REASONING IN TIME: MODELING, ANALYSIS, AND PATTERN RECOGNITION OF TEMPORAL PROCESS TRENDS

Bhavik R. Bakshi

Department of Chemical Engineering
Ohio State University
Columbus, Ohio 43210

George Stephanopoulos

Laboratory for Intelligent Systems in Process Engineering
Department of Chemical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

The plain record of a variable's numerical values over time does not invoke appreciable levels of *cognitive* activity to a human. Although it can cause a fervor of numerical computations by a computer, the levels of cognitive appreciation of the variable's temporal behavior remain low. On the other hand, if one presents the human with a graphical depiction of the variable's temporal behavior, the level of cognition increases and a wave of reasoning activities is unleashed. Nevertheless, when the human is presented with scores of graphs, depicting the temporal behavior of interacting variables, that person's reasoning abilities are severely tested. In such case, the computer will happily continue crunching numbers without ever rising above the fray and thus developing a "mental" model, interpreting correctly the temporal interactions among the many variables.

Reasoning in time is very demanding, because time introduces a new dimension with significant levels of additional freedom and complexity. While the real-valued representation of variables in time is completely satisfactory for many engineering tasks (e.g., control, dynamic simulation, planning and scheduling of operations), it is very unsatisfactory for all those tasks, which require decision making via logical reasoning (e.g., diagnosis of process faults, recovery of operations from large unsolicited deviations, "supervised" execution of startup for shutdown operating procedures).

To improve the computer's ability to reason efficiently in time, we must first establish new forms for the representation of temporal behaviors. It is the purpose of this chapter to examine the engineering needs for temporal decisionmaking and to propose specific models that encapsulate the requisite temporal characteristics of individual variables and composite processes. Through a combination of analytical techniques, such as *scale-space filtering* and *wavelet-based, multiresolution decomposition of functions*, and modeling paradigms from *artificial intelligence* (AI), we have developed a concise framework that can be used to model, analyze, and synthesize the temporal trends of process operations. Within this framework, the modeling needs for logical reasoning in time can be fully satisfied, while maintaining consistency with the numerical tasks carried out at the same time. Thus, through the modeling paradigms of this chapter one may put together intelligent systems that use consistent representations for their logical reasoning and numerical tasks.

I. Introduction

Present-day computer-aided monitoring and control of chemical plants has caused an explosion in the amount of process information that can be conveyed to process operators and engineers. The real-time history of thousands of variables can be displayed and monitored. However, whereas a simple visual inspection of scores of displayed trends is sufficient to allow the operator confirm the process' status during normal, steady-state operations, when the process is in significant transience or crises have occurred the displayed trends and alarms can confound even the best of the operators. When the process variables change with different rates, or are affected by varying transportation lags, or inverse the response dynamics, or information is "suspicious" or lost because of sensors' malfunctioning, it is very difficult for a human operator to carry out routine tasks, such as the following:

- Distinguish normal from abnormal operating conditions (Bastl and Fenkel, 1980; Long and Kanazava, 1980).
- Identify the causes of process trends, e.g., external load disturbances, equipment faults, operational degradation, operator-induced mishandling.
- Evaluate current process trends and anticipate future operational states.
- Plan and schedule sequences of operating steps to bring the plant at the desired operating level, e.g., recover from safety fallback position, return to feasible operation after a fault.

The key recognitive skill required to carry out the above tasks is the formation of a "mental" model of the process operations that fits the current facts about the process and enables the operators to correctly assess process behavior and predict the effects of possible control actions. Correct "mental" models of process operations have allowed operators to overcome the weakness of "lost" sensors and conflicting trends, even under the pressure of an emergency (Dvorak, 1987), whereas most of the operational mishandlings are due to an erroneous perception as to what is going on in the process (O'Shima, 1983).

In order to develop intelligent, computer-aided systems with systematic and sound methodologies for the automatic creation of "mental" models of process operations, we need to resolve the following two and interrelated issues:

- What is the appropriate representational model for describing the "true" process trends, and how is it generated from the process data?
- How does one generate relationships among process trends in order to provide the desired "mental" model of process operations?

In the subsequent paragraphs and sections of this chapter we will see that these two issues impose requirements that transgress the abilities of simple "smoothing" filters and conventional regression techniques.

A. The Content of Process Trends: Local in Time and Multiscale

The time-dependent behavior of measured variables in a chemical process reflects the composite effect of many distinct contributions, coming from the underlying physicochemical phenomena and the status of processing equipment, sensors, and control valves. Thus, basic process dynamics, sensor noise, actuator dynamics, parameter drifts, equipment faults, external load disturbances, and operator-induced actions combine their contributions in some unknown way to form the temporal behavior of the measured operational data. As an example, consider the process signal shown in Fig. 1 (Cheung, 1992). It reflects the composite effect of contributions from five distinct sources; as a slow drift caused by a fouling process (source 1), equipment faults (sources 2 and 3), a periodic disturbance (source 4), and changing sensor noise (source 5). Ideally, we would like to have analytical techniques that can take the observed signal apart and render the exact temporal reproduction of each contributing signal. We know that this is theoretically impossible, but practically acceptable representations of the individual components is feasible and very valuable for the correct interpretation of process trends.

A systematic analysis of a process signal over (1) different segments of its time record and (2) various ranges of frequency (or *scale*) can provide a *local* (in time) and *multiscale* hierarchical description of the signal. Such description is needed if an intelligent computer-aided tool is to be constructed in order to (1) localize in time the "step" and "spike" from the equipment faults (Fig. 1), or the onset of change in sensor noise characteristics, and (2) extract the slow drift and the periodic load disturbance.

The engineering context of the need for multiscale representation of process trends can be best seen within the framework of the hierarchical

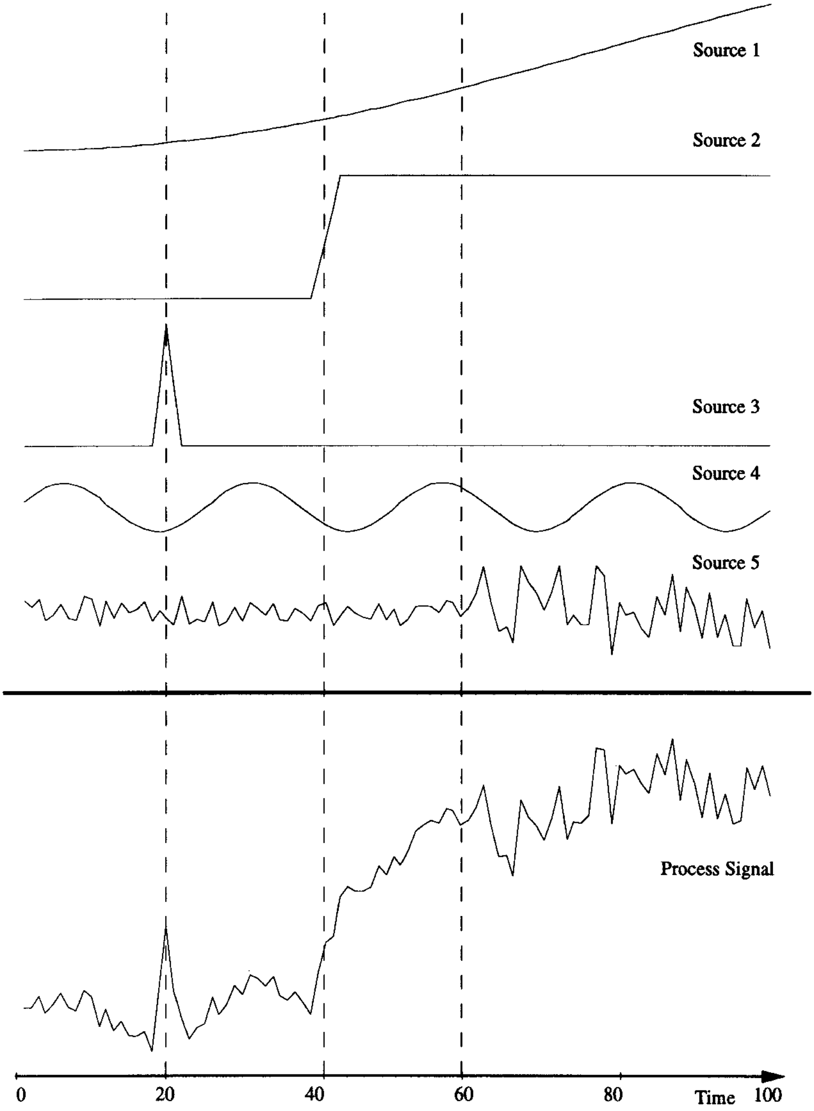


FIG. 1. A process signal and its component.

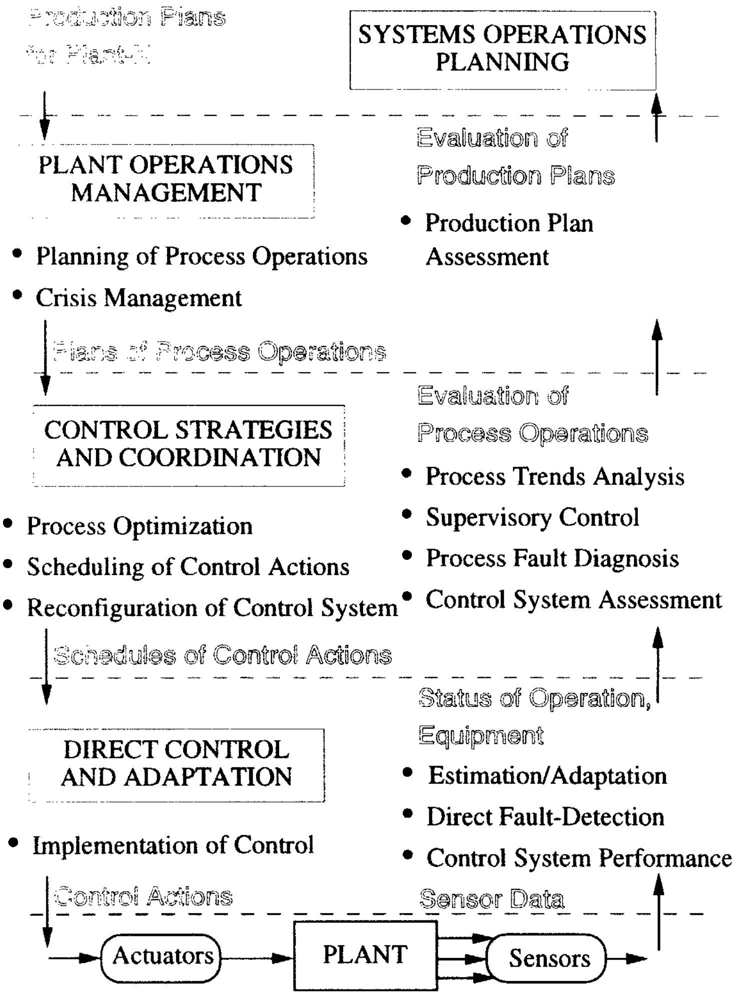

Production Plans
for Plant-K
SYSTEMS OPERATIONS PLANNING
PLANT OPERATIONS MANAGEMENT
Evaluation of Production Plans
• Production Plan Assessment
• Planning of Process Operations
• Crisis Management
Plans of Process Operations
Evaluation of Process Operations
CONTROL STRATEGIES AND COORDINATION
• Process Trends Analysis
• Supervisory Control
• Process Fault Diagnosis
• Control System Assessment
• Process Optimization
• Scheduling of Control Actions
• Reconfiguration of Control System
Schedules of Control Actions
Status of Operation, Equipment
DIRECT CONTROL AND ADAPTATION
• Estimation/Adaptation
• Direct Fault-Detection
• Control System Performance
• Implementation of Control
Control Actions
Sensor Data
Actuators
PLANT
Sensors

FIG. 2. Hierarchy of process operational tasks.

stratification of operational tasks, shown in Fig. 2 (Stephanopoulos, 1990). At the lowest level of abstraction, process data at a scale of seconds or minute are used to carry out a variety of numerical and logical tasks.

B. THE AD HOC TREATMENT OF PROCESS TRENDS

The term, *process trend*, undoubtedly carries an intuitive meaning about how process behavior changes over time. However, the exact meaning has never been formalized to allow one to articulate it in a general and concrete manner. For a long time, a process trend has been considered just as a smooth representation of a noisy signal, and consequently belonged to the province of digital signal processing community. However, the paradigms advanced are mostly ad hoc in character, and rarely can they sustain adequate fundamental physical justification. For example, during the design of digital filters, *the frequency response is selected* and the essential design task is to determine the coefficients of the filter that match the response. How do you come up, though, with the correct frequency response for the discovery of the "true" process trends in a signal? The answers have been essentially empirical and ad hoc, and have led to an explosive proliferation of design techniques. For example, the Fourier approach and over 100 windowing techniques (e.g., the uniform, von Hann, Hamming, and Kaiser windows) can be used for the design of nonrecursive filters, whereas the *impulse-invariant* method, the *bilinear transform*, and various computer-aided, trial-and-error techniques can be used for the design of recursive filters.

Furthermore, the current understanding of temporal process trends, produced by digital filters, is heavily founded on statistical considerations. Strong assumptions are often made about the statistical nature of the trend, in order to obtain "rigorously" optimal detection schemes. For example, the optimal Bayesian detection, the moving-average filters with adaptive controlled width, and the Wiener filters are only *theoretically optimal* (i.e., their practical implementation could yield grossly inadequate results) for (1) known signals with additive Gaussian noise (2) unknown signals with narrow band relative to the noise, and (3) random signals with known spectra, respectively (Van Trees, 1968; Papoulis, 1977). Although it is commonly agreed that a trend will appear if the random elements are removed from the signal, no filtering technique has incorporated a formal notion of the trend itself by eliciting its fundamental properties and its relation to the physicochemical process it describes.

Once a smooth signal has been constructed, how is the trend represented? Most of the available techniques do not provide a framework for the representation (and thus, interpretation) of trends, because their representations (in the frequency or time domains) do not include primitives that capture the salient features of a trend, such as continuity, discontinuity, linearity, extremity, singularity, and locality. In other words, most of the approaches used to "represent" process signals are in fact *data compaction techniques*, rather than trend representation approaches. Furthermore, whether an approach employs a frequency or a time-domain representation, it must make several major decisions before the data are compacted. For frequency-domain representations, assumptions about the origin of the trend must be made, e.g., whether it is stationary (i.e., frequency content independent of time), quasistationary (i.e., frequency content varies slowly with time), or nonstationary. For the curve-fitting representations in the time domain, the selection of the particular functional model and the thresholds for fitting errors are ad hoc decisions.

It is clear from the preceding discussion that *the deficiencies of the existing frequency- and time-domain representations of process trends stem from their procedural character* (Cheung and Stephanopoulos, 1990), i.e. they represent trends as the outputs of a computational process, which quite often bears no relationships to the process physics and chemistry. What is needed is a *declarative* representation, which can capture explicitly all the desirable characteristics of process trends.

C. Recognition of Temporal Patterns in Process Trends

The correct interpretation of measured process data is essential for the satisfactory execution of many computer-aided, intelligent decision support systems that modern processing plants require. In supervisory control, detection and diagnosis of faults, adaptive control, product quality control, and recovery from large operational deviations, determining the "mapping" from process trends to operational conditions is the pivotal task. Plant operators "skilled" in the extraction of real-time patterns of process data and the identification of *distinguishing features* in process trends, can form a "mental" model on the operational status and its anticipated evolution in time.

A formal induction of mappings from measured operating data to process conditions is composed of the following three tasks (Fig. 3):

Task 1. Extraction of pivotal, temporal features from process data.

Task 2. Inductive learning of the relationship between the features of process trends and process conditions.

Task 3. Adaptation of the relationship utilizing future operating data.

