubles its extent in time. The effect of the dilation parameter j on the Fourier transform of $\Psi(t)$ can be derived from the time–frequency scaling relationship of Fourier transforms, that is,

$$F\left[f\left(\frac{t}{2^j}\right)\right]=\frac{1}{2^j}\hat{f}(2^j\omega).$$

This as j increased by one, the center frequency and the bandwidth of the wavelet are cut in half. In the time–frequency plane this results in the dimension of the box being halved along the frequency dimension while being doubled in the time direction (the area of the box is independent of j or k, and is constrained by the uncertainty principle; see Strang (1986)), as well as the center of the box being moved to half the frequency of the original box. Because the parameter j represents a range of frequencies rather than a single one, it is often referred to as a scale parameter instead of a frequency. Note that for discrete signals, the lowest scale (highest frequency range) is set by the sampling rate and we have arbitrarily set this scale to $j=0$. A new subregion of the time–frequency plane is now spanned by holding $j=1$ and again varying k. The entire time–frequency plane can be spanned in this manner, which is a consequence of the completeness of the wavelet basis.

When using the wavelet transform, it is often convenient to work with all of the wavelets at a given scale j, as in the case where all signal information corresponding to a specific frequency bandwidth is required. These wavelets may be grouped together to form a subspace of the space of all square-integrable functions, referred to as W_j. Since, by definition

$$W_j=\text{span}\{\Psi_{jk}(t),-\infty<k<\infty\},$$

the entire set of subspaces W_j span the space of all square-integrable discrete functions, that is,

$$\text{span}\{W_j;0\le j<\infty\}=l^2.$$

Similarly, it is also useful to be able to work with all of the wavelets that are at scales lower than a given scale j, as in the case where a low-frequency description of the signal is needed. These wavelets may be grouped into a subspace of l^2, referred to as V_j. An existing subspace V_j may be broken into a subspace of coarser description (higher scale) and the wavelet subspace that spans the removed portion of the original subspace, or

$$V_j=V_{j+1}\oplus W_j. \tag{7}$$

The subspace V_j is spanned by the translations of the dilation of a single function, $\Phi(t)$, which is referred to as the scaling function, or

$$V_j=\text{span}\{\Phi_{jk}(t);-\infty<k<\infty\},$$

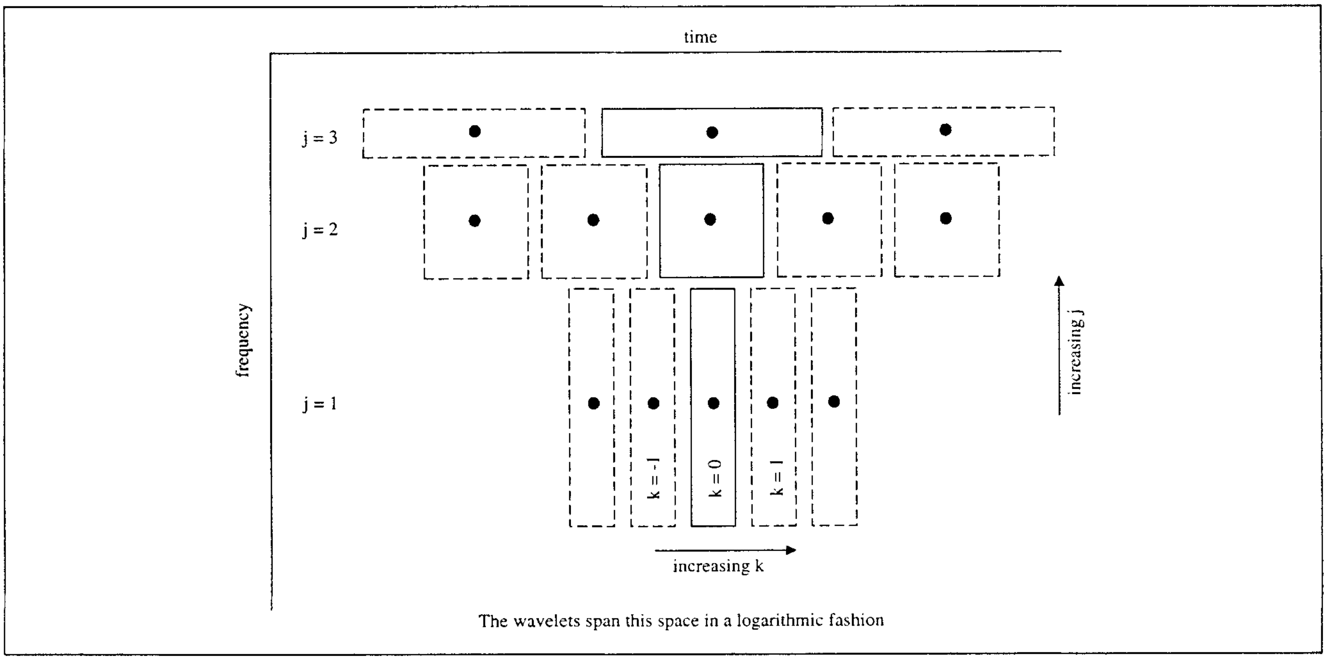


Figure 3. Spanning of time-frequency plane using wavelets.

where

$$\Phi_{jk}(t) = \frac{1}{\sqrt{2^{-j}}}\,\Phi(2^{-j}t - 2^{j}k).$$

Equation 7 implies that, in terms of the subspace V_j, the subspace V_{j+1} is the *complement* of the subspace W_j, thus establishing a relationship between the scaling function, $\Phi(t)$, and the wavelet function, $\Psi(t)$ (Mallat, 1989). Furthermore, Eq. 7 also implies that an arbitrary function, $f(t)$, which belongs to l^2, may be represented by

$$f(t) = \sum_{k=-\infty}^{\infty} c_{m+1,k}\Phi_{m+1,k}(t) + \sum_{j=0}^{m}\sum_{k=-\infty}^{\infty} d_{jk}\Psi_{jk}(t). \quad (8)$$

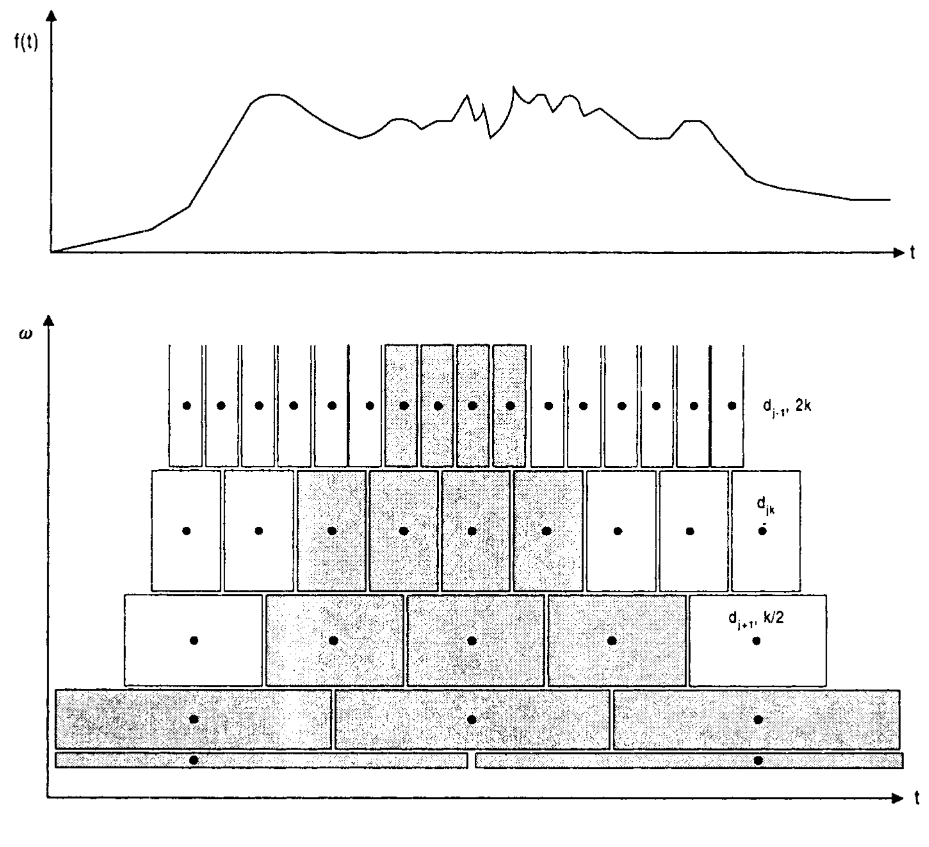
Equation 8 can be used to decompose process signals into smaller pieces that are localized in time and frequency in a highly structured manner, as is demonstrated graphically in the following example. Consider the wavelet expansion of a function $f(t)$ as shown in Figure 4. The numerical values of the coefficients, d_{jk}, have a clear physical interpretation, just as the Fourier coefficients do. A large value of the coefficient d_{jk} indicates that the content of $f(t)$ in the temporal region determined by k is rich in energy over the frequency range j. The function, $f(t)$ contains low-frequency information across its entire time span, while high-frequency events are localized around the center of its domain. This is reflected in large values of the coefficients for the corresponding time–frequency regions, as is shown by the shaded areas in Figure 4.

The preceding description is relevant to the identification of process models from industrial data, because only certain portions of these signals, corresponding to specific time–frequency regions of l^2, should be used when constructing models intended for feedback control from these records of data. A methodology for constructing models of this nature from a basis function description of the process signals is explained in the next section.