Linear, polynomial, or statistical discriminant functions (Fukunaga, 1990; Kramer, 1991; MacGregor *et al.*, 1991), or adaptive connectionist networks (Rumelhart *et al.*, 1986; Funahashi, 1989; Vaidyanathan and Venkatasubramanian, 1990; Bakshi and Stephanopoulos, 1993; third chapter of this volume, Koulouris *et al.*), combine tasks 1 and 2 into one and solve the corresponding problems simultaneously. These methodologies utilize a priori defined general functional relationships between the operating data and process conditions, and as such they *are not inductive*. Nearest-neigh-

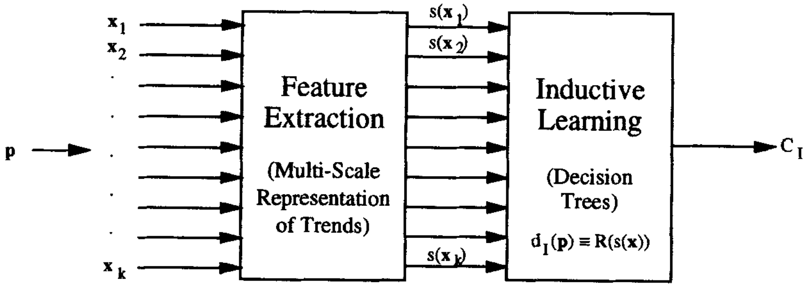


FIG. 3. Inductive generalization of pattern-based relations between process variables and operating conditions.

bor classifiers (Silverman, 1986; Kanal and Dattatreya, 1985), case-based analogical reasoners (Jantke, 1989), or inductive decision trees (Saraiva and Stephanopoulos, 1992), do not assume a functional model and produce truly inductive strategies.

The extraction, though, of the so-called pivotal features from operating data, encounters the same impediments that we discussed earlier on the subject of process trends representation: (1) localization in time of operating features and (2) the multiscale content of operating trends. It is clear, therefore, that any systematic and sound methodology for the identification of patterns between process data and operating conditions can be built only on formal and sound descriptions of process trends.

D. COMPRESSION OF PROCESS DATA

The expression "a mountain of data and an ant hill of knowledge" can be used to signify, besides its obvious message, the significance of the task of maintaining and sorting the vast amounts of data accumulated by present-day, digital process data acquisition systems. Efficient compression, storage, and recovery of historical process data is essential for many engineering tasks. Data compression methodologies must store with minimum distortion of qualitative and quantitative features, independently of whether these features represent behavior *localized in time* or a particular range of frequencies (i.e., *scale*). In other words, the successful compression of process data must necessarily go through an efficient representation of process trends (see Section I, A). Data compression has received surprisingly little attention by the chemical engineering practice, despite the clear articulation of the many benefits that can be drawn from it (Hale and Sellars, 1981; Bader and Tucker, 1987a, b). The *boxcar* and *backward* *slope* methods are fast, and produce piecewise linear interpolations of the data within predefined, acceptable error bounds, but are unable to capture small process transients and upsets. More techniques (Feehs and Arce, 1988; Bakshi and Stephanopoulos, 1995) have been inspired by technologies used for the compression of speech, images, and communications data, e.g., *vector quantization* (Gray, 1984), and *wavelet decomposition* theory (Mallat, 1989).

The primary objective of any data compression technique is to transform the data to a form that requires the smallest possible amount of storage space, while retaining all the relevant information. The desired qualities of a technique for efficient storage and retrieval of chemical process data are as follows:

1. Compacted data should require minimum storage space.
2. Compaction and retrieval should be fast, often in real time.
3. A clear and explicit measure of the quality of the signal, obtained after retrieval, should be available, to be used as a criterion for guiding the compression.
4. The retrieved signal should have minimum distortion and should contain all the desired features.
5. The compaction should be based on physically intuitive criteria and should require minimum a priori assumptions.

None of the practiced compression techniques satisfies all of these requirements. In addition, it should be remembered that compression of process data is not a task in isolation, but it is intimately related to the other two subjects of this chapter: (1) description of process trends and (2) recognition of temporal patterns in process trends. Consequently, we need to develop a common theoretical framework, which will provide a uniformly consistent basis for all three needs. This is the aim of the present chapter.

E. Overview of the Chapter's Structure

Section II introduces the formal framework for the definition and description of process trends at all levels of detail: qualitative, order-of-magnitude, and analytic. A detour through the basic concepts of *scale-space filtering* is necessary in order to see the connection between the concept of process trends and the classical material on signal analysis. Within the framework of scale-space filtering we can then elucidate the notions of "episode," "scale," "local filtering," "structure of scale," "distinguished features," and others.

In Section III we introduce the theory of the multiresolution analysis of signals using *wavelet decomposition*, which is used to provide the scale-space image of a function with correct local characteristics. This localized description of a signal's features allows the correct extraction of distinguished attributes from a signal, a task that forms the basis for the inductive generation of pattern-based logical relationships among input and output variables, or the efficient compaction of process data. Of particular value is the construction of *translationally invariant* wavelet decompositions, which allow the correct formation of temporal patterns in process variables. Using translationally invariant decomposition of signals, Section III contains specific methodologies for the construction of the *wavelet interval–tree of scale*, the extraction of distinguished features and the generalization of process trends.

The ideas presented in Section III are used to develop a concise and efficient methodology for the compression of process data, which is presented in Section IV. Of particular importance here is the conceptual foundation of the data compression algorithm; instead of seeking noninterpretable, numerical compaction of data, it strives for an explicit retention of distinguished features in a signal. It is shown that this approach is both numerically efficient and amenable to explicit interpretations of historical process trends.

In Section V we discuss how the temporal distinguished features of many input and output signals can be correlated in propositional forms to provide logical rules for the diagnosis and control of processes, which are difficult to model. *Inductive decision trees* are used to capture the knowledge contained in previous operating data. The explicit description of process trends and the explicit statement of the knowledge "mined" from the data, overcome many of the real-world obstacles in applying these ideas at the manufacturing floor.

Throughout the remaining four sections of this chapter, we will provide illustrations on the use of the various techniques, using real-world case studies.

II. Formal Representation of Process Trends

Although the term "process trend" emulates a certain intuitive understanding in the minds of the speaker and the listener, this understanding may not be the same. Certainly, we do not have a clear, sound, and unambiguous definition of the term "trend," and this must be the first issue to be addressed.

A. The Definition of a Trend

In order to represent and reason with temporal information, we need to represent time explicitly and concisely. The discrete-time character of computer-aided data acquisition and control dictates that time should be represented as a sequence of strictly increasing time points:

$$t = \{t_{-\infty} \cdots t_i \cdots t_j \cdots t_\infty\},$$

where

$$t_i < t_j \Leftrightarrow i < j \qquad \text{for all integer } i, j.$$

Furthermore, a *time interval*, I_{ij}, is defined to be an open interval of time as follows:

$$I_{ij} = (t_i, t_j), \qquad \text{where} \quad t_i < t_j.$$

Thus, we may represent time as a sequence of open intervals separated by time points, or a sequence of time intervals. We will adopt the second interpretation, considering a time point as an interval of zero duration.

1. From the Quantitative to the Qualitative Representation of a Function

Consider a real-valued function, $x(t)$, with the following properties over a time interval $[a, b]$:

1. $x(t)$ is continuous over $[a, b]$ but is allowed to have a finite number of discontinuities in its value or/ and first derivative.
2. $x'(t)$ and $x''(t)$ are continuous in (a, b), and their one-sided limits exist at a and b.
3. $x(t)$ has a finite number of extrema and inflexion points in $[a, b]$.

Variables that satisfy the above requirements will be called *reasonable variables* (Cheung and Stephanopoulos, 1990). In defining the "reasonableness" of a function, we are concerned only with the properties of the function's value, first and second derivatives. Such definition is less restrictive (it does not require existence of all derivatives), but it is completely general and allows the characterization of a function at different levels (Cheung and Stephanopoulos, 1990). All the physical variables encountered in the operation of a plant are reasonable.

Since we elected to represent time as a sequence of time intervals, we consider that the *state* of a reasonable variable is completely known, if we know the value and the derivative of the variable over a time interval. As the duration of the defining time interval approaches zero, we take the following definition of the state (continuous) of the variable at a given time point (Cheung and Stephanopoulos, 1990):

a. State (Continuous). The state (continuous), $CS(x,t)$ of a reasonable variable, $x:[a,b] \in R \ \forall t \in [a,b]$, is the point value $PtVl(x,t)$ defined by a triplet as follows:

1. If x is continuous at t, then

$$CS(x,t) \equiv PtVl(x,t) = \langle x(t), x'(t), x''(t) \rangle.$$

2. If x is discontinuous at t, then

$$\begin{aligned} CS(x,t) &\equiv PtVl(x,t) \\ &= \langle \langle x_{\mathrm{L}}(t), x_{\mathrm{R}}(t) \rangle, \langle x_{\mathrm{L}}(t), x_{\mathrm{R}}(t) \rangle, \langle x''_{\mathrm{L}}(t), x''_{\mathrm{R}}(t) \rangle \rangle, \end{aligned}$$

where the subscripts L and R denote the left- and right-side limits at the discontinuity.

Consequently, the *trend* (continuous) of a variable can be defined as follows

b. Trend (Continuous). The continuous trend of a reasonable variable, $x:[a,b] \rightarrow R$, is given by the continuous sequence of states over $[a,b]$.

As we go from continuous to discrete-time representations, the time is represented by a strictly increasing sequence of points. In such case, the state (discrete) of a reasonable function and the associated trend are given by the following definitions:

c. State (Discrete). The state (discrete), $DS(x,t)$ of a reasonable function, $x:[a,b] \rightarrow R$, over a set of strictly increasing time points $T = \{a = t_0, \dots, t_j, \dots, t_n = b\} \subseteq [a,b]$, is defined as follows:

1. If $t \in T$, then $DS(x,t) = PtVl(x,t)$.
2. If $t_i < t < t_{i+1}$ for $i = 0, 1, \dots, n$, then $DS(x,t) = \langle PtVl(x,t_i), PtVl(x,t_{i+1}) \rangle$.

d. Trend (Discrete). The discrete trend of a reasonable variable, $x:[a,b] \rightarrow R$, is given by the set of discrete states corresponding to the strictly increasing time points of the time interval $[a,b]$.

Clearly, the quantitative description of the discrete state and of the discrete trend must be declaratively explicit, since we cannot perform differentiation at the single points defining the intervals of $[a,b]$. This is a strict requirement and can be theoretically met only if we know the underlying continuous function that provides the values of the derivatives at the time points of a discrete representation. The availability, though, of such a continuous function is based on a series of ad hoc decisions on the character and properties of the functions, and if one prefers to avoid them, then one must accept a series of approximations for the evaluation of first and second derivatives. These approximations provide a sequence of representations with increasing abstraction, leading, ultimately, to qualitative descriptions of the state and trend as follows (Cheung and Stephanopoulos, 1990):

e. State (Qualitative). Let $x:[a,b] \rightarrow R$ be a reasonable function. $QS(x,t)$, the qualitative state of x at $t \in [a,b]$, is defined as the triplet of qualitative values as follows:

$$QS(x,t) = \left\{\begin{array}{ll} \text{undefined} & \text{if } x \text{ is discontinuous at } t \\ \langle [x(t)], [\partial x(t)], [\partial\partial x(t)] \rangle & \text{otherwise,} \end{array}\right\},$$

where

$$[x(t)] = \begin{cases} + & \text{if} \quad x(t) > 0, \\ 0 & \text{if} \quad x(t) = 0, \\ - & \text{if} \quad x(t) < 0; \end{cases}$$

$$[\partial x(t)] = \begin{cases} + & \text{if} \quad x'(t) > 0, \\ 0 & \text{if} \quad x'(t) = 0, \\ - & \text{if} \quad x'(t) > 0; \end{cases}$$

$$[\partial\partial x(t)] = \begin{cases} + & \text{if} \quad x''(t) > 0, \\ 0 & \text{if} \quad x''(t) = 0, \\ - & \text{if} \quad x''(t) > 0. \end{cases}$$

f. Trend (Qualitative). The qualitative trend of a reasonable variable, $x:[a,b] \rightarrow R$, is the continuous sequence of qualitative states over $[a,b]$.

2. Episodes and Trends

The practical value of the qualitative state and trend lies in the fact that both are very close to the intuitive notions employed by humans in interpreting the temporal behavior of signals. But humans capture the trend as a finite sequence of ordered segments with constant qualitative state over each segment. To emulate a similar notion we introduce the concept of *episode* through the following definition (Cheung and Stephanopoulos, 1990):

a. Episode. Let $x:[a,b] \rightarrow r$ be a reasonable function. For any time interval $I=(t_i,t_j) \subseteq [a,b]$ such that the qualitative state, $QS(x,t)$ is constant for $\forall t \in (t_i,t_j)$, an episode, E, of x over (t_i,t_j) is the pair, $E=\langle I, QS(x,I)\rangle$, with (1) I signifying the temporal extent of an episode and (2) $QS(,I)=QS(x,t)$ $\forall t \in (t_i,t_j)$ characterizing the constant qualitative state over I. Whenever two episodes, defined over adjacent time intervals, have the same qualitative state values, they can be combined to form an episode with broader temporal extent. On the other hand, if a qualitative state is constant over a time interval, then it is also constant over any of its time subintervals. These observations lead to the following definition of a *maximal episode.*

b. Maximal Episode. An episode, E_1, is maximal if there is no episode, E_2, such that (1) E_1 and E_2 have the same qualitative state and (2) $I_1 \subseteq I_2$, i.e., the temporal extent of E_1 is contained in the temporal extent of E_2.

From the definition of the maximal episode we conclude that the maximal episodes occur between adjacent time points, at which $x(t)$, $x'(t)$ or/ and $x''(t)$ change qualitative value. We will call these points *distinguished time points*, and from now on we will employ the following definition of a *trend*:

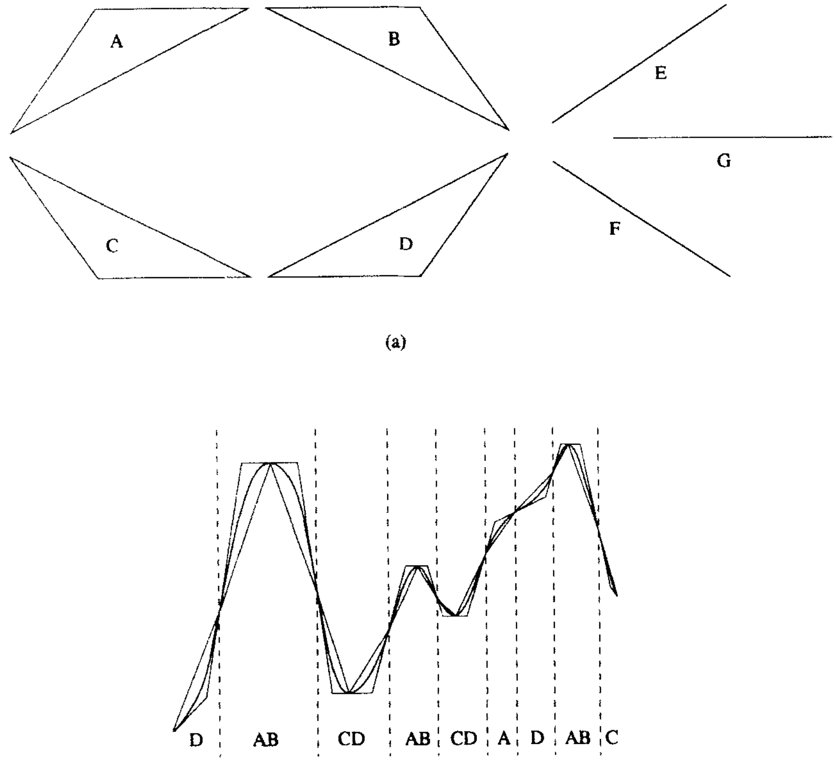
c. Trend. The trend of a reasonable function, $x:[a,b] \rightarrow R$ is a sequence of maximal episodes, defined over time intervals whose distinguished points are strictly ordered in time.

3. Triangular Episode: A Geometric Language to Describe Trends

If a trend is to be described by an ordered sequence of maximal episodes, then we only need to generate a language that contains the declarative value of an episode. Such a language was proposed by Cheung and Stephanopoulos (1990). It is composed of seven primitives (Fig. 4a) and possesses the following properties:

Completeness. Every trend can be represented by a legal sequence of triangular episodes (Fig. 4b).

Correctness. The recursive refinement of triangular episodes allows the description of a trend at any level of detail, converging to the real-valued description of a signal.



(a)

(b)

FIG. 4. (a) The seven primitives of the triangular description of trends; (b) sequence of triangular episodes describing a specific "trend" of a signal.

Robustness. The relative ordering of the triangular episodes in a trend is invariant to scaling of both the time axis and the function value. It is also invariant to any linear transformation (e.g., rotation, translation). Finally it is quite robust to uncertainties in the real value of the signal (e.g., noise), provided that the extent of a maximal episode is much larger than the period of noise.

B. TRENDS AND SCALE-SPACE FILTERING

Consider the continuous function shown in Fig. 5a. Points 1 through 1: are the inflextion points and constitute a subset of all the distinguishec points over the indicated period of time, and define the followin

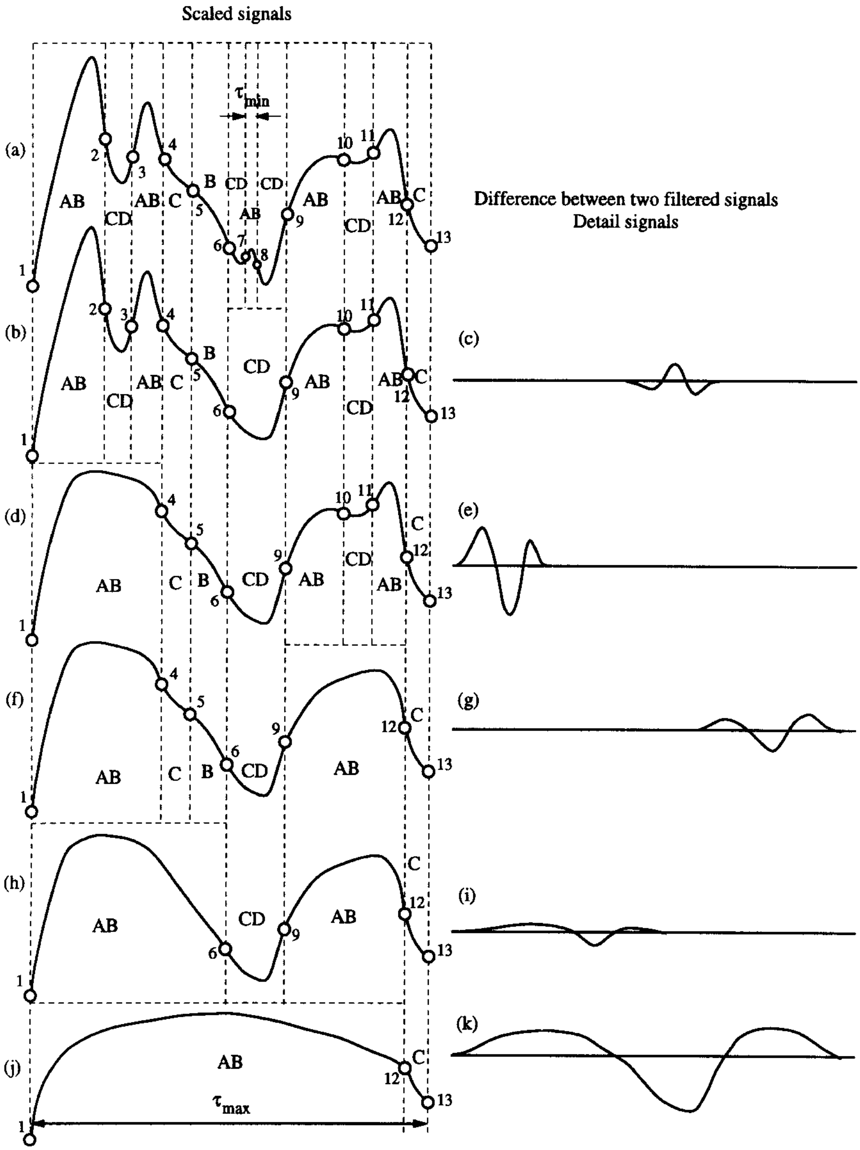


FIG. 5. Multiscale representation of process trends.

sequence of groupings of the triangular episodes:

$$AB(1,2) - CD(2,3) - AB(3,4) - C(4,5) - B(5,6)$$
$$- CD(6,7) - AB(7,8) - CD(8,9) - AB(9,10)$$
$$- CD(10,11) - AB(11,12) - C(12,13).$$

Let us filter the signal of Fig. 5a with a compact filter of variable width of compactness. The first distinguished feature to disappear from the trend of Fig. 5a is the sequence, $CD(6,7) - AB(7,8) - CD(8,9)$ (defined by the inflexion points, 6–9), which is replaced by the $CD(6,9)$ grouping of triangular episodes (see Fig. 5b). The trend of the filtered function (Fig. 5b) differs from the trend of the original one (Fig. 5a) over the time segment defined by the inflexion points 6–9. In an analogous manner, as the width of compactness of the local filter increases, additional features of the trend are replaced with more abstract descriptions. Examples of this process from Fig. 5 are the following:

(a) $\{AB(1,2) - CD(2,3) - AB(3,4)\}$ replaced by $\{AB(1,4)\}$ (Fig. 5d)

(b) $\{AB(9,10) - CD(10,11) - AB(11,12)$ replaced by $\{AB(9,12)]$ (Fig. 5f)

(c) $\{AB(1,4) - C(4,5) - B(5,6)\}$ replaced by $\{AB(1,6)\}$ (Fig. 5h)

(d) $\{AB(1,6) - CD(6,9) - AB(9,12)\}$ replaced by $\{AB(1,12)\}$ (Fig. 5j)

From Fig. 5 we conclude that the original function can be represented by six (6) distinct trends (Fig. 5a, b, d, f, h, j), *each with its own sequence of triangular episodes*. Each successive trend of Fig. 5 contains information at a coarser resolution (scale). The differences among two successive trends are shown in Figs. 5c, 5e, 5g, 5i, and 5k, assuming for presentation purposes perfectly local filters.

The procedure described above is a pictorial approximation of a process called *scale-space filtering* of a function, proposed by Witkin (1983). The surface (e.g., Fig. 6) swept out by a filtered signal as the Gaussian filter's standard deviation is varied, is called *scale-space image* of the signal and is given by

$$F(t,\sigma) = f(t)^{*}g(t,\sigma) = \int_{-\infty}^{+\infty} f(u)\left\{\frac{1}{\sigma\sqrt{2\pi}}\exp\left[\frac{-(t-u)^2}{2\sigma^2}\right]\right\} du, \quad (1)$$

and gives the time-dependent behavior of the signal, $f(t)$, at different scales. As σ increases certain distinguishing features disappear, replaced

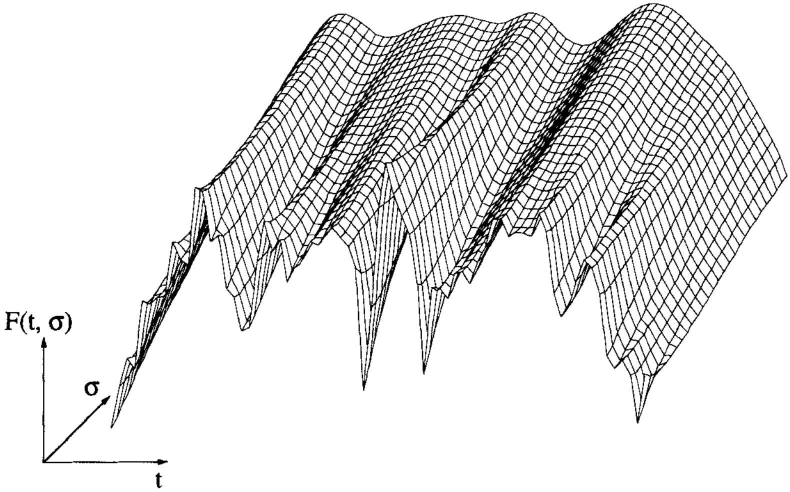


FIG. 6. Scale-space image of a function $F(t)$.

by coarser characteristics, very much in the spirit described by the pictorial examples of Fig. 5.

1. Structure of Scale

A close examination of the scale-space image of Fig. 6 reveals some interesting features:

1. As σ increases, pairs of inflexion points disappear and give rise to a smoother representation of the signal.
2. As σ decreases, a smooth segment of the signal goes through a singular point and gives rise to a "ripple" with the generation of a pair of inflexion points.
3. Some of the inflextion points persist over a large range of σ values, while others disappear. Clearly, the former correspond to *dominant features of the signal* while the latter indicate weak features, e.g., noise, which disappear readily with filtering.
4. The position of the inflexion points at higher values of σ, i.e., higher scales, shifts as a result of the increasingly global effect of the Gaussian filter.
5. If we connect the positions of the same inflexion points over various values of σ by straight lines, we create an *interval tree of scales*, as shown in Fig. 7 for the signals of Fig. 5. The interval tree allows us to generate two very important pieces of information about the trends of a measured variable:

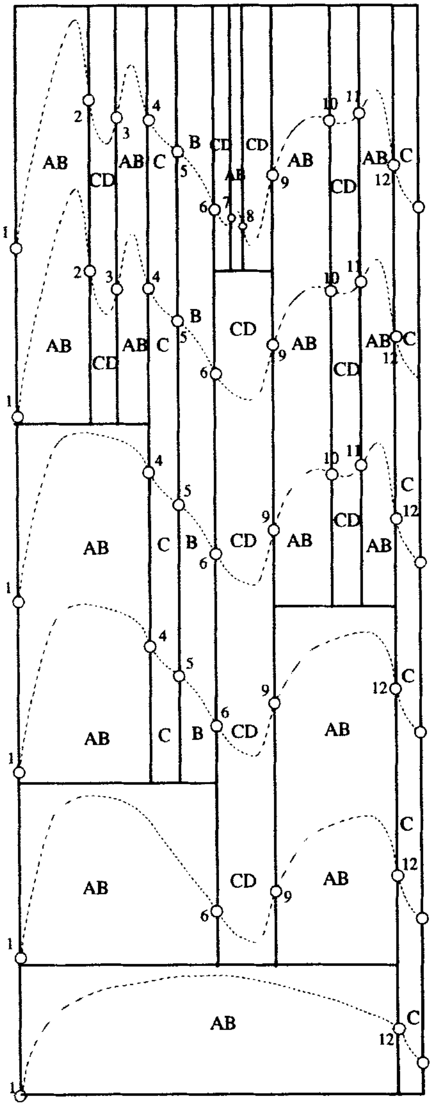


FIG. 7. The interval tree of scales for the signal of Fig. 5.

a. Number of Distinct Trends at Multiple Scales. The disappearance o inflexion points through filtering, or their generation through refinemen changes the sequence of triangular episodes needed for the description o a measured variable. This, in turn, implies that a new trend has bee identified. Having the interval tree of scales, we can easily generate th complete set of trends that can be used to describe a measured signal at various scales of abstraction (or, detail). For example, the signal of Fig. 5 can be described by any of six (6) distinct trends.

b. Dominant Features in a Signal. The inflexion points which persist over extended ranges of σ (or scale), bound dominant features of a signal. For example, consider the signal of Fig. 5. The pairs of inflexion points (1, 6) and (9, 12) bound the features, $AB(1,6)$ and $AB(9,12)$ (see Fig. 5h), respectively, and have "survived" four levels of filtering. They correspond to dominant features of the original signal and as such characterize profoundly the signal itself.

2. *Scaling Episodes and the Second-Order Zero Crossings*

Scale-space filtering provides a multiscale description of a signal's trends in terms of its inflexion points (second-order zero crossings). The only legal sequences of triangles between two adjacent inflexion points are (in terms of triangular episodes):

AB
CD
A (if it is preceded and succeeded by D episodes)
B (if it is preceded and succeeded by C episodes)
C (if it is preceded and succeeded by B episodes)
D (if it is preceded and succeeded by A episodes)

Any of these legal sequences will be called a *scaling episode.*

If a signal is represented by a sequence of triangular episodes, scale-space filtering manipulates the sequences of triangular episodes with very concrete mechanisms. Here is the complete list of syntactic manipulations carried out by scale-space filtering:

1. $AB - CD - AB \Rightarrow AB$
2. $CD - AB - CD \Rightarrow CD$
3. $AB - C - B \Rightarrow AB$
4. $CD - A - D \Rightarrow CD$
5. $D - A - D \Rightarrow D$
6. $C - B - C \Rightarrow C$
7. $B - C - B \Rightarrow B$
8. $A - D - A \Rightarrow A$

The reader should note the following properties:

(a) Filtering occurs over a segment of *three* scaling episodes, producing a segment with *one* scaling episode.

(b) The syntax of the resulting scaling episode is determined by the type of the first and last episodes in the sequence of the three scaling episodes, e.g., $\underline{A}B - CD - A\underline{\underline{B}} \Rightarrow AB$, $\underline{B} - C - \underline{\underline{B}} \Rightarrow BB \equiv B$.

So, within the context of scale-space filtering, it is more convenient to express *a trend as a sequence of scaling episodes*, rather than as a sequence of episodes.

The utility of representing trends in terms of scaling episodes, or equivalently in terms of *second-order zero crossings* (i.e., inflexion points) increases even more, if we can reconstruct a signal from these zero crossings at multiple scales. Marr and Hildreth (1980), using the Laplacian of a Gaussian had surmised that the reconstruction was stable and complete, but gave no proof. Yuille and Poggio (1984) proved that the scale map of almost all signals filtered by a Gaussian of varying σ determines the signal uniquely, up to a constant scaling. Hummel and Moniot (1989) on the other hand have shown that reconstruction using second-order zero crossings alone is unstable, but together with gradient values at the zero crossings is possible and stable, although it uses a lot of redundant information. Mallat and Zhong (1992) have developed an algoalgorithm that converges quickly to stable reconstructions using second-order zero crossings of a signal, derived from the wavelet decomposition of the signal. Nevertheless, the completeness and stability of the reconstruction from second-order zero crossings remains an open theoretical question.

3. *Properties of Scale-Space Filtering*

The interval tree of scale allows us to extract the temporal features contained in a signal and systematically establish the trends at different scales. Witkin (1983) defined *the stability of a temporal episode as the number of scales over which it persists*. The most conspicuous features of a signal are also the most stable. He also developed a heuristic procedure through which he would construct representations of a signal by maintaining the most stable episodes. The stability criterion, though is entirely heuristic, and other criteria may be desired. In Section III we will see that wavelet decomposition of signals offers a sound and much more powerful approach in constructing stable representations of dominant trends. For the time being, it is very important to identify several important properties and shortcomings of the traditional scale-space filtering.

a. Distortion of Signal's Features. As a signal is filtered through a Gaussian, its features become progressively quantitatively more and more distorted. This is a natural effect of the averaging of the smoothing process over increasingly longer segments of a signal. Such distortion inhibits the extraction of accurate distinguished features. Feature extraction by selecting stable episodes from the interval tree involves fitting piecewise quadratic segments to the raw data covered by the stable episodes. This procedure is ad hoc, and creates discontinuities at the inflexion points.

b. Creation of Fictitious Features. Gaussian filtering has two very important properties:

1. Inflexion points never disappear as we move from more abstract to more detailed descriptions.
2. Fictitious inflexion points are not generated as the scale of filtering increases.

Taken together, these properties guarantee that Gaussian filtering never generates fictitious features. For many other filters, this is not the case. It is important that any filter we use maintains the integrity of the original signal.

III. Wavelet Decomposition: Extraction of Trends at Multiple Scales

The wavelet decomposition of continuous and discrete functions has emerged as a powerful theoretical framework over the last 10 years, and has led to significant technical developments in data compression, speech and video recognition and data analysis, fusion of multirate data, filtering techniques, etc. Starting with Morlet's work (Goupillaud *et al.* 1984) on the analysis of seismic data, wavelet decomposition of signals has become extremely attractive for two main reasons: (1) it offers excellent localization of a signal's features in both time and frequency and (2) it is numerically far more efficient than other techniques such as fast Fourier transform of signals (Bakshi and Stephanopoulos, 1994a). It is for the first reason that we discuss in this section the use of wavelet decomposition as the most appropriate framework for the extraction and representation of trends of process operating data.