General Framework for Linear System Identification

In order to implement the identification strategy suggested by Eq. 6, using wavelets to construct the projection operators, the wavelet basis must be represented in finite matrix form. Define the wavelet matrix W as

$$W = \begin{bmatrix} V_{m+1} \\ W_m \\ \vdots \\ W_1 \end{bmatrix},$$



Shaded boxes indicate significantly large values of the corresponding wavelet coefficient

Figure 4. Wavelet representation of a function.

where each submatrix of W represents a single scaling level, and whose rows are the translates of a single wavelet (scaling function) at that level. The manner in which the selection of scale parameter determines the frequency range can be visualized on a Bode plot, as shown in Figure 5. It should be noted that the wavelet subspaces partition the Bode plot into equal sections of the frequency axis. Since the trends of a linear system are invariant (except for translation) on a Bode plot, this equipartitioning on a logarithmic scale is most suitable.

Some of the translates of the wavelet at a given level will have a domain that lies on both sides of one of the two endpoints of the data record. In this article, such "boundary effects" are handled by omitting these "straddlers." The number of basis functions (and hence, number of linear equations) at level j whose domains fall completely within the domain of the data record is given by

$$\text{No. of linear equations at level } j = \frac{z}{2^j} - \frac{l_1}{2} + 1,$$

where z is the length of the data record (equivalent to the number of rows in the A matrix of Eq. 2 and l_1 is the length of the wavelet at level 1. This number is, of course, less than the number of degrees of freedom at each level, given by $z/2^i$, and therefore implies that some of the data are not used at each point. For this application, such a loss is not critical, since the length of the data record is increasing with every sampling instant. The impact of the boundary effects can be further mitigated, if necessary, by using shorter wavelets (at the cost of poorer frequency resolution) and by restricting the number of wavelet levels used.

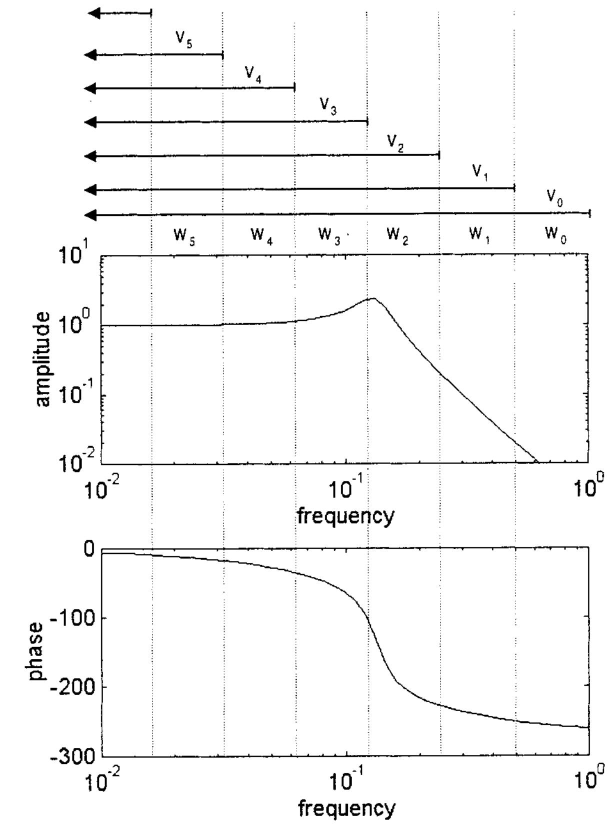


Figure 5. Approximate correspondence of bode plot to wavelet subspaces.

Determination of weighting matrix

The weighting matrix C (see Eq. 6) is considered to be a block-diagonal matrix

$$C = \text{diag}[C_{m+1} C_m \cdots C_1],$$

where each block is a diagonal matrix and contains the weighting functions for the wavelets at each scaling level. The selection of "optimal" values for the weighting constants requires knowledge of the open-loop crossover frequency and the rate at which the process response varies over the relevant frequency regime. This information resides in the input–output data set, and requires extraction. Therefore, selection of the weighting constants is generally an iterative process. The advantage specific to wavelets in this application is that the relationship between the magnitude of the coefficients and the objective to be achieved by adjustment of these coefficients is explicit, and provides direct control over the iterative model-building process.

For example, the frequency region that contains the open-loop crossover frequency may be found by setting each of the submatrices in turn to the identity, while forcing all others to zero. Each resulting model is then evaluated as to whether it contains a phase shift of $-180°$. The highest scale model that meets this criterion defines the open-loop crossover frequency. A model of order three or greater should be used to ensure that the model structure is sufficient to capture the crossover frequency.

Similarly, nonstationary noise may be easily removed from the data without impacting the overall identification process, simply by setting the weighting coefficient, corresponding to the wavelet that is projected onto the noise, to zero. This technique is especially useful in removing "spikes" from the data, which may be deduced to be extraneous effects from process fundamentals.

Finally, both frequency and temporal prediction accuracy of the model may be directly controlled by proper selection of the weights. For example, selecting a submatrix of the form

$$C_i = \text{diag}[\cdots \lambda^3 \lambda^2 \lambda\ 1]; \qquad 0 < \lambda < 1$$

for each frequency level of interest will result in a model that may be regarded as a *frequency-localized, time-varying model* with exponential forgetting.

Calculation of bounds on modeling error

Doyle and Stein (1981) suggested that the unstructured modeling error, which is a function of frequency, that is,

$$l_a(\omega) = |G(\omega) - \hat{G}(\omega)|$$

should be bounded at each frequency. ($G(\omega)$ is the transfer function of the real process, while $\hat{G}(\omega)$ is the transfer function of the assumed model.) The bounding function can be estimated from input–output data using the following relations (LaMaire et al., 1987; Kosut, 1987):

$$|G(\omega)-\hat{G}(\omega)| \leq |G(\omega)-\overline{G}(\omega)| + |\overline{G}(\omega)-\hat{G}(\omega)|, \tag{12}$$

where $\overline{G}(\omega)$ is a nonparametric (ETFE) estimate of the transfer function. The first term on the righthand side of Eq. 12 is given by (Jenkins and Watts, 1968)

$$|G(\omega)-G(\omega)|^2 \leq \frac{2}{\upsilon-2} F_{2,\upsilon-2}(\alpha)\frac{\phi^s yy(\omega)}{\phi^s uu(\omega)},$$

where $F_{2,\upsilon-2}(\alpha)$ is the $(1-\alpha)$ Fisher statistic form 2 and $n-2$ degrees of freedom, υ is the number of degrees associated with the spectral window, and $\phi yy(\omega), \phi uu(\omega)$ are the autospectra of y and u, respectively. The use of wavelets to clean the temporal input–output process signals from noise and unscheduled disturbances, provides more reliable estimates of the autospectra at various frequency ranges (see Example 2, below).

Comparison of the Wavelet-Based Process Identification Method Against Standard Techniques Using LTI and LTV Processes

The previous sections introduced the wavelet transform, emphasized its time–frequency localization characteristics, and indicated that wavelets could form a very attractive set of modulating functions for process identification. The examples presented in this section have been selected to demonstrate how these properties can be used to develop reduced-order models tailored for use in a control framework by taking into account, among other factors, such realities as noise, disturbances, and time-varying process characteristics.

Example 1: Identification of Reduced-Order Models for LTI Systems Using the Wavelet Decomposition. Consider the common system identification problem of identifying a reduced-order model from data. The actual underlying system has a transfer given by

$$g(z) = 10^{-5}\frac{0.0052z^4+0.1316z^3+0.3307z^2+0.1302z+0.005}{z^5-4.72z^4+9.14z^3-9.08z^2+4.63z+0.969},$$

which has 4 zeros and 5 poles located at

zeros: $[-22.7, -2.29, -0.431, -0.0435]$

poles: $[0.869+0.475i, 0.869-0.475i, 0.998+0.05i, 0.998-0.05i, 0.990]$.

This system has two pairs of lightly damped poles, which is evident from its Bode plot in Figure 6a. There is a pair in the region of the crossover frequency, such that it is necessary that the dynamics of this pair is accurately modeled. The data are generated using a PRBS signal $[-1,1]$ of length 4096. The data are fitted to a reduced-order model with two zeros and three poles.

Least-Squares Parameter Estimation. The reduced-order model corresponding to a least-squares fit is

$$g_{LS}(z) = 10^{-5}\frac{0.0733z^2+0.212z+0.627}{z^3-2.95z^2+2.91z-0.955}$$

with

zeros: $[-1.45+2.54i, -1.45-2.54i]$

poles: $[0.994+0.0497i, 0.994-0.0497i, 0.963]$.

Parameter Estimation Through Modulation with Wavelets. The parameters for the reduced-order model were estimated using the wavelets at the scale $j=6$, which covers the frequency region that includes the open-loop crossover frequency. The reduced-order model is found to be

$$g_w(s) = 10^{-4}\frac{0.243z^2-0.325z+0.339}{z^3-2.98z^2+2.97z-0.986}$$

with zeros and poles at the following locations:

zeros: $[0.669+0.975i, 0.669-0.975i]$

poles: $[0.998+0.0499i, 0.998-0.0499i, 0.988]$.

Least-Squares Parameter Estimation with Band-Pass Filtering of Data. For comparison, parameter estimation was done by first filtering the data with a fourth-order Butterworth filter with band-pass in the frequency range [0.01, 0.1] Hz, before performing a least-squares estimation. The reduced-order model for this case is

$$g_{bf}(s) = 10^{-3}\frac{-0.160z^2+0.323z-0.139}{z^3-2.975z^2+2.95z-0.977}$$

whose zeros and poles are at the following locations:

zeros: $[1.40, 0.618]$

poles: $[0.998+0.0498i, 0.998-0.0478i, 0.979]$.

To compare the results of the three methods, let us examine the Bode plots of the resulting dynamic systems shown in Figure 6a. Note that the least-squares model attempts to provide a balanced estimation over the whole frequency range. This is not an optimal model, since accurate modeling is not required at frequencies higher than the desired closed-loop bandwidth. This can be rectified by using wavelets as modulating functions. The band-pass nature of the wavelets mitigates the effects of irrelevant low and high frequencies. For this process, the open-loop bandwidth of the process corresponds to the wavelets at scale level $j=6$. The wavelet model thus gives a much better fit of the process at the relevant frequencies as compared to least squares, which can be seen from the lower bounds on the additive modeling error, as shown in Figure 6b. It can be seen that the error bounds near

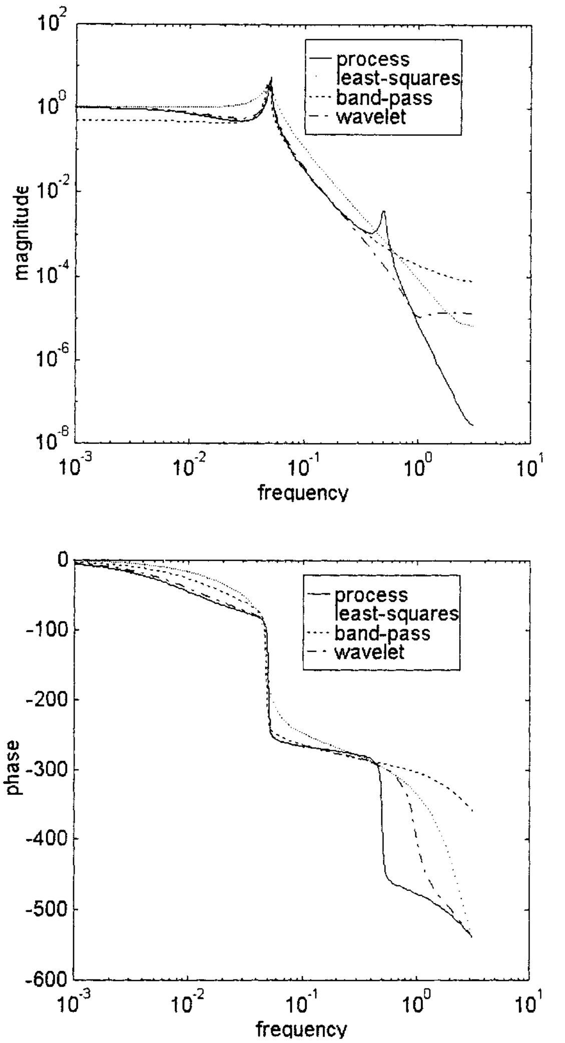


Figure 6a. Reduced-order models of LTI process (Example 1).

the crossover frequency are approximately an order of magnitude greater for the least-squares estimate than for the wavelet estimate, which implies that the robust control loop designed using the wavelet model will have a higher closed-loop bandwidth than that of a robust controller designed using the least-squares model. Not surprisingly, the band-pass filter performs similarly to the wavelet transform. However, due to the fact that it cannot provide simultaneous localization in both time and frequency domains, it is not expected to fare as well in the identification of systems with time-varying or nonlinear characteristics (see Example 3, below).

Example 2: Effects of Low-Frequency Disturbances and High-Frequency Noise on Model Identification. The presence of noise and disturbances in the operating data can lead to poor models for control if the data are not properly treated. The wavelet transform is an excellent framework for performing reduced-order model identification under these conditions. Consider the process from Example 1 corrupted by a low-frequency ramp disturbance and Gaussian white noise, or

$$y(z) = g(z)u(z) + d(z) + n(z)$$

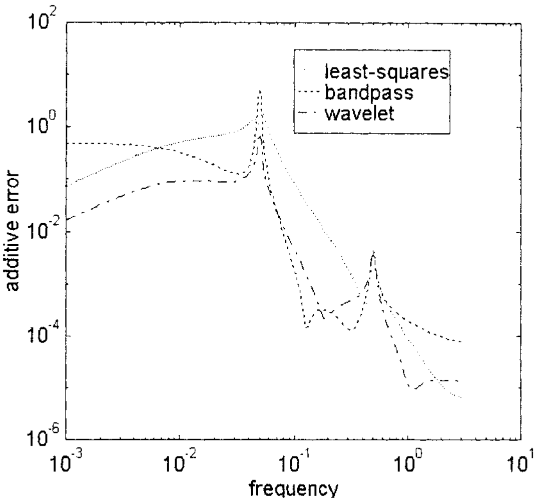


Figure 6b. Modeling error associated with different methods.

where

$$d(t) = \frac{t}{N}, \quad t = 1, \dots, N$$

and

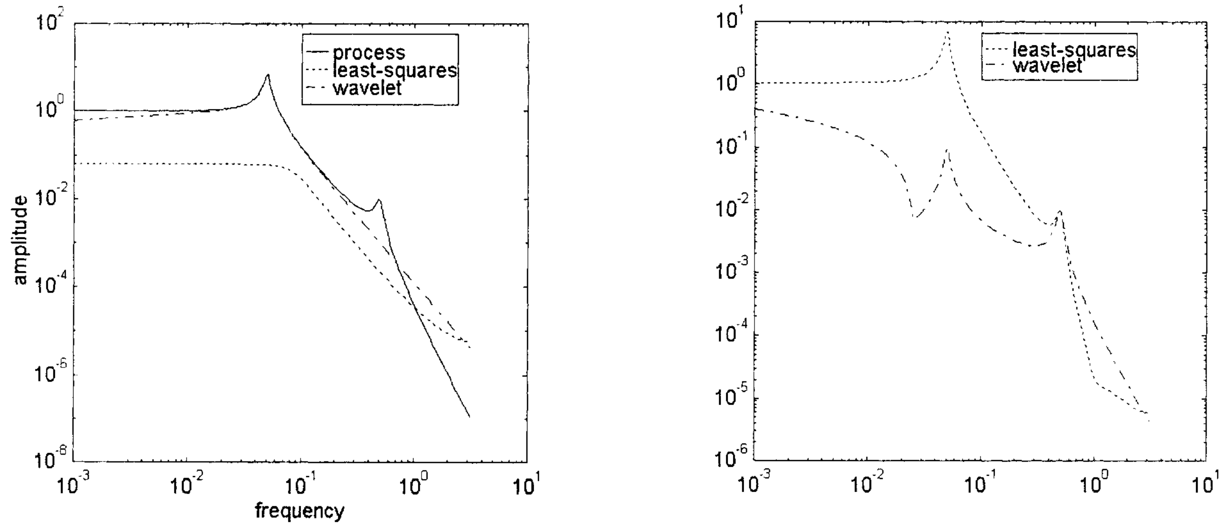
$$E[n(t)] = 0; \qquad E[n(t)n(\tau)] = 10^{-3}\delta(t-\tau)$$

The process data are corrupted by a low-frequency disturbance and have a low signal-to-noise ratio (SNR) at high frequencies due to noise. Therefore, these frequencies contain little information about process behavior and should not be used for model identification. This is easily accomplished using wavelets as modulating functions in the same way as in Example 1 and by omitting the contribution of the wavelets at the scale where the noise is concentrated. Regarding the separation of the disturbance effects, it is important to note that the crossover frequency is at a different frequency region from that where the disturbance contains significant levels of energy and the gain from input to output is still high with respect to the magnitude of the disturbance. (Clearly, theoretical restrictions make it impossible to separate out the effects of the disturbance, if the frequency region of the disturbance's main power coincides with the region containing the crossover frequency, whatever filtering approach one were to use.) The third-order model identified by the wavelet transform at the scale corresponding to the crossover frequency is given by

$$g_w(s) = 10^{-4}\frac{1.278z^2 - 2.754z + 1.737}{z^3 - 2.975z^2 + 2.953z - 0.978},$$

and at the region of the crossover frequency provides a more accurate depiction of process dynamics than the third-order model generated from the corresponding least-squares approach and that is given by

$$g_{LS}(s) = 10^{-5}\frac{-0.8935z^2 + 3.603z - 0.1653}{z^3 - 2.985z^2 + 2.972z - 0.9874}.$$



a) Comparison of Methods

(b) Frequency Localization Prevents Corruption of Model Performance Noise and Distubances at Irrelevant Frequencies

Figure 7. Reduced-order models of LTI process with noise and disturbances (Example 2).

The Bode plots for the models identified by least-squares and wavelet-based modulations of the input and output data are shown in Figure 7a, while the corresponding modeling errors are depicted in Figure 7b.

Example 3: Identification of Reduced-Order Models for LTV Systems Using the Wavelet Decomposition. Example 1 is extended to the more general case where the actual process is of unknown order and changes with time. Initially, there is a pair of lightly damped poles in the region of the crossover frequency. As time progresses, these poles become more heavily damped until the end of the time record, at which point the poles are critically damped. The location of these poles is in the region of the crossover frequency. It is therefore important that these poles be accurately tracked as they change with time, in order to maintain a stable feedback loop with good performance characteristics. The initial process is given by

$$g_3(z) = 10^{-3}\frac{0.1284z^4+2.712z^3+6.094z^2+2.422z+0.1026}{z^5-2.427z^4+2.798z^3+2.637z^2+1.993z-0.7154},$$

and the Bode plots of the process at the beginning and the end of the data record are shown in Figure 8a. The data record was generated using a PRBS signal of length 4096 as the input. Various techniques were used to fit the data to a reduced-order model with three poles and two zeros. Figure 8b displays the additive error between the actual process and the models, identified by the various techniques at the process crossover frequency as a function of time.

Not surprisingly, the least-squares method gives the least accurate result, as the method incorporates all the data including that which is outdated and at irrelevant frequencies. The Kalman filter with an exponential forgetting factor $\lambda = 0.97$ gives similar results to the least-squares because too much modeling effort is spent on the high-frequency underdamped pair of poles. Prefiltering the data with a fourth-order Butterworth filter with band-pass [0.01, 0.1] provides for no time localization and thus gives an average of the frequency response of the process in the band-pass region. As the process moves from lightly damped to critically damped, it passes through this average, which explains the minimum encountered in the additive error plot. The wavelet identification technique using the ten most recent wavelets as modulating functions is able to capture the process behavior with an error that is approximately an order of magnitude, or more, lower than the other techniques.