A. The Theory of Wavelet Decomposition

A family of wavelets is a family of functions with all its members derived from the translations (e.g., in time) and dilations of a single, mother function. If $\psi(t)$ is the mother *wavelet*, then all the members of the family are given by

$$\psi_{su}(t) \equiv \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right) \qquad \text{for} \quad (s,u) \in R^2. \tag{2}$$

the parameters s and u that label the members of a wavelet family are known as the *dilation* and *translation* parameters, respectively. A wavelet function, $\psi(t)$, belongs to the set of square-integrable functions, i.e., $\psi(t) \in L^2(t)$, and satisfies the following admissibility condition

$$C_\psi = \int_0^\infty \frac{|\hat{\psi}(\omega)|}{\psi}\, d\omega < +\infty, \tag{3}$$

with $\hat{\psi}(\omega)$ signifying the Fourier transform of $\psi(t)$. Condition (3) implies that as $\omega \to 0$, $\hat{\psi}(\omega)$ approaches zero faster and thus represents a *band-pass filter*. $\psi(t)$ is either compactly supported in time, or has some fast decay at infinity implying that condition (3) is equivalent to the vanishing of the first p moments (see Section III of third Chapter in this volume):

$$\int_{-\infty}^{\infty} t^m \psi(t)\, dt = 0, \qquad m = 0,1,2,\ldots,p-1 \tag{3a}$$

where p is the order of approximation considered for $\psi(t)$. Figure 8 shows examples of various wavelet functions.

1. *Resolution in Time and Frequency*

In the time domain, a wavelet placed at a translation point, u, has a standard derivation, σ_u, around this point. Similarly, in the frequency domain a wavelet is centered at a given frequency, ω (determined by the value of the dilation parameter) and has a standard deviation, σ_ω, around the specific frequency. The values of σ_u and σ_ω are given by

$$\sigma_u^2 = -\int_{-\infty}^{+\infty} t^2 |\psi(t)|^2\, dt \qquad \sigma_\omega^2 = \int_{-\infty}^{+\infty} \omega^2 |\hat{\psi}(\omega)|^2\, d\omega \tag{4}$$

As the dilation parameter, s, increases, the wavelets have significant value: over a broader time segment and the resulting resolutions in the time anc frequency domains change as $s\sigma_u$ and σ_ω / s, respectively. Thus, at large

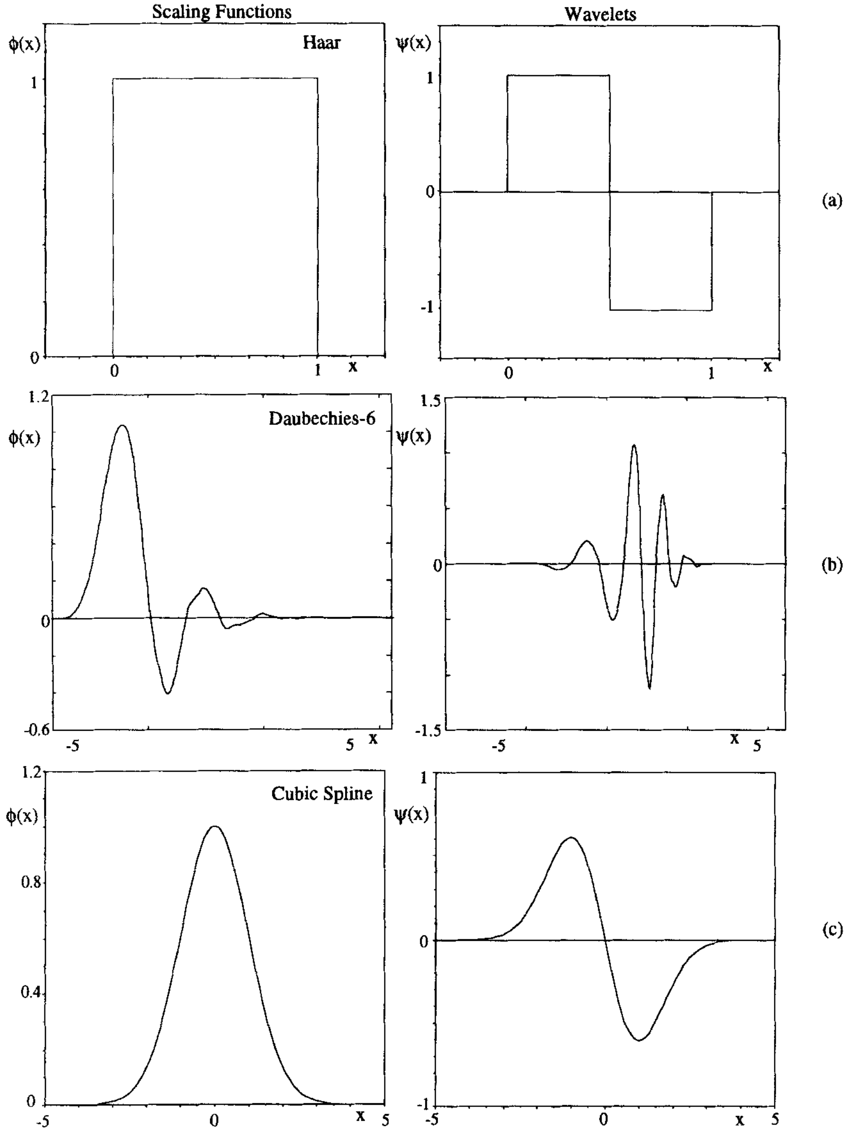


FIG. 8. Typical wavelets and scaling functions: (a) Haar, (b) Daubechies-6, (c) cubic spline.

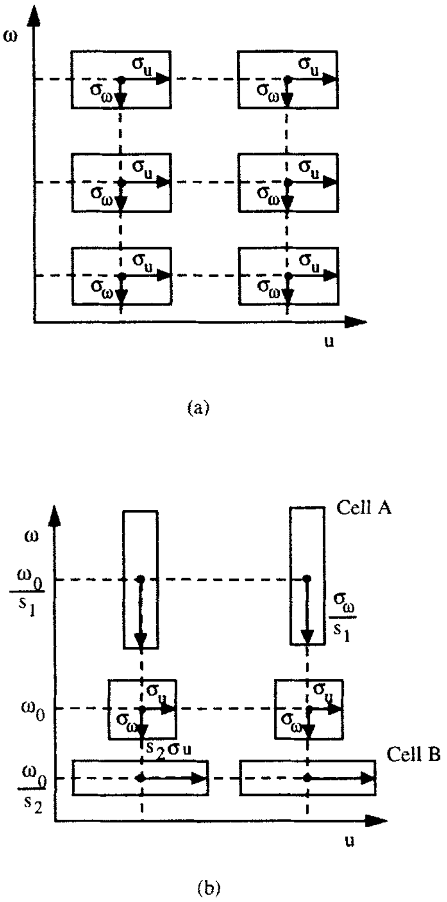


FIG. 9. Resolution in scale space of (a) window Fourier transform and (b) wavelet transform.

dilations (i.e., *scales*) the resolution in the time domain becomes coarser, while in the frequency domain, it becomes finer. The dimensions of the resolution cell (Fig. 9b) in the *scale-space* is equal to

$$[u_0 - s\sigma_u, u_0 + s\sigma_u] \times \left[\frac{\omega_0}{s} - \frac{\sigma_\omega}{s}, \frac{\omega_0}{s} + \frac{\sigma_\omega}{s}\right],$$

where u_0 is the translation point of a wavelet and consequently the center of its energy's concentration in the time domain, and ω_0/s is the frequency where the wavelet's energy is concentrated in the frequency domain. The variable dimensions of the resolution cell (Fig. 9b) are characteristic of the wavelet's ability to provide a variable-window framework for the localization in both time and frequency (i.e., scale), of the features contained in a signal. By comparison, the resolution cell of the windowed Fourier transform (Fig. 9a) maintains constant dimensions and thus produces inefficient representation of the signal's features.

Decomposition and Reconstruction of Functions

Let $F(t)$ be a square-integrable function, i.e., $F(t) \in L^2(t)$. The projection of $F(t)$ on a wavelet, $\psi_{su}(t)$, at dilation s and translation point, u, is given by

$$\alpha_{su} \equiv [W_{su}F(t)] = \int_{-\infty}^{+\infty} F(t)\psi_{su}(t)\,dt. \tag{5}$$

Using these projections over all wavelets with $(s,u) \in R^2$, we can reconstruct the function by the following equation:

$$F(t) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \alpha_{su}\psi_{su}(t)\,ds\,du. \tag{6}$$

If we consider a wavelet with compact support, then the projection given by Eq. (5) reflects the information content of $F(t)$ only in the resolution cell of the corresponding wavelet. For example, projections of $F(t)$ on wavelets with large dilation, s, represent the large-scale (i.e., low-frequency) components of $F(t)$. Consequently, the reconstruction of $F(t)$ through Eq. (6) indicates that we can put the functions together from "pieces," each of which represents the content of a specific segment of $F(t)$ over time [determined by $(u_0 - s\sigma_u, u_0 + s\sigma_u)$], and in terms of frequencies in a specific range [determined by $(\omega_0/s) - (\sigma_\omega/s), (\omega_0/s) + (\sigma_\omega/s)$. For example, if a function contains a sharp feature, e.g., spike, then the bulk of the informational content carried by the spike will be reflected by the projection of the function on a "narrow" wavelet, whose resolution cell is narrow in u and long in ω (e.g., cell A in Fig. 9b). On the other hand, the content of a function reflected by a slowly rising trend will be reflected by the projection of the function on a "wide" wavelet with a resolution cell such as that of cell B in Fig. 9b.

3. Discretization of Scale

When we analyze the scale-space image of a function (see Section II, B) in order to extract the trends of process variables, we are interested only in a finite number of distinct trends, as they are defined by the interval tree of scale.

Usually, the dilation parameter s is discretized along the dyadic sequence,

$$s = 2^m \qquad m \in Z,$$

where Z is the space of integer numbers. The resulting family of wavelets with a discretized dilation parameter is represented as follows:

$$\psi_{2^m u}(t) = \frac{1}{\sqrt{2^m}} \psi\left(\frac{t-u}{2^m}\right), \qquad m \in Z. \tag{2a}$$

The projection of $F(t)$ to all wavelets of the above form with $m \in Z$ and $u \in R$, yields the so-called *dyadic wavelet transform* of (t), with the following components:

$$\alpha_{2^m u} \equiv \left[W_{2^m u} F(t)\right] = \int_{-\infty}^{+\infty} F(t) \left\{ \frac{1}{\sqrt{2^m}} \psi\left(\frac{t-u}{2^m}\right) \right\} dt, \qquad m \in Z, \quad u \in R. \tag{5a}$$

The discretization of the dilation parameter need not be dyadic. Discretization schemes with integer or noninteger factors are possible and have been suggested. The reconstruction of $F(t)$ is given by

$$F(t) = \sum_{m=-\infty}^{+\infty} \int_{-\infty}^{+\infty} \alpha_{2^m u} \psi_{2^m u}(t)\, du. \tag{6a}$$

To ensure that no information is lost on $\hat{F}(\omega)$ as the dilation is discretized, the scale factors 2^m for $m \in Z$ must cover the whole frequency axis. This can be accomplished by requiring the wavelets to satisfy the following condition:

$$\sum_{m=-\infty}^{+\infty} |\hat{\psi}(2^m \omega)|^2 = 1. \tag{7}$$

Equation (6a) implies that the scale (dilation) parameter, m, is required to vary from $-\infty$ to $+\infty$. In practice, though, a process variable is measured at a finite resolution (sampling time), and only a finite number of distinct scales are of interest for the solution of engineering problems. Let $m = 0$ signify the finest temporal scale (i.e., the sampling interval at which a variable is measured) and $m = L$ be coarsest desired scale. To capture the information contained at scales $m > L$, we define a *scaling function*, $\phi(t)$, whose Fourier transform is related to that of the wavelet, $\psi(t)$, by

$$|\hat{\phi}(\omega)|^2 = \sum_{m=1}^{+\infty} |\hat{\psi}(2^m \omega)|^2. \tag{8}$$

Since $\hat{\psi}(\omega)$ must satisfy condition (3), then from Eqs. (7) and (8) we conclude that

$$\lim_{\omega \to 0} |\hat{\phi}(\omega)| = 1.$$

In other words, the energy of $\phi(t)$ can be seen as a lowpass filter.

Let us now create a family of scaling functions through the dilation and translation of $\phi(t)$

$$\phi_{2^m u}(t) = \frac{1}{\sqrt{2^m}} \phi\left(\frac{t-u}{2^m}\right), \quad m \in Z, \quad u \in R \tag{9}$$

Consequently, the projection of $F(t)$ on $\phi_{2^m u}(t)$ produces a function whose content is completely stripped of high frequencies and includes information only at scales higher than m. As m increases, more and more details are removed from $F(t)$.

Through the employment of the scaling function we can reconstruct $F(t)$ using a finite number of scales, as follows:

$$F(t) = \sum_{m=0}^{L} \int_{-\infty}^{+\infty} \alpha_{2^m u} \psi_{2^m u}(t)\, du + \int_{-\infty}^{+\infty} \beta_{2^L u} \phi_{2^L u}(t)\, du, \tag{10}$$

where, $\beta_{2^L u}$ is the projection of $F(t)$ on the scaling function, $\phi_{2^L u}$, and is given by

$$\beta_{2^L u} \equiv [S_{2^L u} F(t)] = \int_{-\infty}^{+\infty} F(t) \phi_{2^L u}(t)\, dt. \tag{11}$$

4. Dyadic Discretization of Time

Having completed the decomposition and reconstruction of a function at a finite number of discrete values of scale, let us turn our attention to the discretization of the translation parameter, u, dictated by the discrete-time character of all measured process variables. The classical approach, suggested by Meyer (1985–1986), is to discretize time over dyadic intervals, using the sampling interval, τ, as the base. Thus, the translation parameter, u, can be expressed as

$$u = k(2^m \tau), \qquad \text{with} \quad (m,k) \in Z^2, \tag{12}$$

where k is the translation parameter. At the initial dilation level $m = 0$, and $u = k\tau$, $k = 0, 1, 2, \ldots$, i.e., the translation points coincide with the sampling instants, i.e., $0, \tau, 2\tau, \ldots$. At the next dilation level, $u = (2\tau)k$, i.e., the distance between two adjacent wavelets double, and the translation points are $0, 2\tau, 4\tau, \ldots$.

As a result of the dyadic discretization in dilation and translation, the members of the wavelet family are given by

$$\psi_{mk}(t) \equiv \frac{1}{\sqrt{2^m}}\psi\left(\frac{t}{2^m} - k\tau\right), \qquad (m,k) \in Z^2.$$

A discrete-time signal is decomposed into the following set of projections:

$$\alpha_{mk} \equiv \left[W_{2^m, k(2^m \tau)}F(t)\right] = \sum_{k=-\infty}^{+\infty} F(t)\psi_{mk}, \qquad 1 \leq m \leq L, \tag{5b}$$

$$\beta_{Lk} \equiv \left[S_{2^L, k(2^L \tau)}F(t)\right] = \sum_{k=-\infty}^{+\infty} F(t)\phi_{Lk}, \tag{11a}$$

from which it can be completely reconstructed:

$$F(t) = \sum_{m=1}^{L} \sum_{k=-\infty}^{+\infty} \alpha_{mk}\psi_{mk} + \sum_{k=-\infty}^{+\infty} \beta_{Lk}\phi_{Lk}. \tag{6b}$$

For practical purposes, the wavelet decomposition can only be applied to a finite record of discrete-time signals. If N is the number of samples in the record, and $\tau = 1$, then the maximum value of the translation parameter can be found from Eq. (12), by setting $u = N$, and is equal to $k_{\max} = N/2^m$. Consequently, the decomposition and reconstruction relations [Eqs. (5b), (11a), (6b)] take the following form:

$$\alpha_{mk} \equiv \left[W_{2^m, k2^m}F(t)\right] = \sum_{k=1}^{k_{\max}} F(t)\psi_{mk}, \qquad 1 \leq m \leq L, \tag{13}$$

$$\beta_{Lk} \equiv S_{2^m, k2^m}F(t) = \sum_{k=1}^{k_{\max}} F(t)\phi_{Lk}, \tag{14}$$

$$F(t) = \sum_{m=1}^{L} \sum_{k=1}^{k_{\max}} \alpha_{mk}\psi_{mk} + \sum_{k=1}^{k_{\max}} \beta_{Lk}\phi_{Lk}. \tag{15}$$

Equation (15) provides a discrete, complete and nonredundant representation of the function $F(t)$, since it requires the computation of N coefficients, if N is the number of discrete-time samples describing $F(t)$.

5. *Uniform Discretization of Time*

An alternative approach to the discretion of the translation parameter u involves uniform sampling of the measured signal at all scales, i.e., $u = k\tau$, with $k \in Z$. The resulting decomposition algorithm is of complexity $O(N \log N)$, and the associated reconstruction requires the computation of $N \log N$ coefficients, i.e., it contains redundant information.