It should be noted that a numerically equivalent solution can be generated through an appropriate combination of a least-squares regressor with properly selected noise filters and forgetting factors. However, the wavelet-based approach provides a unifying framework for the specification of such a "cocktail," which otherwise would have remained an ad hoc procedure.

Modeling of Nonlinear Processes Using the Wavelet Transform

For many industrial problems, linear modeling techniques are insufficient because the process displays significant nonlinear behavior over the desired operating region. These problems have been the impetus for a rich variety of tools and techniques that could be considered to fall under the loose title of nonlinear identification and control. Two of the more important of these techniques are the following.

- Gain scheduling
- Feedback linearization and its adaptations for robust control.

These techniques require a nonlinear model of the process. Nonetheless, a nonlinear model can be represented as a set of linear models, each corresponding to a different operating region (Banerjee and Pearson, 1995).

In this section we introduce a system identification technique based on the wavelet transform that is capable of modeling nonlinear systems yet accomplishes this goal while

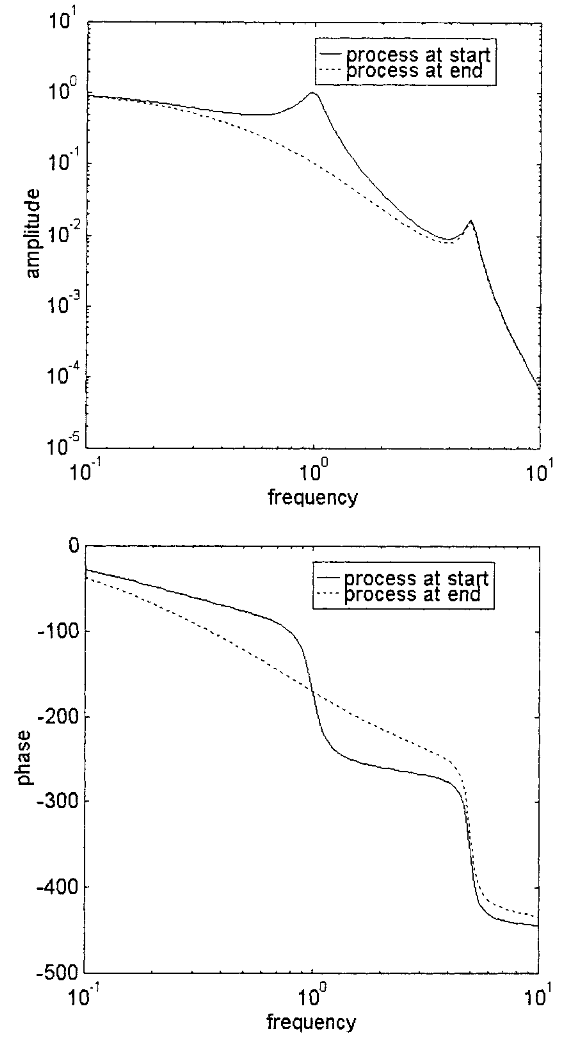


Figure 8a. Comparison of models constructed for the control of an LTV process.

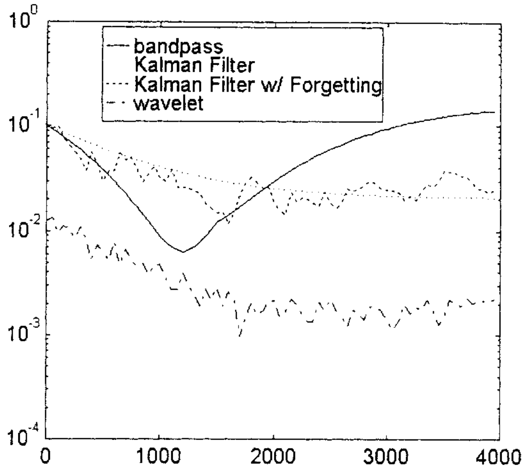


Figure 8b. Time and frequency localization required for construction of accurate control model.

maintaining a reasonable trade-off between model structure complexity/size and effectiveness. The underlying concept arises from the practical application of linear system theory to real processes. Since these processes are in general nonlinear, the linear system theory is invalid. However, if the process does not show significant nonlinearities over the *desired* range of operation, a linear feedback controller may prove to be quite adequate.

This idea can be derived from taking a Taylor series of a nonlinear, continuous system and noting that the system will behave linearly in some finite neighborhood surrounding the point of linearization. If this neighborhood contains the desired operating region, a linear model should be suitable for feedback control, as stated earlier. If it does not, the desired operating region is divided into a sufficient number of subregions such that the process behavior is approximately linear in each one of them.

The wavelet provides an excellent basis for a modeling task of this nature, because of its orthogonality and time–frequency localization properties. The key to identifying nonlinear systems is noting that the *state* level (which may be used as the point of linearization for the process model) is correlated with *time*. For example, as pointed out in the fourth section, the wavelet coefficient d_{jk} contains information about the signal at a frequency region determined by j, and a temporal extent specified by k. During this time period, the range of the function is bounded as shown in Figure 9. The wavelet coefficient d_{jk} contains information about the process response that is specific to (1) the particular operating region, and (2) the main frequency range(s) contained in the signal. A linear model that reflects the process behavior only over a limited operating region and bandwidth relevant to feedback control can then be constructed by selecting values of j and k, which correspond to an output level in both the operating region and bandwidth, respectively.

Modeling scheme for nonlinear processes

This procedure is based on the aforementioned concept that a continuous nonlinear process may be adequately modeled as linear over a particular operating region. In this section, the procedure of deciding whether a process is behaving in a linear fashion over a given finite operating region is for-

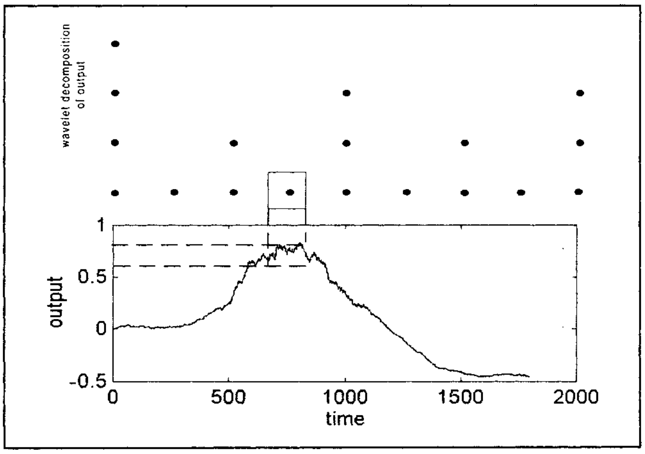


Figure 9. Decomposition of the process signal into segments localized in frequency and state level using the wavelet transform.

malized, and methods for dividing up a given operating region into subregions, over which the process behaves in a more linear fashion are suggested.

It is necessary to develop a criterion to assess whether the process is behaving in a nonlinear fashion over a given operating region from input–output data. Given a set of input–output data, it is postulated that the process can be adequately represented as being linear over the operating region of the data. Fitting these data to a linear model is not sufficient to test for nonlinearities because the model prediction error is a result of not only nonlinearities, but also of reduced-order modeling, disturbances, and noise. The differentiating characteristic of errors due to nonlinearities from the others is that the nonlinear effects are correlated with state level. This property can be exploited to detect and model nonlinear effects from the input–output data alone. Note that, as in the linear case, the model order must first be estimated using a technique such as cross-validation (Stone, 1974; Snee, 1977).

Due to its independence with respect to state level, all estimates of a linear process will converge to the same model, regardless of how the data are subdivided. This is not true for nonlinear systems. Thus, the hypothesis that the system is behaving linearly over the range of input–output data may be tested by splitting the data record with respect to output level and by comparing the average prediction error of the new models to that of the original. For models constructed through modulation with wavelet functions, the average model prediction error (which is calculated over the modeling region of interest) is simply the average of the minimum filtered error from the prediction of the modeling parameters.

The error bound for each of these linear models is compared with the error bound for a single linear model. If the process behaves linearly, there will be no statistically significant difference between any of these estimates (the size of the error bounds is then a function of noise and disturbances); a decrease in the error bounds of the two linear models with respect to the error using one model implies that the system is behaving nonlinearly over the operating region. For the latter case, the two linear models are accepted as a better model of the process than the original single linear model, which is discarded.

The algorithm is then repeated in each of the subregions until one of the following conditions is reached:

1) The error rate remains constant as the number of models is increased (this error rate reflects the amount of noise and disturbance in the signal).

2) There is insufficient data for further iterations.

The preceding operation requires a method for dividing a data region into two subregions. As this split is arbitrary, it can be done in several ways, as discussed in the following paragraphs.

Mathematically rigorous approach

The mathematically rigorous method is based on minimizing the prediction error over the operating region given two linear models with undetermined coefficients, that is

$$\min_{A,B}\left[\sum_{i}(y_i^A-\hat{y}_i^A)^2+\sum_{j}(y_i^B-\hat{y}_i^B)^2\right].$$

This formulation may result in a nonlinear optimization with local minima depending on the inherent process nonlinearities. It may be more practical to use other "suboptimal" schemes such as the following:

1) Split the operation regime in half; the major advantage of this method is its simplicity.

2) Split the operating regions into two subregions that have the same number of data points; this approach is also simple and maintains data sufficiency.

3) Make the split using *a priori* knowledge from either fundamental principles or previous operating experience.

The first of these three practical approaches is used in this article for the illustrative examples.