Nevertheless, uniform discretization of time at all scales leads to representations that are highly suitable for feature extraction and pattern recognition. More on this subject in a subsequent paragraph.

6. Practical Considerations

Let us now see how the theory of the wavelet-based decomposition and reconstruction of discrete-time functions can be converted into an efficient numerical algorithm for the multiscale analysis of signals. From Eq. (6b) it is easy to see that, given a discrete-time signal, $F_0(t)$ we have

$$F_0(t) = D_1(t) + F_1(t), \tag{16}$$

where

$$F_1(t) = \sum_{k=-\infty}^{+\infty} \beta_{1k}\phi_{1k}(t) \quad \text{and} \quad D_1(t) = \sum_{k=-\infty}^{+\infty} \alpha_{1k}\psi_{1k}(t). \tag{17}$$

$F_1(t)$ is called the *scaled signal* and is derived from the filtering of $F_0(t)$ with the lowpass scaling function. It represents a smoother version of $F_0(t)$. $D_1(t)$ is called the *detail signal* and is derived from the filtering of $F_0(t)$ with the bandpass wavelet functions. It represents the information that was filtered out of $F_0(t)$ in producing $F_1(t)$.

Generalizing Eqs. (16) and (17), we take

$$F_m(t) = D_{m+1}(t) + F_{m+1}(t), \qquad m \in Z, \tag{18}$$

with

$$F_{m+1}(t) = F_m(t)^*H_m \quad \text{and} \quad D_{m+1}(t) = F_m(t)^*G_m, \tag{19}$$

where the operator, *, signifies the convolution operation. Filter H_m is a lowpass filter, emulating the effects of the scaling function, and G_m is a bandpass filter affecting the influence of the wavelet function. Equations (18) and (19) imply that the wavelet-based decomposition and reconstruction of a discrete-time signal can be carried out through a cascade of convolutions with filters H and G (see Fig. 10). Figure 11 shows the scaled and detail signals generated from the wavelet decomposition of a given signal using dyadic and uniform sampling.. The detailed methodological aspects of the cascaded convolution of signals, described above, can be found in Mallat (1989) for the case of dyadic sampling of time and in Mallat and Zhong (1992) for uniform sampling.

For real-world signals with finite records, the convolution processes leading to decomposition and reconstruction of a signal, require data in regions beyond the signal's endpoints. Assuming a mirror image of the

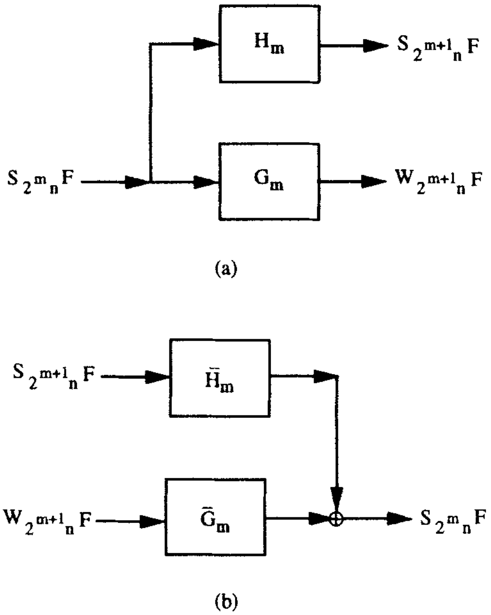


FIG. 10. Methodology for multiscale (a) decomposition and (b) reconstruction, using wavelets, with uniform sampling $(m, n) \in Z^2$.

original signal beyond both ends (Mallat, 1989), causes numerical inaccuracies in the computation of the wavelet coefficients. Bakshi and Stephanopoulos (1995) have studied various approaches for handling end-point effects and they have suggested certain practical solutions in this problem.

B. EXTRACTION OF MULTISCALE TEMPORAL TRENDS

In Section II we defined the trend of a measured variable as a strictly ordered sequence of scaling episodes. Since each scaling episode is defined by its bounding inflexion points, it is clear that the extraction of trends necessitates the localization of inflexion points of the measured variable at various scales of the scale-space image. Finally, the interval tree of scale (see Section II) indicates that there is a finite number of distinct sequences of inflexion points, implying a finite number of distinct trends. The question that we will try to answer in this section is, "How can you use the wavelet-based decomposition of signals in order to identify the distinct sequences of inflexion points and thus of the signal's trends?"

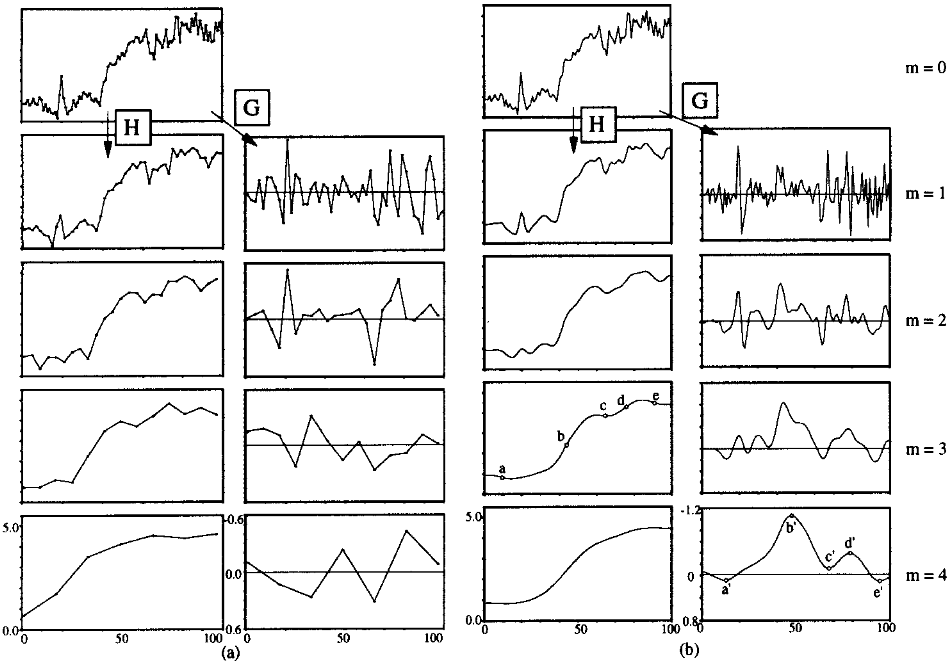


FIG. 11. Wavelet decomposition (a) dyadic sampling using Daubechies-6 wavelet; (b) uniform sampling using cubic spline wavelet.

1. Translationally Invariant Representation of Measured Variables

Measured variables may have identical features in different segments of the time record. Any analysis method should be able to identify and extract such common features, independently of the time segment in which they appear. Consequently, the wavelet decomposition of a time-translated signal should produce wavelet coefficients that are translated in time but equal in magnitude to the wavelet coefficients of the original (untranslated) signal. For example, if $F(t) \in L^2$ is the original signal and $F_\theta(t)$ is the same function translated in time by θ units, i.e., $F_\theta(t) = F(t - \theta)$, then

$$[W_{2^m u} F_\theta(t)] = [W_{2^m u} F(t-\theta)], \qquad m \in Z, \quad u \in R. \tag{20}$$

Since the translation variable, u, is also discretized (dyadically, or uniformly at each scale), relationship (20) will hold only if the signal is translated in time by a period that is an integer multiple of the discretiza-

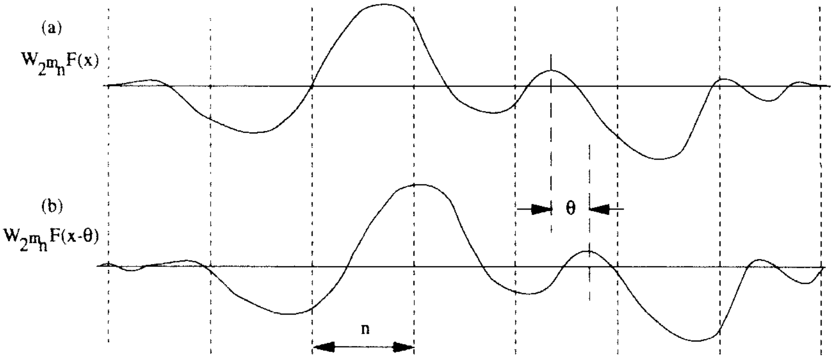


FIG. 12. Translation of the wavelet transform of $F(x)$ by θ, with sampling rate, n (Mallat, 1991).

tion interval at all scales. For example, in the case of dyadic sampling of time, translational invariance of the wavelet transform requires that $\theta = r(2^m\tau)$, $r \in Z$, whereas uniform sampling necessitates that $\theta = r\tau$, $r \in Z$, where τ is the sampling period of the original signal.

Unfortunately, the requirements for translational invariance of the wavelet decomposition are difficult to satisfy. Consequently, for either discretization scheme, comparison of the wavelet coefficients for two signals may mislead us into thinking that the two trends are different, when in fact one is simply a translation of the other.

On the other hand, the value of (zero-order) zero crossings and local extrema (first-order zero crossings) of a dyadic wavelet transform, $W_{2^m u}F(t)$, do translate when the original signal is translated (Fig. 12). Therefore, we can generate translationally invariant representations of a signal if we employ zero-order and first-order zero crossings of the signal's wavelet transform.

2. *Detection of Inflexion Points*

Within the framework of scale-space filtering, inflexion points of $F(t)$ appear as extrema in $\partial F(t)/\partial t$ and zero crossings in $\partial^2 F(t)/\partial t^2$. Thus, filtering a signal by the Laplacian (second derivative) of a Gaussian will generate the inflexion points at various scales (Marr and Hildreth, 1980). In the same spirit, if the wavelet is chosen to be the first derivative of a scaling function, i.e., $\psi(t) = d\phi(t)/dt$, then from Eqs. (5a) and (11) we can easily show (Mallat and Zhong, 1992) that

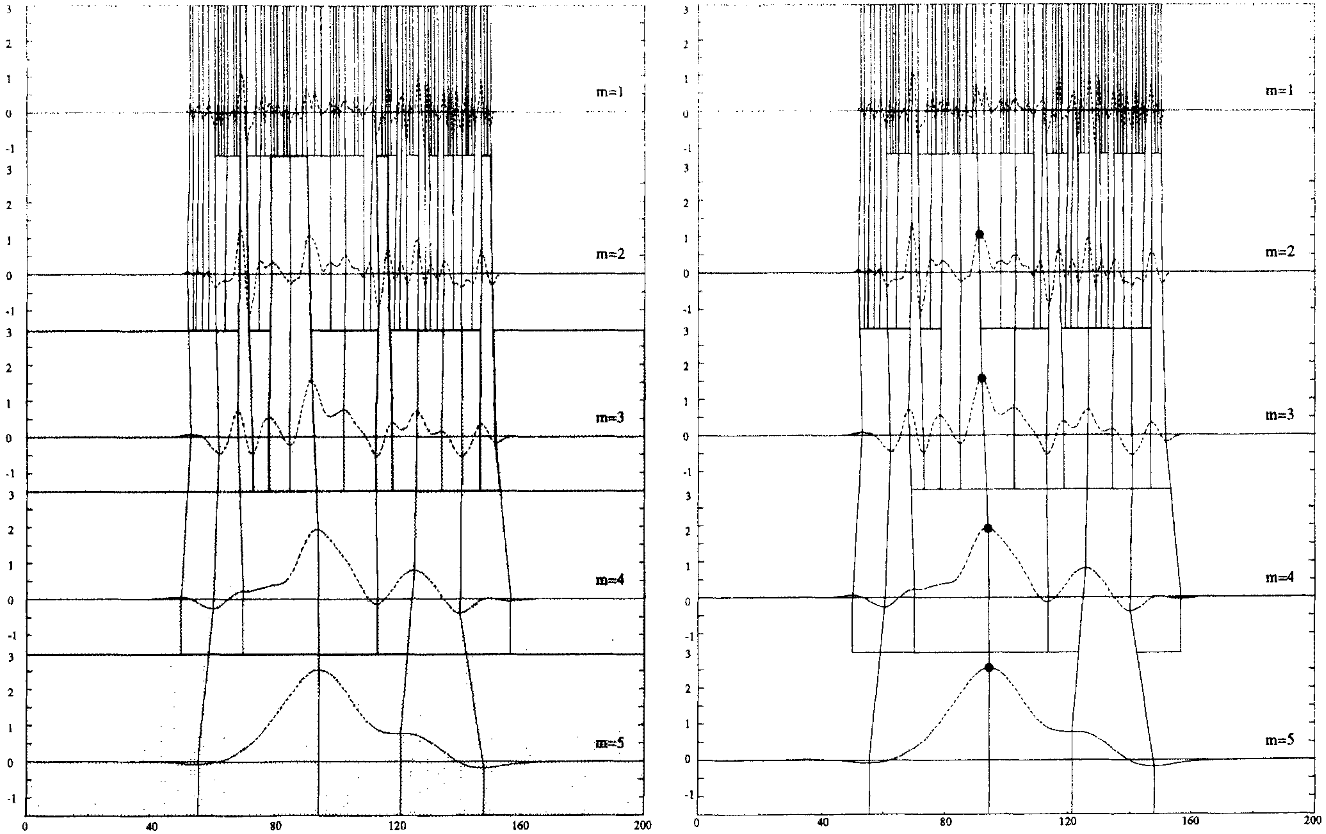
$$[W_{2^m u}F(t)] = 2^m \frac{d}{dt}[S_{2^m u}F(t)]. \tag{21}$$

A wavelet defined as above is called a *first-order wavelet*. From Eq. (21) we conclude that the extrema points of the first-order wavelet transform provide the position of the inflexion points of the scaled signal at any level of scale. Similarly, if $\psi(t) = d^2\phi(t)/dt^2$, then the zero crossings of the wavelet transform correspond to the inflexion points of the original signal smoothed (i.e., scaled) by the scaling function, $\psi(t)$ (Mallat, 1991).

An obvious choice for the scaling function is the Gaussian, but it is not a compactly supported function, and the corresponding filter requires a large number of coefficients. The scaling function and associated wavelet chosen in this chapter to support the extraction of process trends were suggested by Mallat and Zhong (1992). The scaling function is a cubic spline with compact support of size 4 (i.e., the corresponding filter requires four coefficients), whereas the associated wavelet is not compactly supported but decreases exponentially at infinity (Fig. 8c). Using these functions, we can generate efficient representations of signals. The scaled signals at various scales are equivalent to the cubic spline fits of the data at the various scales. Unfortunately, since the scaling function is not a Gaussian, we cannot guarantee that spurious oscillations with fictitious inflexion points will not be generated at coarser scales. Nevertheless, such fictitious inflexion points, if generated, are short-lived, i.e., disappear quickly with increasing scale, and thus they will never be stable and distinguished features of any process trend. It is also worth noting that the variation in the magnitude of local extrema of the wavelet decomposition across several scales provides useful information about the nature of the inflexion points, and thus offers a measure of the signal's regularity.

3. *The Wavelet Interval–Tree of Scale*

Consider the given signal and its wavelet decomposition, as shown in Fig. 11. Since the wavelet is a first-order wavelet, the inflexion points of the original function appear as extrema of the detail signal at the corresponding scale (Fig. 11). If we connect the extrema of the detailed signal across various scales, we generate a structure similar to the interval tree of scale, discussed in Section II, B. We will call this tree, the *wavelet interval–tree of scale* (Fig. 13). It gives the evolution of the features of a signal in the scale space, and can be used to generate the trends of measured variables with various combinations of features. The wavelet



(a)

(b)

FIG. 13. Wavelet interval–tree of scales for the signal of Fig. 1; (a) extrema in shaded region reconstruct stable trend at $m = 3$ (see Fig.

interval–tree of scale presents certain advantages over the classical scale-space filtering, which are very important in the reconstruction of process trends. These advantages are (Bakshi and Stephanopoulos, 1994a):

1. The wavelet interval–tree of scale is constructed from $\log_2 N$ distinct representations, where N is the number of points in the record of measured data. This is a far more efficient representation than that of scale-space filtering with continuous variation of Gaussian σ.
2. In scale-space filtering, trends consisting of combinations of features at various scales are constructed by approximating the stable features through piecewise continuous segments, like parabolas. Such an approach leads to discontinuities in the trends. On the contrary, the trends generated from the wavelet interval-tree of scale with a mixture of features from different scales, are continuous up to the degree of continuity of the scaling function, and are based on formal and sound analysis.
3. The detail signals at the various scales of the tree provide the features of the trend that are distinct to the corresponding range of scales.
4. The evolution of the inflexion points, as given by the local extrema of the detail signals, also characterize the regularity of the original signal's inflexion points.