Example 4: Modeling of a Nonlinear Exothermic CSTR Using the Wavelet Transform. A first-order irreversible reaction in a constant-volume, nonadiabatic CSTR can be described by the following dimensionless nonlinear equations (Hoo and Kantor, 1985):

$$\frac{dx_1}{d\tau}=-x_1+Da(1-x_1)\exp\left(\frac{x^2}{1+\frac{x_2}{\gamma}}\right) \tag{13}$$

$$\frac{dx_2}{d\tau}=-x_2+BDa(1-x_1)\exp\left(\frac{x_2}{1+\frac{x_2}{\gamma}}\right)-\beta(x_2-x_c), \tag{14}$$

where

$$B=\frac{(-\Delta H)C_{Ai}\gamma}{\rho C_pT_o}\qquad \beta=\frac{UA_c}{\rho C_pQ_o}\qquad \gamma=\frac{E_o}{RT_{fo}}$$

$$Da=\frac{V}{Q_o}k_oC_{Ai}^2\exp(-\gamma)$$

with the corresponding dimensionless variables given by

$$\tau=\frac{Q_o}{V}t\qquad x_1=\frac{C_{Ai}-C_A}{C_{Ai}}\qquad x_2=\frac{T-T_i}{T_i}\qquad x_c=\frac{T_c-T_f}{T_f}$$

The term x_c is the dimensionless temperature of the cooling jacket and is the manipulated variable.

The control input is the cooling water temperature, and the measured output is the outlet concentration of reactant. It consists of a low-frequency "pulse" that is of sufficient magnitude to excite the nonlinear dynamics of the system plus a PRBS signal of magnitude ± 0.05 to ensure persistent excitation. The output data are correlated to the input data with second-order ARMAX model at scale level $j=5$, which has a filtered prediction error of 1.17×10^{-6}.

The wavelet coefficients at $j=5$ are then split into two groups with respect to the average operating level, which is depicted in Figure 10. Note that disjoint regions of data may be combined to construct a single linear model, as is the case with the first model in this set, indicated as Model-1 on Figure 10. The resulting linear models have error bounds which are 1.06×10^{-6} and 0.957×10^{-7} for the upper and lower op-

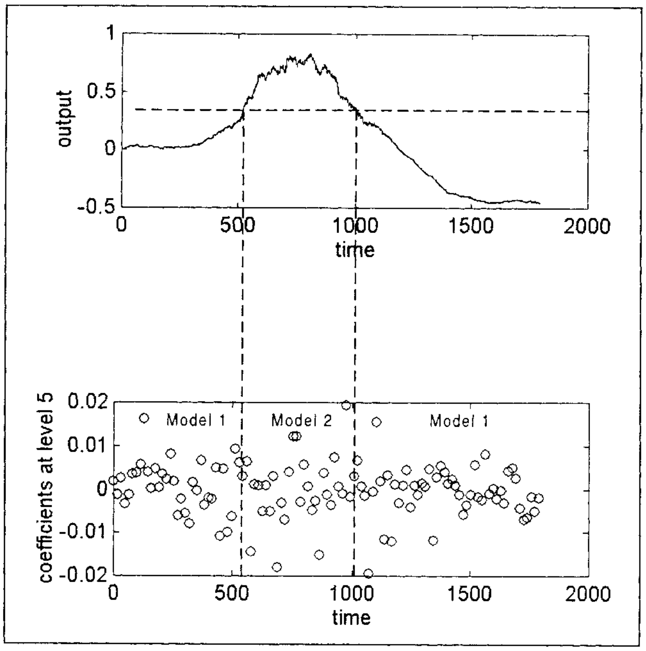


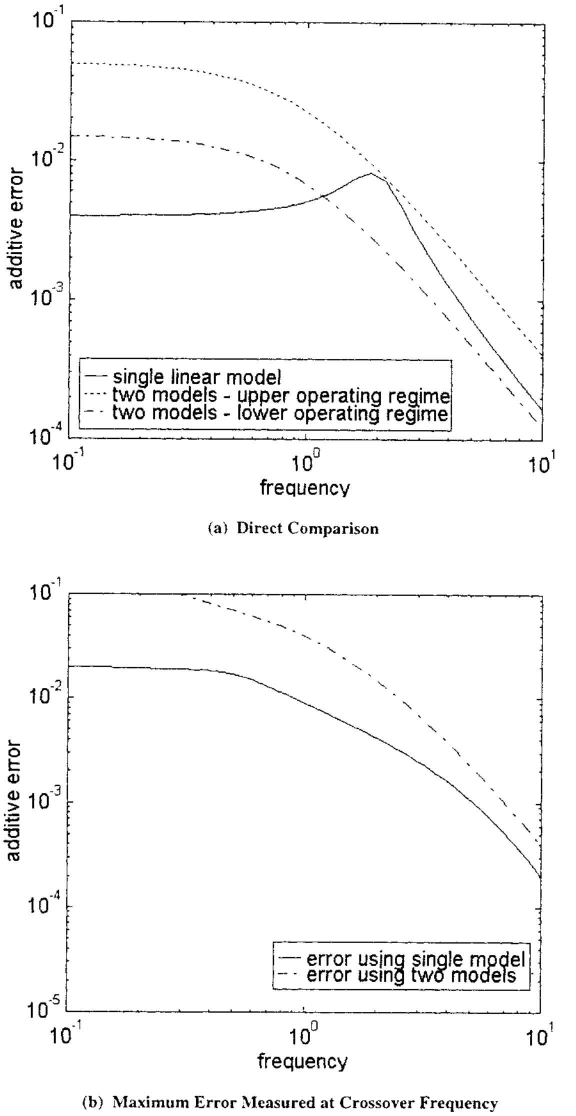
Figure 10. Construction of linear model set using the wavelet transform to represent a nonlinear process.

erating region, respectively, which shows some improvement over the original single model. For comparison, the Bode plots for all three models are shown in Figure 11a, along with the maximum and minimum frequency response for all linear approximations of Eqs. 13 and 14 over the operating region, $x_1 \in [0.057, 0.101]$. The residual errors, displayed in Figure 11b clearly demonstrate that for this case, "two models are better than one." This procedure may be repeated on either of the two models if, for performance reasons, a lower residual error is required, provided that there is a sufficient excitation in the data subrecord that is to be further subdivided.

Example 5: Modeling the Dynamics of an Industrial Distillation Column. Consider the distillation column shown in Figure 12. The feed to the column is a wild stream consisting of essentially methane, ethylene, and propylene, and it is desired to separate the methane from the heavier components. The methane can be drawn off as a gas or dissolved in a condensed phase from the condenser (the intermediate pressure methane stream). The temperature at an intermediate plate is maintained by manipulating the stream rate to the reboiler, and the pressure at the top of the column is maintained by manipulating the off-gas flow rate.

It is desired to keep the ethylene in the methane vapor stream (the *slip*) below a specified level, as well as to control the level of the accumulator tank. The available manipulated variables are the reflux flow rate and the speed of the compressor used in the refrigeration-based condenser. Although this is a multivariable system, for the purposes of this illustration we will concern ourselves with the control of the ethylene in the slip by changing the reflux flow rate. For more detailed studies on the multivariable system, see Carrier (1995).

Due to its compact extent in time, the Haar family of wavelet/scaling functions (Strang, 1989) has been selected for use as the modulating function. A large number of real-time operating data were available, spanning the operation of the column over a period of six months, but most of the planned experiments were corrupted by the influence of unpredicted external disturbances. Consequently, the first task was to analyze all the available records of data and identify those segments, which yielded "clean" input–output data for parameter identification. The wavelet decomposition of the input–output data led to the identification of several records where there is a step in reflux while the compressor speed (disturbance) remains constant. This provided an excellent opportunity to assess the effect of reflux on the percentage of ethylene in the slip independently of compressor speed. The five records corresponding to these conditions are shown in Figure 13. Note that each of these records corresponds to a different operating region based on output level, which we shall maintain as separate entities based on the expectation that the process behaves in a nonlinear manner over this operating region.



(a) Direct Comparison

(b) Maximum Error Measured at Crossover Frequency

Figure 11. Effect of model set size on approximation of a nonlinear process.

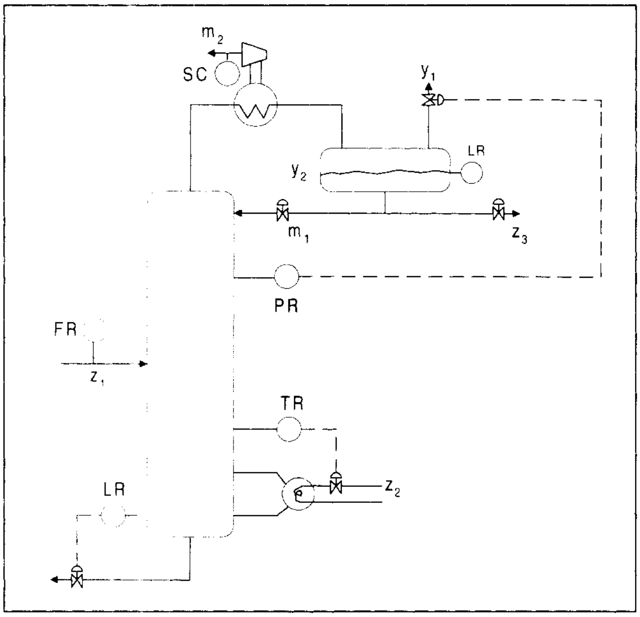


Figure 12. Industrial distillation column used in Example 5.

Manipulated Variables
m_1 = reflux rate
m_2 = compressor speed

Controlled Variables
y_1 = ethylene in slip
y_2 = reflux drum level

Measured Variables
z_1 = total feed
z_2 = steam rate to reboiler
z_3 = liquid ethylene flow rate from reflux drum

Thus, in light of the previous physical insights, we immediately look for varying performance with respect to output level. At level $j = 1$, 35 wavelet and scaling function coefficients can be calculated, and this number falls by a factor of 2 for each increase in scaling level. Thus, for instance, at level 4 a fourth-order linear model can be constructed (although a greater number of wavelet coefficients than unknown parameters is required if noise is present). For the data corresponding to the selected sets, the scaling functions at level $j = 2$ were identified as being the most appropriate set of basis functions for the identification. The selection of model order was then determined using standard techniques (Ljung, 1987); it was found that a second-order ARMAX model was suitable in all cases. The step responses for each of these models is displayed in Figure 14. As expected, the process gain is negative and decreases with respect to increasing output level. It is also evident from Figure 14 that the dominant time constant of the process increases with increasing output level.