4. The Algorithm for the Extraction of Trends

Let us recall that *a trend is a strictly ordered sequence of scaling episodes*, whose bounding inflexion points can be identified by the local extrema of the detail signals, as discussed in the previous paragraph. Clearly, the more stable a scaling episode (over a range of scales), the more distinguished is the corresponding feature in describing the trend. Consider, for example, the noise-corrupted pulse shown in Fig. 14 along with the associated detail signals, resulting from its decomposition with a first-order wavelet. The extrema, A and B, of the detail signal at the smallest scale persist at all scales, indicating the presence of a very stable scaling episode. This scaling episode is bounded by the inflexion points on the left and right edges of the pulse and reveals the most dominant feature of the original signal, i.e., the pulse itself. On the contrary, the inflexion points corresponding to the fluctuations of the noise in the original signal disappear very quickly and do not represent distinguished scaling episodes, i.e., features. The range of each extremum point is equal to the range of the G filter and is equal to $2^{m+1}q+1$, where m is the scale of the corresponding detail signal and $2q+1$ is the length of the filter. By

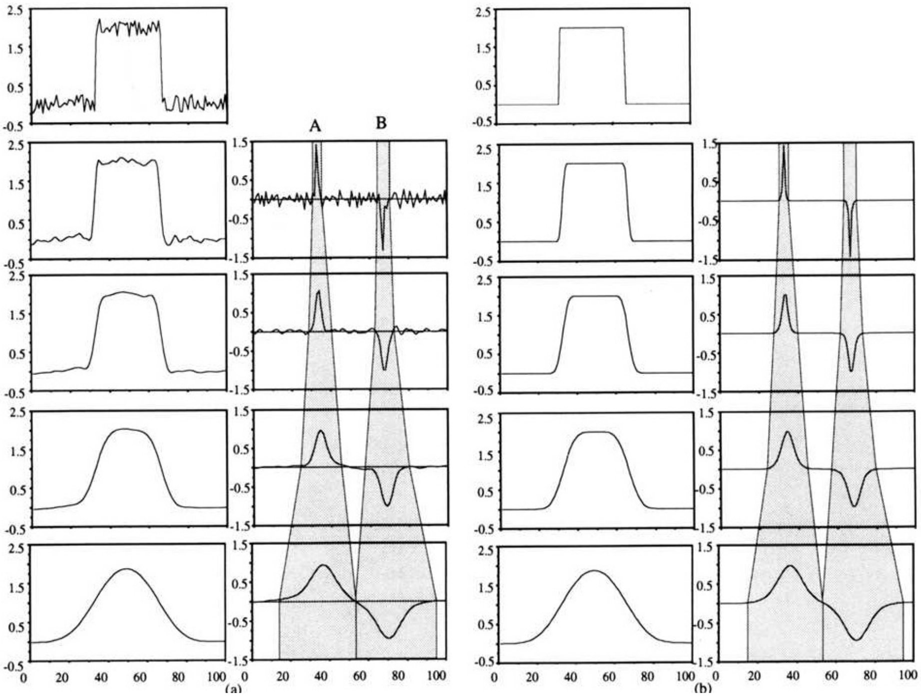


FIG. 14. Extracting distinguishing features from noise pulse signal. Wavelet coefficients in shaded regions represent stable extrema. (a) Wavelet decomposition of noisy pulse signal; (b) wavelet decomposition of pulse signal. (Reprinted from Bakshi and Stephanopoulos, "Representation of process trends, Part III. *Computers and Chemical Engineering*, **18**(4), p. 267, Copyright (1994), with kind permission from Elsevier Science Ltd., The Boulevard, Langford Lane, Kidlington OX5 1GB, UK.)

utilizing the wavelet coefficients of the stable extrema we can reconstruct a signal that contains the distinguished features at several scales. In Fig. 14, by utilizing the wavelet coefficients of the shaded region, we can completely reconstruct the pulse with minimum quantitative distortion.

The algorithm for extracting the most stable description of trends at each level of scale, proceeds as follows:

Step 1. Generate the finite, discrete dyadic wavelet transform of data using Mallat and Zhong's (1992) cubic spline wavelet (Fig. 8c).

Step 2. Generate the wavelet interval-tree of scale.

Step 3. For each scaling level, m, and for each scaling episode, (a) determine the range of stability and (b) collect the wavelet coefficients at all scales in the range of the episode's stability.

Step 4. Reconstruct the signal using the last scaled signal and the wavelet coefficients collected in step 3.

Once the stable reconstruction of a signal has been accomplished, its subsequent representation can be made at any level of detail, i.e. qualitative, semi-quantitative, or fully real-valued quantitative. The triangular episodes (described in Section I, A) can be constructed to offer an explicit, declarative description of process trends.

5. *Illustrative Examples*

a. Example 1: Generating Multiscale Descriptions. Consider the process signal and its wavelet decomposition shown in Fig. 11. Quantitatively accurate representations at each of the dyadic scales may be generated by selecting the wavelet transform extrema for the most stable episodes at each scale. For example, the extrema selected for reconstructing the process trend at scale, $m = 3$, are shown shaded in Fig. 13a. Note that wavelet coefficients at scales < 3 are needed to provide a stable representation. The reconstructed signal is shown in Fig. 15b, and is qualitatively equivalent to the scaled signal at $m = 2$, but is quantitatively much more accurate, since features such as the spike at $t = 20$ are undistorted. Similarly, the stable process trend at scale $m = 4$ is shown in Fig. 15c. The step change at $t = 40$ is not brought out accurately in the trends in Figs. 15b and 15c, due to the presence of other extrema in the range of the extremum corresponding to the step change. Most of the extrema in its range are much smaller than the extremum at $t = 40$. An empirical criterion, based on the relative magnitudes of wavelet transform extrema may be used for selecting large extrema, resulting in the reconstructed signal shown in Fig. 15d. The extremum at $t \approx 40$, at scales $m = 2$–5 used for this reconstruction, as shown in Fig. 13b. The step change is brought out quite accurately in this trend at $m = 5$.

Another powerful heuristic for extracting process trends is Witkin's stability criterion (Witkin, 1983). Starting with the episodes at the desired scale, lower-scale descriptions are considered only if the lower-scale episodes have a higher mean stability than their parent episode. The episodes selected using this criterion at scale, $m = 5$, and the reconstructed signal are shown in Fig. 16. The reconstructed signal contains all the conspicuous features in the raw data, without quantitative distortion. Witkin's stability criterion is very effective for extracting the most relevant features from a process signal.

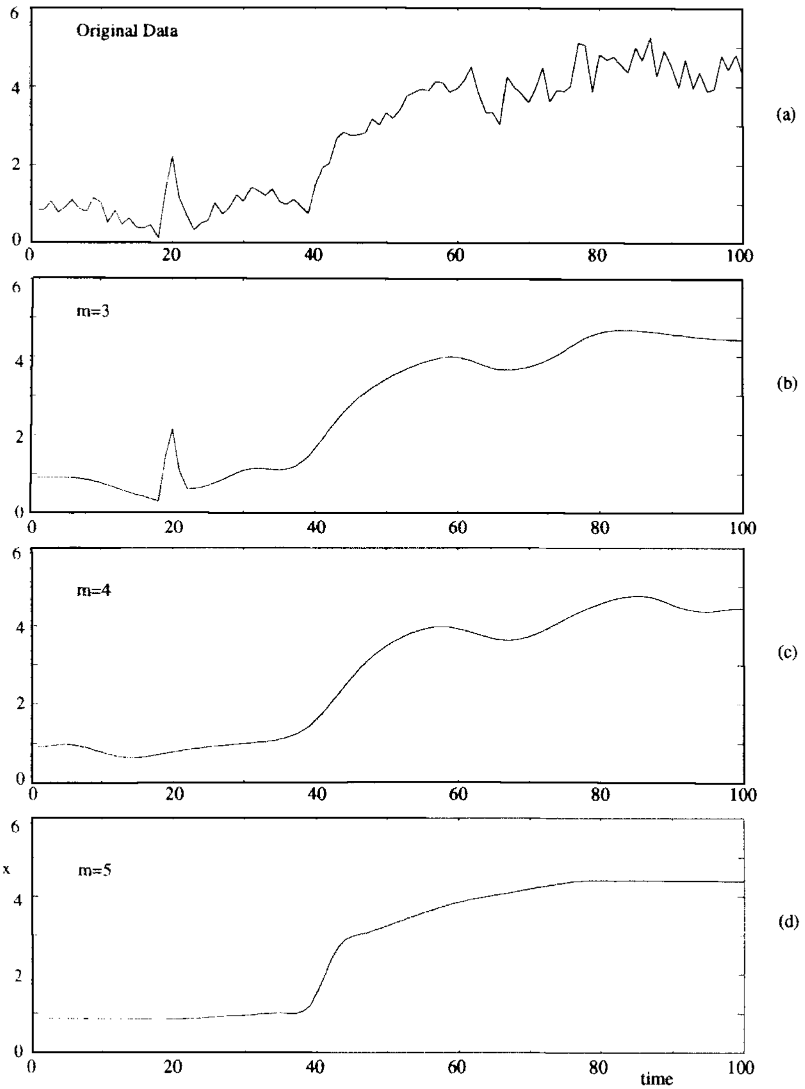


FIG. 15. Process trends of the signal in Fig. 1, extracted at various scales: (a) original data; (b) stable trend at $m = 3$ (see Fig. 13); (c) stable trend at $m = 4$; (d) stable trend at $m = 5$, neglecting small extrema.

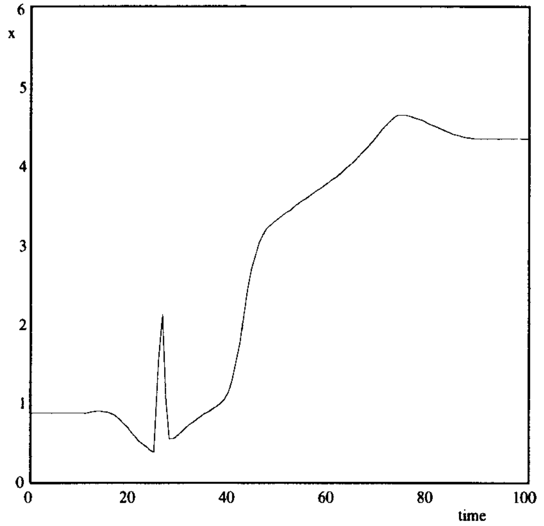


FIG. 16. Stable trend of the signal in Fig. 1, generated through the use of Witkin's stability criterion, and the wavelet coefficients shown in Fig. 13b.

b. Example 2. Generalization of process trends. Comparison of process signals obtained from several examples representing different process conditions is essential for evaluating the relevance of the extracted features. The representation of process trends at multiple scales provides a convenient, hierarchical technique for evaluating the features. Consider the signals shown in Fig. 17a obtained from three different batches of the fed-batch fermentation process described in Section V. At the level of the raw data, it is difficult to compare the three signals. On representing the signal at multiple scales, the differences in the signals disappear, and at a coarse-enough scale, the signals become qualitatively identical, as shown in Figs. 17–17g. The discovery of a generalized description provides a means of comparing the qualitative and quantitative features contained in the signals, and allows matching of features in the trends, facilitating easy extraction of qualitative differences. If the information in the signals at coarse scales is inadequate for distinguishing between the signals, then information in trends at finer scales is considered. As lower scales are considered, only the matched features need to be compared with each other. This provides a natural decomposition to the learning problem, and simplifies it significantly. The utility of generalized descriptions of process signals will be exploited further in the learning of input/ output mappings, as described in Section V. The process trends obtained by applying Witkin's stability criterion at the coarsest scale are shown in Fig. 17h. These process trends are also quantitatively very similar, and the extracted features are physically meaningful and thus interpretable, as we will see in the illustrations of Section V (Figs. 21–23).

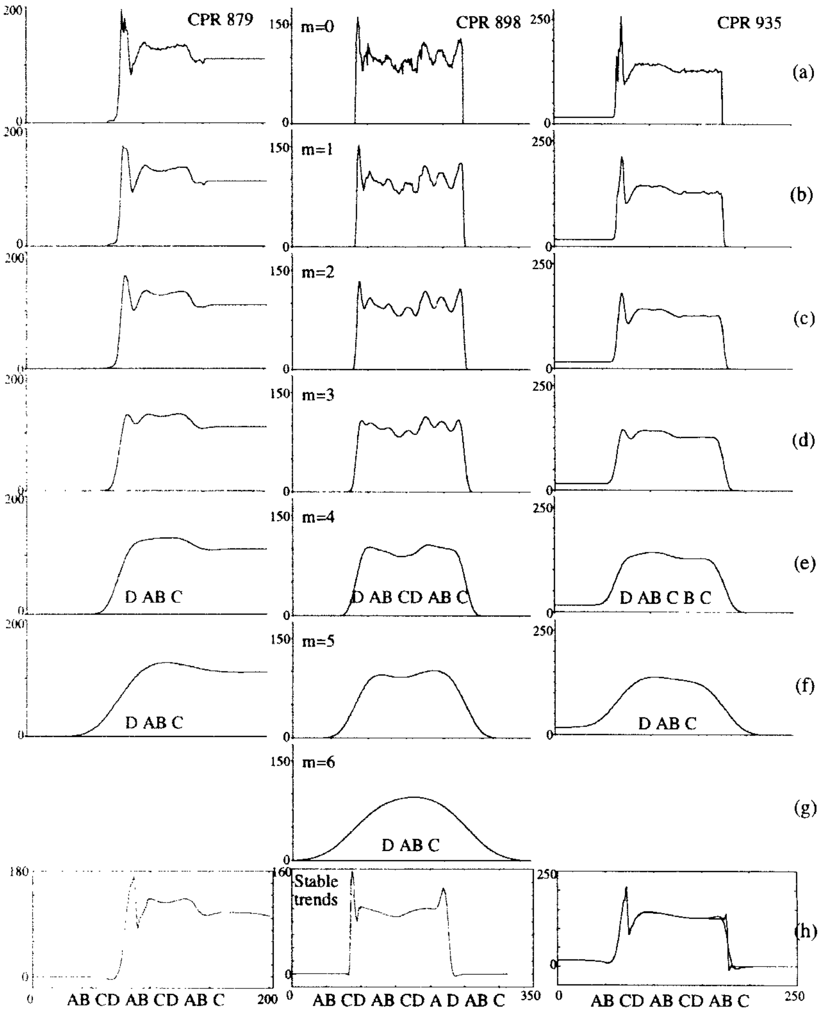


FIG. 17. Generalization of process trends for three distinct records; (a) raw data; (b)–(g) scaled signals; (h) stable trends.

IV. Compression of Process Data through Feature Extraction and Functional Approximation

With increasing computerization and improvements in sensor technology, it is relatively easy and inexpensive to collect large quantities of process data. Since measured data are useful for performing a variety of analytical and decision support engineering tasks, it is essential to store the data in historical records for future use. Efficient storage techniques are needed for two primary reasons: (1) to reduce the space required for the historical records and (2) to retrieve the data in a manner than renders the data easily interpretable for the execution of engineering tasks. In this section we will examine how both of these needs can be satisfied within the same theoretical framework of wavelet decomposition for the representation of measured signals.

A. Data Compression through Orthonormal Wavelets

Data compaction involves representation of the measured signal as an approximation that requires less storage space, at the cost of losing some information from the original signal. The data compaction problem may be stated as an approximation problem (Bakshi and Stephanopoulos, 1995) as follows:

1. Definition: Data Compaction Problem

Determine the approximate representation $\tilde{F}$ of a discrete-time function, $F(t)$ so as to either (1) minimize the error of the approximation given by

$$e_{L_p} = \frac{\|F - \tilde{F}\|_{L_p(I)}}{N} = \frac{\left(\int_I |F(t) - \tilde{F}(t)|^p \, dt\right)^{1/p}}{N}, \qquad 0 < p < \infty,$$

where N is the number of data points contained in the original signal, for a given compression ratio, $C_{\mathrm{R}} = (N/\tilde{N})$, where $\tilde{N}$ is the number of points stored in the compacted representation; or (2) maximize the compression rate, C_{R}, for a given error of approximation, $e_{L_p}^2$.