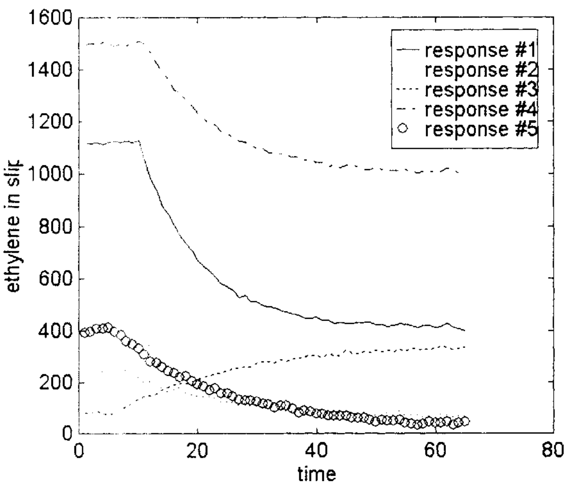


Figure 13. Response of ethylene in slip to five distinct experiments in step changes of the reflux rate.

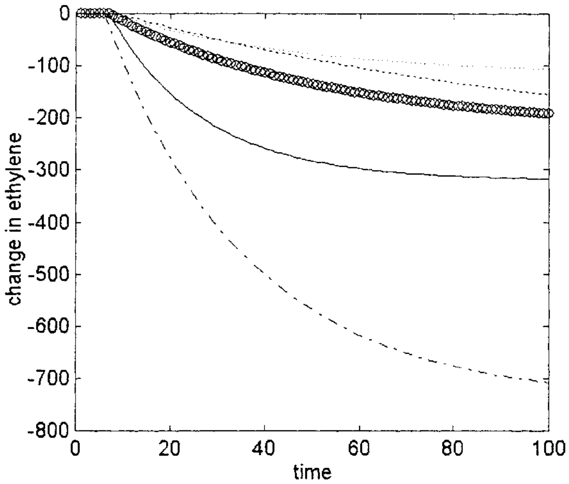


Figure 14. First-order-models constructed from the wavelet-decomposed input output data for the column of Example 5.

Similar studies were carried out for the transfer functions between ethylene slip and compressor speed, as well as the transfer functions between accumulator level (output) and the two inputs (reflux, compressor speed). The details of these studies, which do not add any new aspects for the purposes of this article and thus were omitted, can be found in Carrier (1995).

Identification of Multivariable Systems

The preceding wavelet identification technique for SISO systems using the modulating functions method can be extended to MIMO system identification, as was done in Co and Ydstie (1990). As is typical with MIMO systems, there are added complexities in comparison with the SISO system, such as minimal representation, multiple time scales of interest, and multirate sampling. However, stability and robustness of a MIMO feedback loop is still dependent on the frequency response of the system, which is the basis for such stability and robustness analysis techniques as the MIMO Nyquist criterion, principle gains (Maciejowski, 1989), and μ-analysis (Doyle, 1982). Therefore, analogous to SISO identification for control, frequency localization is an imperative, as is time localization. The wavelet transform plays the same critical role in the identification of MIMO systems for feedback control as in the SISO case.

Before a modulation function method for multivariable systems using the wavelet transform is derived, the topics of multirate sampling and minimal realization must be addressed.

Multirate processes

The construction of models from data sampled at different rates is an important industrial problem, and is receiving considerable attention in the process-control literature (Gopinath and Bequette, 1991; Ohshima and Hashimoto, 1992). The topic of multirate processes naturally arises in the multivariable context, because different outputs often respond over different time scales. These time scales determine the recommended sampling rates. It may also be neither possible nor desirable for the individual sensors to sample at identical rates because this may lead to oversampling on outputs that operate over larger time scales. This oversampling increases computational load and can push the zeros of the system outside of the unit circle (Åström and Wittenmark, 1984). Thus, it is desirable to devise a method that can accept multirate input–output data and transform them into a consistent set of information in a form proper for control-relevant identification. Due to its hierarchical structure of increasing time scales (and the corresponding frequency interpretation), the wavelet transform provides an elegant solution to the multirate sampling problem.

The sampling rate of a signal determines the points at which the wavelets and scaling functions are defined. Signals that are sampled at different rates belong to different discrete functional spaces. Thus, the basis set to be projected against both the input and output, as shown in Eq. 10a, is not uniquely defined. This drawback can be remedied by selecting a single wavelet-based functional space, which is large enough to contain both the input and the output. Consider a SISO system where the input u is sampled at a rate t_u and the output rate is sampled at a rate t_y. Let

$$t_o = \gcd\{[t_u, t_y]\},$$

where gcd is the *greatest common denominator*. In this way, the sampled points of both the input and output exist only at multiple values of t_o.

Let V_o be the space spanned by the translations of the scaling function $\Phi_o(t)$, that is,

$$V_o = \text{span}\left(\sum_k \Phi_0(t - kt_o)\right).$$

Any signal sampled at a rate that is an integer multiple of t_o may then be projected into a subspace of V_o. This process is represented graphically in Figure 15. This projection is uniquely defined for all scales greater than or equal to $j_{\min}$, where

$$j_{\min} = \sup\left(\log_2\left(\frac{t_k}{t_o}\right)\right), \tag{15}$$

where t_k is the sampling period of the signal. The approximation of the signal that is contained in the space V_j, $j \geq j_{\min}$ can be constructed as follows. The resolution of a signal $y(t)$ at scale j can then be written as

$$F_j[y(t)] = \sum_k a_{jk}\Phi_{jk}(t).$$

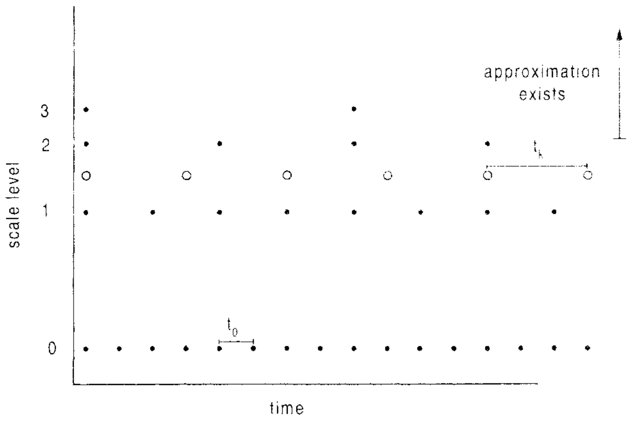


Figure 15. Generation of the binary tree of nodes for the consistent merging of input-output information in multirate systems.

The coefficients a_{jk} can be found by minimizing the sum of the prediction errors at the sampling points

$$\min \sum_{t_k} [y(t) - F_j(y(t))]^2,$$

which leads to the following set of linear equations

$$\sum_{t_k} y(t_k)\Phi_{mk}(t_k) = \sum_{\ell} a_{k\ell} \sum_{k} \Phi_{mk}(t_k)\Phi_{m\ell}(t_\ell),$$

which can be represented in matrix form as

$$\underline{y} = A\underline{x}$$

The finite temporal extent of the wavelets gives A a banded structure; A is also symmetric, because the scaling function, $\Phi(t)$, is symmetric. For the special case where

$$j = \log_2\left(\frac{t_k}{t_o}\right),$$

the off diagonal terms for an orthogonal wavelet basis are equal to zero, or

$$\sum_k \Phi_{jk}(t_k)\Phi_{j\ell}(t_\ell) = 0; \qquad j \neq 1.$$

In this case, A is reduced to a diagonal matrix.

It is now straightforward to estimate the parameters of a linear model using the methodology presented in the fourth section, with the stipulation that any wavelets or scaling functions selected as modulating functions must belong to scales that are greater than or equal to the minimum scales for both the input and output, as defined by Eq. 15.

Aliasing and the selection of sampling rate

Although the preceding method is suitable for any set of sampling rates, it is necessary that the signals are sampled at a sufficiently high rate that extends over the frequency bandwidth of the signal. Consider a signal sampled at evenly spaced intervals of t_o. This signal lies in the subspace V_o, whereas the actual underlying continuous signal may contain elements in the wavelet subspaces W_o, W_{-1}, W_{-2}, ..., and so on. The projection of the sampled data onto the subspace V_o causes these elements at higher resolutions to be misidentified as belonging to subspaces of lower resolution, thus resulting in aliasing. Since the scaling function acts as a low-pass filter, it is necessary to sample at a rate that is at least twice the bandwidth of the scaling function to prevent aliasing, or to prefilter the underlying signal with a proper aliasing filter before sampling (Oppenheim and Schafer, 1989).

Example 6: Estimation of Transfer Function from Multirate Data. Consider the system whose dynamics are shown in Figure 16. The input is sampled at a rate of 3 s, and the output is sampled at a rate of 5 s. The gcd of $\{3,5\} = 1$, which defines the distance between samples of the scaling function at level $j = 0$ to be $\frac{1}{2}$ s. The supremum of $\log_2(3)$ and $\log_2(5)$ is 2 and 3, respectively. Therefore, the input and output are each approximated by scaling functions at level 3. After the scaling coefficients at level 3 are determined, the signals are reconstructed at time intervals of 1 s, where the scaling functions are used for interpolation. These signals are then used in conjunction with the methodology discussed in the fourth section to construct a transfer function model with 2 zeros and 3 poles at level $j = 6$, which corresponds to the crossover frequency. The amplitude plot of the generated reduced-order model is also shown in Figure 16.