A popular technique for approximation consists of representing the data as a weighted sum of a set of basis functions. As described in Section III, A, wavelets form a convenient set of basis functions to represent signals consisting of a variety of features. A signal decomposed on an orthonormal basis, using dyadic discretization of the translation and dilation parameters may be represented in terms of its wavelet coefficients and the coefficients of the last scaled signal, as given by Eq. (15). Since the wavelets at all translations and dilations, and scaling functions at a given dilation are orthonormal, the total energy in the signal is equal to the sum of the square of the coefficients:

$$\|F(t)\|_{L_2}^2 = \sum_{m=1}^{L} \sum_{k=1}^{k_{\max}} \alpha_{mk}^2 + \sum_{k=1}^{k_{\max}} \beta_{Lk}^2 .$$

Compression may be achieved if some regions of the time–frequency space in which the data are decomposed do not contain much information. The square of each wavelet coefficient is proportional to the least-squares error of approximation incurred by neglecting that coefficient in the reconstruction;

$$e_N^2 = \frac{\|F(t) - \tilde{F}(t)\|^2}{N^2} = \frac{\sum_{m=1}^{L} \sum_{k=1}^{k_{\max}} \alpha_{mk,\text{neglected}}^2 + \sum_{k=1}^{k_{\max}} \beta_{Lk,\text{neglected}}^2}{N^2}, \tag{22}$$

where $\alpha_{mk,\text{neglected}}$ and $\beta_{mk,\text{neglected}}$ are the coefficients that are not stored in the compacted representation. Similarly, the local error in a region consisting of $(2r + 1)$ points, bounded by the interval $[l - r, l + r]$ is given by

$$e_{2r+1}^2 = \frac{\|F(t) - \tilde{F}(t)\|^2}{(2r + 1)^2} = \frac{\sum_{m=1}^{L} \sum_{k=l-r}^{l+r} \alpha_{mk,\text{neglected}}^2 + \sum_{k=l-r}^{l+r} \beta_{Lk,\text{neglected}}^2}{(2r + 1)^2}. \tag{23}$$

As $r \to 0$, the error of approximation tends to the L^∞ norm. Equations (22) and (23) are useful for data compaction with orthonormal wavelets.

Several complete orthonormal bases may be used for data compression. The selection of the best basis may be performed by utilizing several different criteria suggested by Coifman, Wickerhauser, and coworkers (Coifman and Wickerhauser, 1992; Wickerhauser, 1991). Some of the most interesting basis selection criteria include those discussed in the following paragraphs.

a. Number above a Threshold. The set of wavelets with the minimum number of coefficients above a threshold, ε, is selected as the best basis. This gives the best basis to represent a signal to precision ε. This measure is similar in principle to the error measure for the box car and backward slope methods.

b. Entropy. The statistical thermodynamical entropy is given by

$$H(\alpha) = -\sum_j p_j \log p_j,$$

where

$$p_j = \frac{|\alpha_j|^2}{\|\alpha_j\|^2} \quad \text{and} \quad p \log p = 0 \quad \text{for} \quad p = 0.$$

Coefficients of the wavelet expansion are indicated by α_j. Entropy is not an additive cost function, but it can be written as

$$H(\alpha) = \|\alpha\|^{-2}\lambda(\alpha) + \log\|\alpha\|^2,$$

where

$$\lambda(\alpha) = -\sum_j |\alpha_j|^2 \log|\alpha_j|^2$$

is additive, and can be minimized. The entropy measure provides a physically meaningful criterion for selecting the appropriate coefficients for data compression because $\exp[H(\alpha)]$ is proportional to the number of coefficients needed to represent the signal to a fixed mean square error.

c. Number Capturing Given Percentage of Signal Energy. This measure explicitly selects the smallest number of coefficients necessary to represent the signal with a given least-squares error. The cost function is the number of coefficients, N_e, with the largest absolute value that capture a percentage, e, of signal energy. These N_e coefficients represent the signal with the least-squares error, e, in the local region covered by the wavelet in the most compact form:

$$\sum_{j=1}^{N_e} \alpha_j^2 \geq \frac{e}{100} \sum_{j=1}^{N} \alpha_j^2,$$

where N is the total number of coefficients in the given wavelet decomposition. This measure evaluates the number of coefficients necessary to approximate the signal with a given least-squares error of approximation.

B. Compression through Feature Extraction

Compression of process data through feature extraction requires

1. Techniques for representing features in the process data in an explicit manner to allow selection of the relevant features.
2. Criteria for determining what features in a process signal are relevant and worth storing, and what features may be lost due to compression.

Multiresolution representation of a process signal allows satisfaction of both these requirements. Decomposition of a signal using derivative wavelets and uniform sampling of the translation parameter provides a technique for representing the dominant features in a signal at various scales, as discussed in Section III, B. The relevance of features in the process signal may be determined based on their persistence over multiple scales. These properties of derivative wavelets are based on the work of Mallat and Zhong (1992), and have been exploited for extracting features from process data by Bakshi and Stephanopoulos (1994a) and were discussed in Section III, B.

C. Practical Issues in Data Compression

The practical, implementational issues that arise during the utilization of the data compression techniques, presented in the previous two subsections, are discussed in the following paragraphs.

1. Compression in Real Time

The speed with which the data need to be compressed depends on the stage of data acquisition at which compression is desired. In intelligent sensors it may be necessary to do some preliminary data compression as the data are collected. Often data are collected for several days or weeks without any compression, and then stored into the company data archives. These data may be retrieved at a later stage for studying various aspects of the process operation.

In order to compress the measured data through a wavelet-based technique, it is necessary to perform a series of convolutions on the data. Because of the finite size of the convolution filters, the data may be decomposed only after enough data has been collected so as to allow convolution and decomposition on a wavelet basis. Therefore, point-by-point data compression as done by the boxcar or backward slope methods is not possible using wavelets. Usually, a window of data of length 2^m, $m \in Z$, is collected before decomposition and selection of the appropriate features for storage are carried out. The optimal window size depends on the nature of the signal, the compression parameter used, and the decomposing wavelet. The window size should be greater than the duration of the longest discardable or irrelevant even in the signal. For a measured variable, the best basis may be different for compressing data in different windows. Usually, the time–frequency characteristics of the signal from a given measured variable do not vary significantly with time. Therefore, a best basis may be selected from a few sets of data and then used for new data. The feasibility of the selected best basis may be evaluated at regular intervals, if necessary.

2. *Inaccuracies Due to End Effects*

A practical problem in decomposing and reconstructing signals using wavelets is the errors introduced due to boundary effects. Both wavelet decomposition and reconstruction with limited amounts of data require assumptions about the signal's behavior beyond its endpoints. Usually, the signal is assumed to have its mirror image beyond its boundaries. This assumption often introduces an unacceptable error in the reconstructed signal, particularly near its endpoints. For processes, where the signal is compressed by analyzing windows of finite sizes, this error may be eliminated by augmenting the signal beyond its endpoints by segments of constant value. Thus, for a signal of length 128 points, the decomposition is performed on a signal that is augmented by constant segments of length equal to 64 points on either end. The resulting decomposition has $L + 1$, i.e., one additional scale that is created because of the augmentation. For data compression, this additional scale is disregarded, and the scaling function and wavelet coefficients are selected from L scales only. This simple procedure results in highly accurate reconstruction by elimination of the boundary effects. The quality of the compression is also unaffected since the wavelet coefficients for the augmented portions are constant or zero. The computational complexity of decomposition and reconstruction of the augmented signal is $O(2N)$, and still linear in the length of the original signal.

3. *Selecting the Mother Wavelet and Compression Criteria*

Several types of wavelets have been developed, but no formal criteria exist for selecting the mother wavelet for compressing a given signal. Some qualitative criteria and experience-based heuristics are normally used. In approximating a signal by wavelet or derivative wavelet transform extrema, the smoothness of the reconstructed signal depends directly on the nature of the basis functions. Several orthonormal wavelets with different degrees of smoothness have been designed. The orthonormal wavelets with compact support designed by Daubechies (1988) are reasonably smooth for orders greater than 6. The Haar wavelets provide piecewise constant approximation which may be adequate for some process signals, such as those of manipulated variables. Among derivative wavelets, quadratic and cubic wavelets are described by Mallat and Zhong (1992). The first derivative of a Gaussian is an infinitely differentiable wavelet. A priori knowledge of the nature of the signal may be used to select the mother wavelet based on its smoothness.

The accuracy of the error equations (Eqs. (22) and (23)] also depends on the selected wavelet. A short and compactly supported wavelet such as the Haar wavelet provides the most accurate satisfaction of the error estimate. For longer wavelets, numerical inaccuracies are introduced in the error equations due to end effects. For wavelets that are not compactly supported, such as the Battle–Lemarie family of wavelets, the truncation of the filters contributes to the error of approximation in the reconstructed signal, resulting in a lower compression ratio for the same approximation error.

D. An Illustrative Example

A typical example of half a day's process data from a distillation tower and its wavelet decomposition using an orthonormal wavelet (Daubechies, 1988), with dyadic sampling are shown in Fig. 18. The raw process data represent pressure variation measured every minute (Takei, 1991). The raw signal is augmented by a constant segment of half the length of the original signal to avoid boundary effects. The augmentation results in an additional level of the decomposition, which is disregarded in the selection of coefficients, and the signal reconstruction. The wavelet coefficients are normalized to be proportional to their contribution to the overall signal. From Fig. 18 it is clear that many of the wavelet coefficients have very small values, which may be neglected without significant loss of information. The reconstructed signal with compression ratios of 2, 4, and 24 are shown in Fig. 19. The performance of orthonormal wavelet-based data compaction is compared with that of the conventional techniques such as backward slope and boxcar methods. As shown in Fig. 20, the wavelet-based method outperforms both the boxcar, and the backward slope methods. The quality of the reconstructed signal is significantly better for the wavelet-based method for similar compression ratios. Process data compaction using biorthogonal wavelets, wavelet packets and wavelet trans-

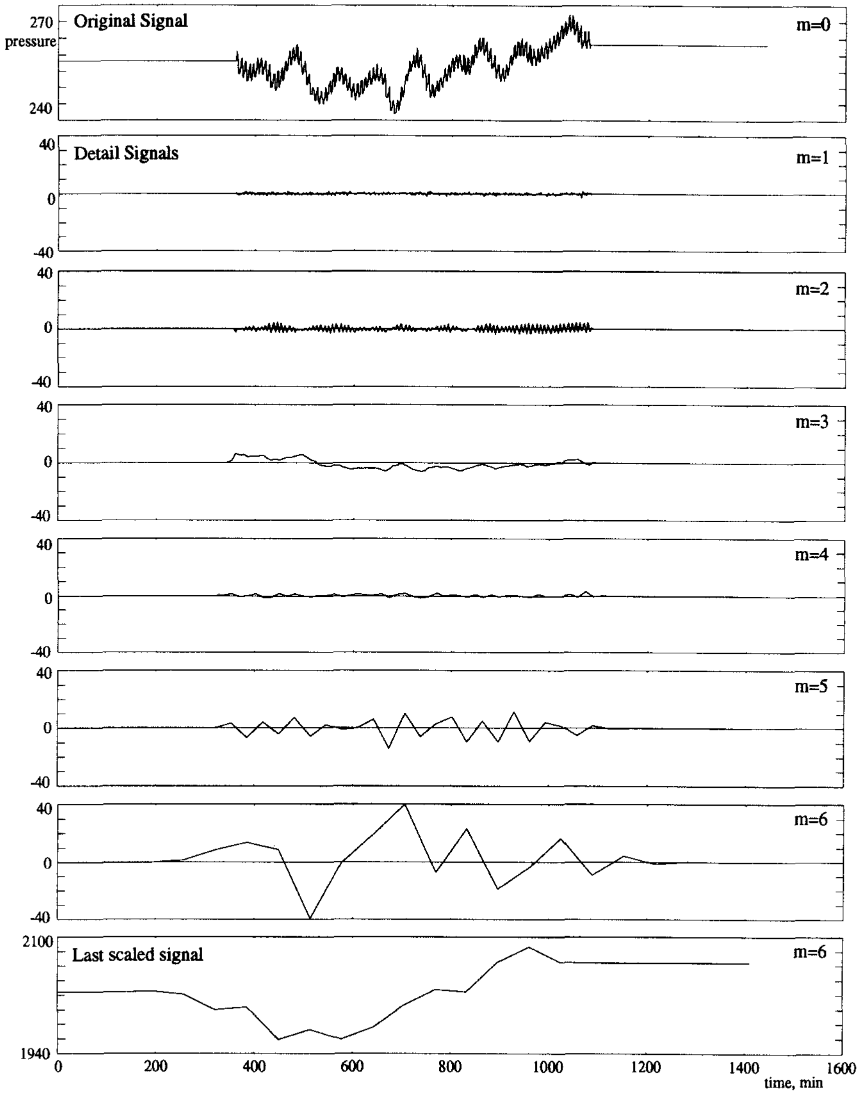


FIG. 18. Wavelet decomposition of a pressure signal, using Daubechies-6 wavelet.

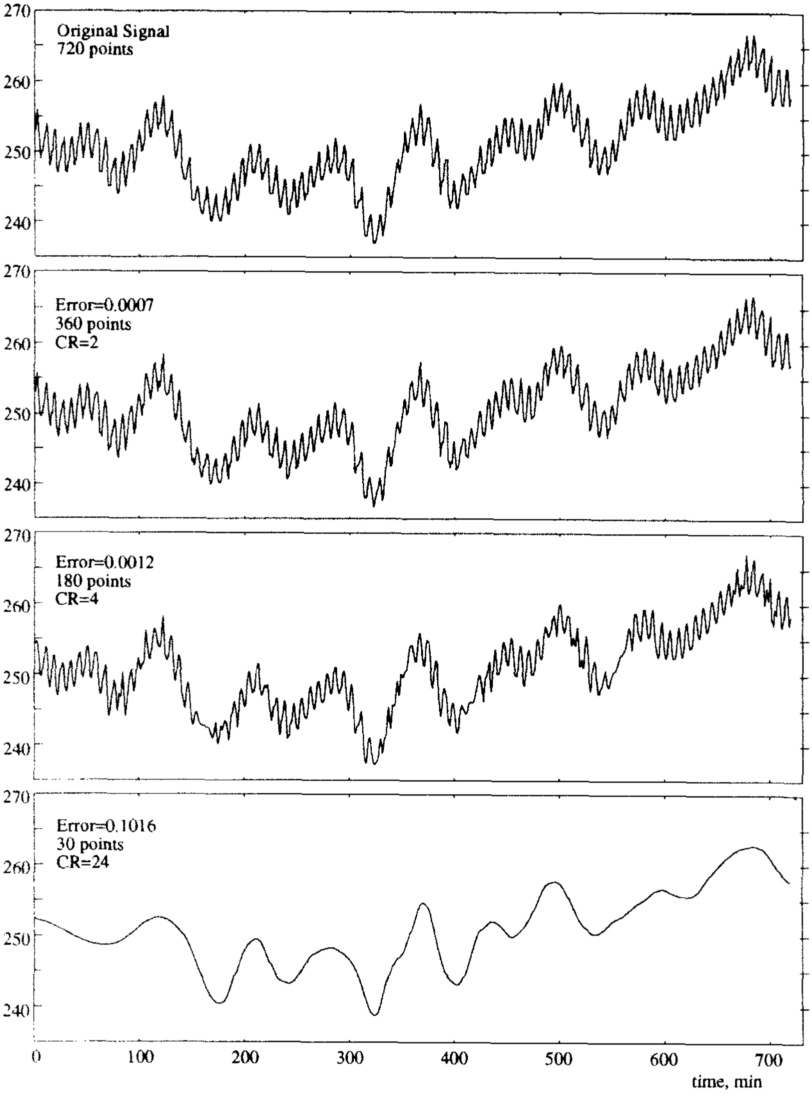


FIG. 19. Reconstruction of compressed signal from the wavelet decomposition of Fig. 18

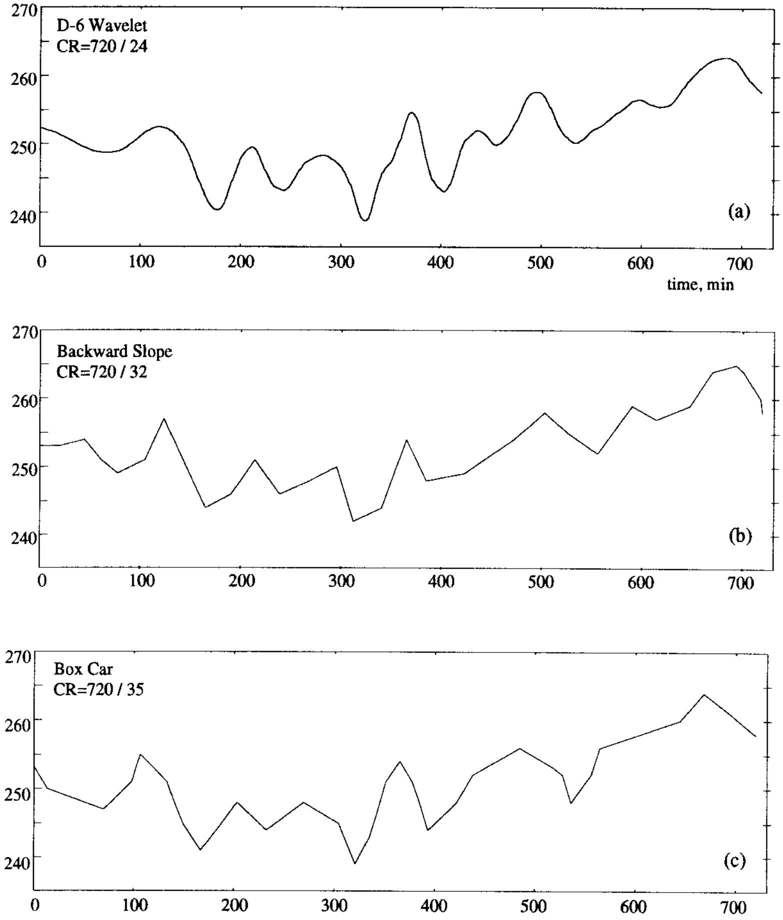


FIG. 20. Performance of data compression techniques: (a) orthonormal wavelet; (b) backward slope; (c) boxcar.

form extrema, and several more examples are presented in Bakshi and Stephanopoulos (1995).