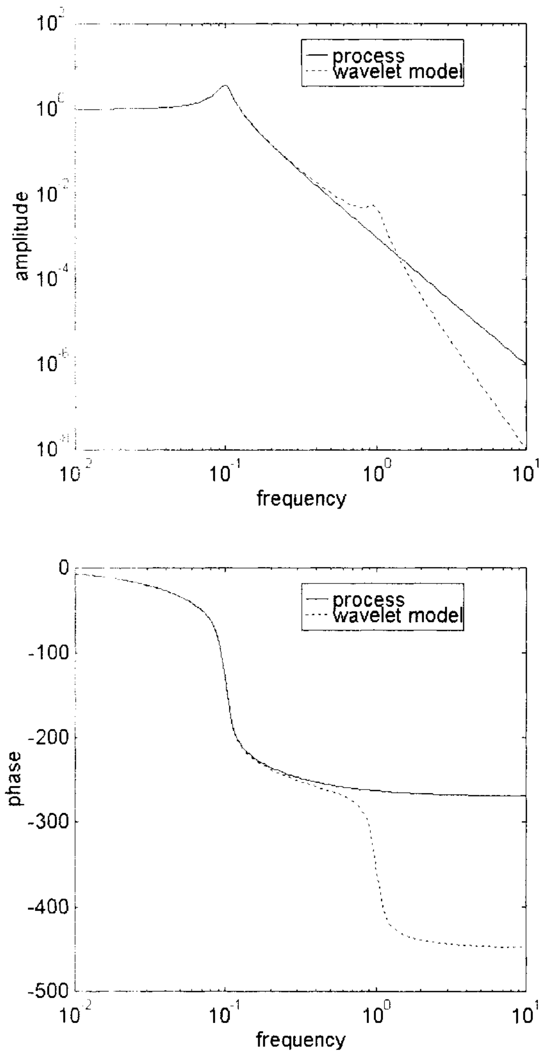


Figure 16. Bode plots of the underlying dynamics for the system of Example 6.

Wavelet-modulated parameter estimation for MIMO systems

Once a state-space or transfer-function matrix structure has been specified, the evaluation of the model parameters for a MIMO system is a simple extension of the SISO case. The major difficulty inherent in MIMO system identification is that achieving a realization that is minimal (or equivalently, identifying pole-zero cancellation) is not a trivial exercise. The added degrees of freedom result in a poorer fit as well as an increased computational burden for the same amount of data.

In this section, we review the standard methods that have been developed for the identification of linear multivariable systems from input-output data. All of these methods represent the MIMO system as a set of finite difference equations that can be used to estimate the model parameters by the pseudoinverse. It is then straightforward to apply the method of modulating functions to this set of linear equations.

Finally, it is shown that the method that results in a minimal representation (and hence the smallest number of parameters for the same input-output representation) is based on the assumption that all data are sampled at the same rate. Accordingly, it is necessary to use a nonminimal realization for systems with multisampling/multiple time scales, which decouples the identification output-by-output.

The emphasis of the paper by Gautier and Landau (1978) was to develop a MIMO system identification technique that would be an extension of SISO techniques. A linear multivariable system may be represented in the form

$$D(z)\underline{y}(z) = N(z)\underline{u}(z),$$

which is known as the matrix fraction decomposition (MFD; Kailath, 1980). $D(z)$ is a polynomial $r \times r$ matrix and $N(z)$ is a $r \times q$ polynomial matrix, where q is the number of inputs, and r is the number of outputs. The MFD is related to the transfer-function representation via

$$D^{-1}(z)N(z) = H(z),$$

where

$$\underline{y}(z) = H(z)\underline{u}(z).$$

Note that the matrices $D(z)$ and $N(z)$ are not unique; for example, an equivalent set is

$$H(z) = [N(z)X(z)][D(z)X(z)]^{-1} = \tilde{N}(z)\tilde{D}(z)^{-1}, \quad (16)$$

where $X(z)$ is any invertible matrix.

The particular representation of a multivariable linear system, as given in Eq. 16, is relevant for identification in that it is equivalent to a set of finite difference equations, which can be used to estimate the model parameters in an analogous manner to SISO systems.

Several different methods based on Eq. 16 have been suggested. Each of these methods differs in the form that the matrix $\boldsymbol{D}(z)$ is allowed to take. Typical cases are:

Method 1.

$$\boldsymbol{D}(z)=d(z)\boldsymbol{I}.$$

This method forces each individual transfer-function element to have the same denominator. This leads to identical roots in the numerator and denominator, which implies that this is a nonminimal realization.

Method 2. $\boldsymbol{D}(z)$ is diagonal. The diagonal structure of $\boldsymbol{D}(z)$ decouples the identification problem into r subproblems, where r is the number of outputs. Therefore, each row i of transfer functions in $\boldsymbol{H}(z)$ is determined based upon measurements of y_i only. This representation is still nonminimal, because every row in $\boldsymbol{H}(z)$ is forced to have the same denominator, while having less pole-zero cancellations than Method 1.

Method 3. The set $[\boldsymbol{D}(z)\ \boldsymbol{N}(z)]$ forms an irreducible pair. The realization $[\boldsymbol{N}(z),\ \boldsymbol{D}(z)]$ of the system is irreducible if, and only if, Eq. 16 is satisfied for unimodular $\boldsymbol{X}(z)$. Proofs and algorithms for calculating irreducible $\boldsymbol{N}(z)$ and $\boldsymbol{D}(z)$ are given in Kailath (1980). For our purposes, it is sufficient to realize that an irreducible pair $(\boldsymbol{N}(z),\ \boldsymbol{D}(z))$ is a minimal realization and that procedures for finding such minimal realizations exist. $\boldsymbol{D}(s)$ and $\boldsymbol{N}(s)$ are polynomial matrices that satisfy the following conditions (Guidorzi, 1973):

$$\deg[N_{ij}(z)]\le\deg[N_{ii}(z)] \qquad \text{for} \qquad j<i$$

$$\deg[N_{ij}(z)]<\deg[N_{ii}(z)] \qquad \text{for} \qquad j>i$$

$$\deg[N_{ij}(z)]<\deg[N_{jj}(z)] \qquad \text{for} \qquad i\neq j$$

$$\deg[M_{ij}(z)]<\deg[N_{ii}(z)].$$

Typically, $\boldsymbol{D}(z)$ is a full matrix. The off-diagonal elements capture internal couplings between the outputs. This methodology has been developed for both standard (Guidorzi, 1975) and recursive (Gauthier and Landau, 1978) algorithms for the case where all inputs and outputs are sampled at a single rate.

In algebraic form, the linear difference equations that represent the dynamics of a MIMO system are

$$y_i(t)=-\sum_{k=1}^{g_{ij}}d_{iik}y_i(t-k)-\sum_{j=1}^{r}\sum_{k=1}^{g_{ij}}d_{ijk}y_j(t-k)+\sum_{j=1}^{q}\sum_{k=1}^{g_{ij}}n_{ijk}u_j(t-k)\quad \text{for}\quad t=1,\ldots,T;\ \ i=1,\ldots,r. \tag{17}$$

For Method 1 and Method 2, the off-diagonal terms of $\boldsymbol{D}(z)$, represented by d_{ijk}, $i\neq j$, are equal to zero.

Analogous to Eq. 5, the wavelet identification method may be applied by projecting both sides of Eq. 17 against the weighted modulating functions $c_p^{1/2}\varphi_p(t)$, $p=1,\ \ldots,\ L$, or

$$c_p^{1/2}\sum_t\varphi_p(t)y_i(t)=c_p^{1/2}\sum_t\varphi_p(t)\left\{-\sum_{k=1}^{g_{ij}}d_{iik}y_i(t-k)-\sum_{j=1}^{r}\sum_{k=1}^{g_{ij}}d_{ijk}y_j(t-k)+\sum_{j=1}^{q}\sum_{k=1}^{g_{ij}}n_{ijk}u_j(t-k)\right\}, \tag{18}$$

where the desired modulating functions (i.e., wavelets) are selected with respect to the modeling region of interest in an analogous manner to the SISO case.

It follows from Eq. 18 that in the general case the prediction of the output $y_i(t)$ is also a function of all other outputs, y_j, $j\neq i$, as well. The presence of more than a single output in the same equation places the restriction that these outputs must be evaluated at identical time scales. This is clearly not desirable for the case where the processes respond over different time scales, nor is it generally possible for processes sampled at multiple rates because the projection term $\Sigma_t\ \varphi_p(t)y_j(t-k)$ may not exist for all j. Therefore, for MIMO systems with multiple time scales and multisampling, Method 2, which decouples the identification into p independent subproblems, where p is the number of outputs, provides a properly formulated problem. This method is sometimes referred to as the transfer function output-by-output (TFOO) method. Thus, for a single output the model parameters may be estimated by projecting the set of modulating functions against the following set of linear equations:

$$y_i(t)=-\sum_{k=1}^{g_{ij}}d_{iik}y_i(t-k)+\sum_{j=1}^{q}\sum_{k=1}^{g_{ij}}n_{ijk}u_j(t-k). \tag{19}$$

It should be noted that due to the nonminimal realization, the estimated parameters in Eq. 19 may not be independent. This dependence expresses itself in terms of common zeros and poles, which will cancel under conditions of perfect modeling. In the presence of disturbance, noise, and model reduction, there will be near cancellation of these common zeros and poles.

Conclusions and Future Directions

In this article, the significance of the employed basis functions in process model identification was emphasized as the single most important source of effectiveness in modeling and possible problems in the ensuing controller design task, as well as how the properties of the basis functions influence the properties of the model. The need for process models, accurate over specific frequency regions and for certain time segments of data record, was similarly emphasized. Wavelets offer an excellent analytic framework for the construction of banks of band-pass filters, which can then be tuned to the needs of the control-relevant identification tasks, that is, select the segments of time records to be used in identification, and select the range of frequencies that are of interest for control purposes and weight appropriately the input–output information contained in these frequency ranges. The result is a system identification methodology that (1) can incorporate directly the engineering specifications on the system identification subtask, and (2) is applicable for nonstationary systems, including time-varying and nonlinear systems. The wavelet's logarithmic spanning of the time-frequency plane has been exploited to construct feedback control models for multirate systems as well.

The system identification techniques proposed in this article may be readily extended and combined with standard techniques in the literature (Lee and Chikkula, 1995) to construct adaptive control systems, with the property that the adaptive model will be constructed in the frequency range of interest, and the ability to discard data that have been corrupted by disturbances. In this regard, it is conceivable that certain of the weaknesses of the standard adaptive controllers, related to model adaptation based on corrupted information, or information at irrelevant frequencies, can be avoided.

Another area of potential future developments is related to the construction of true multiscale models for the description of process dynamics. The identification techniques proposed in this article have concentrated on the efficient decomposition of input–output data in the time-frequency space, while maintaining the discrete-time models as the essential framework for describing process dynamics. It is conceivable that the arguments made for the wavelet decomposition of the input–output signals in system identification could be extended to the decomposition of the model itself, thus leading to models that are defined over the binary tree of time–frequency space. Such models would allow the explicit incorporation of physical insights (pertaining to the multiscale character of the physical processes themselves), as well as the optimal fusion of measurements at different scales for the estimation of these multiscale process models. Typical example of inquiries in this direction is the work of Stephanopoulos et al. (1997).

Notation

$E[\cdot]$ = expected value
u = manipulated input
y = measured output
δ = Kronecker delta
λ = forgetting factor (Kalman filtering)

Literature Cited

Åström, K. J., and B. Wittenmark, *Computer Controlled Systems*, Prentice Hall, Englewood Cliffs, NJ (1984).

Åström, K. J., and B. Wittenmark, *Adaptive Control*, Addison-Wesley, Reading, MA (1989).

Bakshi, B., and G. Stephanopoulos, "Representation of Process Trends. IV. Induction of Real-time Patterns From Operating Data for Diagnosis and Supervisory Control," *Comput. Chem. Eng.*, **18**(4), 303 (1994).

Banerjee, A., and R. Pearson, "Multiple Model Based Estimation of Nonlinear Systems," AIChE Meeting, Miami Beach, FL, p. 183f (1995).

Bengtsson, G., "Output Regulation and Internal Models—A Frequency Domain Approach," *Automatica*, **13**, 333 (1977).

Braatz, R. D., and G. Mijares, "Control Relevant Identification and Estimation," AIChE Meeting, Miami Beach, FL, p. 183a (1995).

Carrier, J. F.,"The Application of Time-Frequency Techniques to Identification and Control," PhD Thesis, MIT, Cambridge, MA (1995).

Co, T. B., and B. E. Ydstie, "System Identification Using Modulating Functions and Fast Fourier Transforms," *Comput. Chem. Eng.*, **14**(10), 1051 (1990).

Cutler, C. R., and B. L. Ramaker, "Dynamic Matrix Control—A Computer Control Algorithm," *JACC Proc.*, San Francisco, CA (1980).

Daubechies, I., "Orthonormal Bases of Compactly Supported Wavelets," *Commun. Pure Appl. Math.*, **41**, 909 (1988).

Desoer, C. A., and Y. T. Wang, "Linear Time-Invariant Robust Servomechanism Problem: A Self-Contained Exposition," *Control and Dynamic Systems*, C. T. Leondes, ed., Vol. 16, Academic Press, New York, p. 81 (1980).

Doyle, J. C., "Analysis of Feedback Systems with Structured Uncertainties," *Proc. IEE, Part D.*, **129**, 242 (1982).

Doyle, J. C., and G. Stein, "Multivariable Feedback Design: Concepts for a Classical/Modern Synthesis," *IEEE Trans. Autom. Contr.*, **AC-26**, 4 (1981).

Gabor, D., "Theory of Communication," *J. Inst. Elec. Eng., London*, **93**, 429 (1946).

Gaikwad, S. V., and D. E. Rivera, "Control Relevant Identification of Ill-Conditioned Systems: Two High-Purity Distillation Case Studies," AIChE Meeting, Miami Beach, FL, p. 183i (1995).

Gautheir, A., and I. D. Landau, "On the Recursive Identification of Multi-Input, Multi-Output Systems," *Automatica*, **14**, 609 (1978).

Gopinath, R. S., and B. W. Bequette, "Multirate Model Predictive Control of Unconstrained Single Input-Single Output Processes," *Proc. Amer. Control Conf.*, Vol. 3, p. 2042 (1991).

Guidorzi, R., "Canonical Structures in the Identification of Multivariable Systems," *Automatica*, **11**, 361 (1975).

Hoo, K. A., and J. C. Kantor, "An Exothermic Continuous Stirred Tank Reactor Is Feedback Equivalent to a Linear System," *Chem. Eng. Commun.*, **37**, 1 (1985).

Jenkins, G. M., and D. G. Watts, *Spectral Analysis and Its Applications*, Holden-Day, San Francisco (1968).

Kailath, T., *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ (1980).

Kosut, R. L., "Adaptive Uncertainty Modeling: On-line Robust Control Design," *Proc. American Control Conf.*, Minneapolis, MN, p. 245 (1987).

LaMaire, R. O., L. Valavani, M. Athans, and G. Stein, "A Frequency-Domain Estimator for Use in Adaptive Control Systems," *Proc. American Control Conference*, Minneapolis, MN, p. 238 (1987).

Lee, J. H., and Y. Chikkula, "Improving Computational Efficiency of Model Predictive Control Algorithm Using Wavelet Transformation," *Int. J. Control*, **61**(4), 859 (1995).

Ling, W.-M., and D. E. Rivera, "Control Relevant Model Reduction of Volterra Series Models," AIChE Meeting, Miami Beach, FL, p. 183c (1995).

Ljung, L., *System Identification Theory for the User*, Prentice Hall, Englewood Cliffs, NJ (1987).

Maciejowski, J. M., *Multivariable Feedback Design*, Addison-Wesley, Reading, MA (1989).

Mallat, S. G., "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **11**(7), 674 (1989).

Matelinsky, V., "Identification of Continuous Dynamical Systems with Spline-Type Modulating Function Method," *IFAC Cong. on Parameter Identification and Parameter Estimation*, Darmstadt, Germany, p. 275 (1979).

Muske, K. R., and J. B. Rawlings, "Receding Horizon Recursive State Estimation," *Proc. Amer. Control Conf.*, San Francisco, CA, Vol. 1, p. 900 (1993).

Oppenheim, A. V., and R. W. Schafer, *Digital Signal Processing*, Prentice Hall, Englewood Cliffs, NJ (1989).

Oshima, M., and I. Hashimoto, "Multi-Rate Multivariable Model Predictive Control and Its Application to a Semi-Commercial Polymerization Reactor," *Proc. Amer. Control Conf.*, Chicago, IL, Vol. 2, p. 1576 (1992).

Palavajjhala, S., R. L. Motard, and B. Joseph, "Process Identification Using Discrete Wavelet Transforms: Design of Prefilters," *AIChE J.*, **42**(3), 777 (1996).

Pearson, A. E., and F. C. Lee, "Parameter Identification of Linear

Differential Systems via Fourier-Based Modulating Functions," *Control Theory Adv. Technol.*, **1**, 239 (1985).

Preisig, H. A., and D. W. T. Rippin, "Theory and Application of the Modulating Function Method—I. Review and Theory of the Method and Theory of the Spline-Type Modulating Functions," *Comput. Chem. Eng.*, **17**(1), 1 (1993a).

Preisig, H. A., and D. W. T. Rippin, "Theory and Application of the Modulating Function Method—II. Algebraic Representation of Matelinsky's Spline-Type Modulating Functions," *Comput. Chem. Eng.*, **17**(1), 17(1993b).

Preisig, H. A., and D. W. T. Rippin, "Theory and Application of the Modulating Function Method—II. Algebraic Representation of Matelinsky's Spline-Type Modulating Functions," *Comput. Chem. Eng.*, **17**(1), 17 (1993b).

Qin, S. J., and T. A. Badgwell, "An Overview of Industrial Model Predictive Control Technology," *Int. Conf. Chemical Process Control, CPC-V*, C. Garcia and J. Kantor, eds., *AIChE Symp. Ser. 316*, Vol. 93, 232 (1997).

Rioul, O., and M. Verletti, "Wavelets and Signal Processing," *IEEE Signal Process. Mag.*, **8**(4), 14 (1991).

Rivera, D. E., "Control-Relevant Parameter Estimation: A Systematic Procedure for Prefilter Design," *Proc. American Control Conf.*, Boston, MA, p. 237 (1991).

Shinbrot, M., "On the Analysis of Linear and Nonlinear Systems," *Trans. ASME*, 79, 547 (1957).

Skelton, R. E., "Model Error Concepts in Control Design," *Int. J. Control*, **49**, 1725 (1989).

Snee, R. D., "Validation of Regression Models. Methods and Examples," *Technometrics*, **19**, 415 (1977).

Stephanopoulos, G., M. Dyer, and O. Karsligil, "Multi-Scale Modeling, Estimation and Control of Processing Systems," *Proc. PSE-ESCAPE '97* (1997).

Stone, M., "Cross-Validity Choice and Assessment of Statistical Predictors," *J. Roy. Stat. Soc., Ser. B*, **36**, 111 (1974).

Strang, G., *Introduction to Applied Mathematics*, Wellesley-Cambridge Press, Wellesley, MA (1986).

Strang, G., "Wavelets and Dilation Equations," *SIAM Rev.*, **31**, 613 (1989).

Ziegler, J. G., and N. B. Nichols, "Optimum Settings for Automatic Controllers," *Trans. ASME*, **64**, 759 (1942).

Manuscript received Jan. 6, 1997, and revision received Nov. 19, 1997.