V. Recognition of Temporal Patterns for Diagnosis and Control

The pivotal component of many engineering tasks—e.g., process fault diagnosis and product quality control—is the recognition of certain distinguishing temporal patterns (i.e., features) during the operation of the plants, and the development of associations between process operating patterns and process conditions. This two-step process is depicted in Fig. 3. Pattern recognition is the process through which a given pattern, $\mathbf{p}_i$, is assigned to the correct output class, C_I. A *pattern*, $\mathbf{p}$, is designated as the N-dimensional vector of inputs $(x_1, x_2, \dots, x_N)$ in the input space, which may be partitioned into regions indicated by $(C_1, C_2, \dots, C_K)$. The inputs may represent the value of manipulated and measured process variables, measured external disturbances, violation of output constraints, and/or set points at a given time point, or over a time interval. The feature extraction phase transforms the pattern $\mathbf{p}$, to a feature space, S_x, where only the most relevant parts of the inputs are retained. The extracted features may include the values of computed variables such as derivatives, integrals, or averages over time of the measured variables. Any pattern $\mathbf{p}_i = (x_1, x_2, \dots, x_N)_i^{\mathrm{T}}$ corresponds to a particular class of operating situations, such as sensor or actuator failure, process equipment failure, process parameter changes, activation of unmodeled dynamics, and effect of unmodeled disturbances. During inductive learning, the feature space S_x is partitioned into K mutually exclusive regions, $S_x^{(I)}$, with $I = 1, 2, \dots, K$. Thus

$$S_x^{(I)} \cap S_x^{(J)} = 0, \qquad I, J = 1, 2, \dots, K, \quad \text{but} \quad I \neq J,$$

and

$$\bigcup_{I=1}^{K} S_x^{(I)} = S_x.$$

This mapping from S_x to the classes C_I is determined by the discriminant functions that define the boundaries of regions $S_x^{(I)}$, $I = 1, 2, \dots, K$ in S_x. Let $d_I(\mathbf{p})$ be the discriminant function associated with the Ith class of operating situations, where $I = 1, 2, \dots, K$. Then, a pattern of measurements $\mathbf{p}$, implies the operating situation C_I iff

$$d_I(\mathbf{p}) > d_J(\mathbf{p}) \qquad \text{for all } J \neq I.$$

Also, the boundary of the region $S_x^{(I)}$ with another region $S_x^{(J)}$ is given by

$$d_I(\mathbf{p}) - d_J(\mathbf{p}) = 0.$$

The inductive learning process determines the discriminant functions, using prior examples of $(\mathbf{p}, C_I)$ associations.

For solving the pattern recognition problem encountered in the operation of chemical processes, the analysis of measured process data and extraction of process trends at multiple scales constitutes the feature extraction, whereas induction via decision trees is used for inductive learning. A formal framework for the multiscale analysis of process data and the extraction of qualitative and quantitative features at various scales, based on the mathematical theory of wavelets, has already been developed (see Section III, B). The formal and efficient framework of wavelet decomposition provides a translationally invariant representation, and enables the extraction of qualitative and quantitative features at various scales and temporal locations with minimum distortion.

Having represented measured process data at multiple scales, we can develop consistent models for the various operational tasks. Learning input/ output mappings between features in measured data and process conditions involves determining features that are most relevant to the process conditions. Inductive learning via decision trees provides explicit input/output mappings by identifying the most relevant qualitative and quantitative features from the measured variables. The mapping is easily expressed as *if-then rules* and may even be physically interpretable.

Several techniques from statistics, such as partial least-squares regression, and from artificial intelligence, such as artificial neural networks have been used to learn empirical input/ output relationships. Two of the most significant disadvantages of these approaches are the following:

1. If the input to the learning procedure consists of raw process data from, e.g., several production batches, and the corresponding product yields, the learning technique has to overcome the "curse of dimensionality." The raw data consists of information, much of which may be irrelevant to the learning task, since the sensor data are measured at a scale much smaller than that of the events that may be relevant to the process yield. Such extraneous information increases the complexity of the learning process and may necessitate a large number of training examples to achieve the desired error rate.
2. The model learned is usually a "blackbox" and does not provide any insight into the physical phenomena and events influencing the process outputs.

These disadvantages are overcome by the methodology we will describe in the subsequent paragraph developed by Bakshi and Stephanopoulos. Effects of the curse of dimensionality may be decreased by using the hierarchical representation of process data, described in Section III. Such a multiscale representation of process data permits hierarchical development of the empirical model, by increasing the amount of input information in a stepwise and controlled manner. An explicit model between the features in the process trends, and the process conditions may be learned by *inductive learning* using *decision trees* (Quinlan, 1986; Breiman *et al.*, 1984), as described later on in this section.

A. Generating Generalized Descriptions of Process Trends

Consider a measured operating variable, $x(t)$, and its M distinct measurement records, $[x(t)]_i$, $i = 1, 2, \ldots, M$ over the same range of time. Using the multiscale decomposition of measured variables, discussed in Section III, we can represent each measurement record, $[x(t)]_i$, $i = 1, 2, \ldots, M$ by a finite state of trends, where each trend is a pattern of triangular episodes;

$$[p]_i^k = \{T_1, T_2, \ldots, T_{m_{i,k}}\}_i^k, \qquad k = 1, 2, \ldots, l_i,$$

where superscript k indicates the kth representation of the record at some temporal scale, and l_i is the number of distinct ranges of scale generated by the wavelet interval-tree of scale. $T_1, T_2, \ldots, T_m$ denote the primitive triangles, describing the monotonic temporal behavior of $[x(t)]_i$ over a certain period of time, and could be any of the seven types shown in Fig. 4a;

$$T \in \{A, B, C, D, E, F, G\}.$$

1. *Qualitatively Equivalent Patterns*

Two patterns are qualitatively equivalent if and only if their corresponding sequences of triangular episodes are qualitatively equivalent episode by episode; i.e., the condition

$$\{T_1, T_2, \ldots, T_{m_{i,k}}\}_i^k \langle QE \rangle \{T_1, T_2, \ldots, T_{m_{i,j}}\}_i^j$$

implies that

$$\{T_r\}_i^k \langle QE \rangle \{T_r\}_i^j, \qquad r = 1, 2, \ldots, m_{i,k}; \qquad m_{i,k} = m_{i,j},$$

where $\langle QE \rangle$ denotes the qualitative equivalence.

2. *Generalized Description of Trends*

Consider M distinct batch records of the same measured variable. Let $[p]_i^k$; $i = 1, 2, \ldots, M$ be the appropriate representation of the ith record at some scale of abstraction. If $[p]_1^{k_1} \langle QE \rangle [p]_2^{k_2} \langle QE \rangle \ldots [p]_i^{k_i} \langle QE \rangle \ldots [p]_M^{k_M}$, then the common qualitative trend of all records is called the *generalized description* of trends contained in *M* records of a measured variable.

Figure 17 shows the raw data and the scaled signals for three records (CPR 879, 898, and 935) of carbon dioxide (CO_2) production rate (CPR) from a fed-batch fermentor. The description of the three records at the coarsest scale is $\{DABC\}$ and is identical for the three batches, thus leading to a general description of the variation of CPR. Representation at scales coarser than that of the generalized description will also be qualitatively equivalent. Figure 21 shows the generalized description of six (6) operating variables. These generalized descriptions were generated

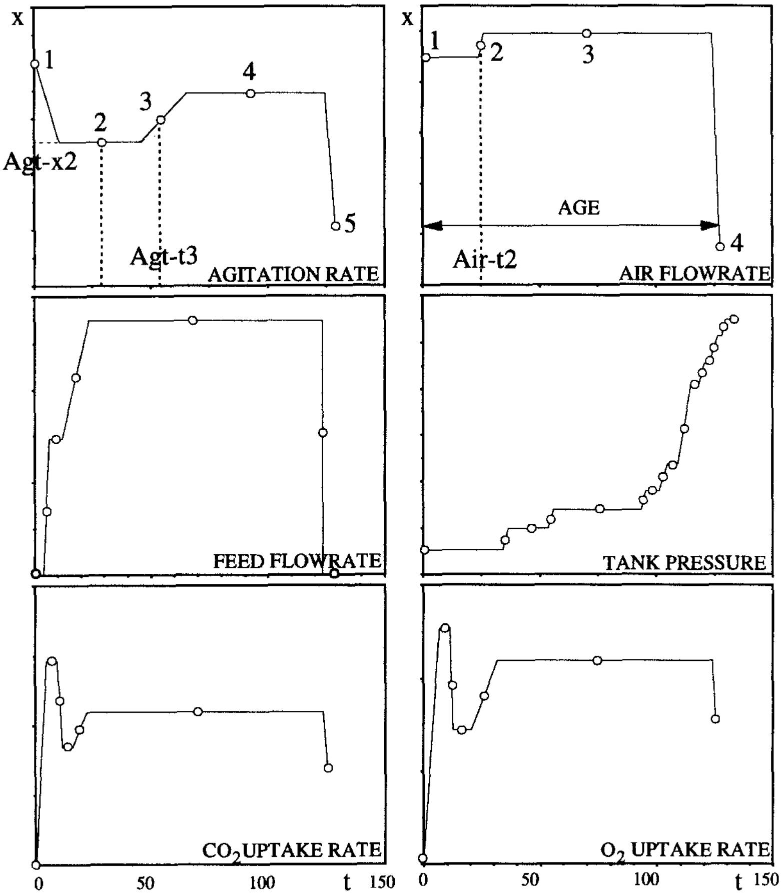


Fig. 21. Generalized description of fed-batch fermentation process data.

from 32 records of each variable, using Witkin's stability criterion to extract the most stable trends.

3. *Pattern Matching of Multiscale Descriptions*

Extraction of the most relevant parts of a process trend requires pattern matching of features that arise from the same physical phenomena, and are qualitatively identical. Matching of qualitatively identical features allows comparison of the qualitative and quantitative features at each scale in an organized, hierarchical manner, and helps fight the curse of dimensionality.

Consider the process data shown in Fig. 17a. These data represent the CPR for the fed-batch fermentation process for three batches giving different yields, as shown in the figure. The raw data are quite different, both in their qualitative and quantitative features. At coarser representations, the three signals start looking similar, and finally, at the coarsest scale, they are qualitatively identical, as shown in Figs. 17f and 17g, resulting in a unique generalized description: $\{D - AB - C\}$, and matching qualitatively identical features is straightforward.

The descriptions obtained using Witkin's stability criterion are shown in Fig. 17h. Pattern matching of the features in these descriptions results in one of the following matches:

Match 1: $AB - CD - AB - CD - A$ $\quad B - C$ CPR 879, CPR 935
$AB - CD - AB - CD - A - (D - A)$
$B - C$ CPR 898
Match 2: $AB - CD - AB - C$ $\quad D - AB - C$ CPR 879 CPR 935
$AB - CD - AB - C(D - A) - D - AB - C$ CPR 898

The qualitative feature, $\{D - A\}$, distinguishes the variable CPR 898 from the other two, and should be evaluated for its relevance to solving the classification problem. The pattern matching task may not be straightforward, especially at lower scales, due to greater detail in the trends. Heuristics may then be used to organize the matches in terms of their likelihood of being physically meaningful. Often, domain-specific information about the process is available for matching features in trends from different examples. For example, information about the sequence and type of features in a trend under normal operation, may be available. Such information simplifies the pattern matching problem, and eliminates several infeasible matches. In the absence of domain-specific information, all matches are equally likely, and need to be evaluated for their relevance to solving the learning problem. For more details on the technical aspects of pattern matching see Bakshi and Stephanopoulos (1994b).

TABLE I

DATA FOR ILLUSTRATING INDUCTION USING DECISION TREES[a]

	Input Features		Output Features	
Example #	Pressure	Temperature	Color	Production Quality
1	N	99	N	Good
2	N	105	A	Bad
3	H	108	N	Good
4	L	92	N	Bad
5	L	106	N	Bad
6	N	106	N	Good
7	L	104	A	Bad
8	N	95	A	Bad

[a] Key: H—high; L—low; N—normal; A—abnormal. *Source*: reproduced from Bakshi and Stephanopoulos (1994b), by permission.

B. INDUCTIVE LEARNING THROUGH DECISION TREES

Once the several records of a process variable have been generalized into a pattern, as indicated in the previous paragraph, we need a mechanism to induce relationships among features of the generalized descriptions. In this section we will discuss the virtues of inductive learning through decision trees.

Inductive learning by decision trees is a popular machine learning technique, particularly for solving classification problems, and was developed by Quinlan (1986). A decision tree depicting the input/output mapping learned from the data in Table I is shown in Fig. 22. The input information consists of pressure, temperature, and color measurements of

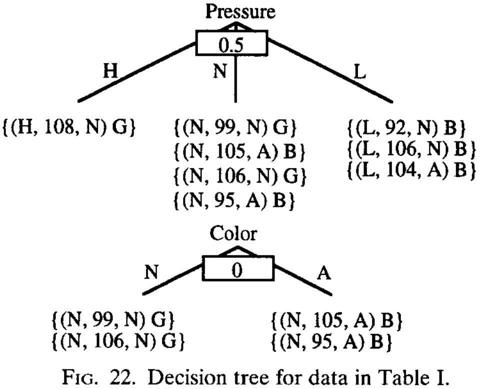


FIG. 22. Decision tree for data in Table I.

a chemical process. The output variable is the product quality corresponding to each set of measured variables. The decision tree in Fig. 22 provides the following information for obtaining product of good quality:

1. If pressure is normal and color is normal, then quality is good.
2. If pressure is low, then quality is bad.
3. If pressure is normal and color is abnormal, then quality is bad.

The model induced via the decision tree is not a blackbox, and provides explicit and interpretable rules for solving the pattern classification problem. The most relevant variables are also clearly identified. For example, for the data in Table I, the value of the temperature are not necessary for obtaining good or bad quality, as is clearly indicated by the decision tree in Fig. 22.

The procedure for generating a decision tree consists of selecting the variable that gives the best classification, as the root node. Each variable is evaluated for its ability to classify the training data using an information theoretic measure of entropy. Consider a data set with K classes, C_I, $I = 1, 2, \ldots, K$. Let M be the total number of training examples, and let M_{C_I} be the number of training examples in class C_I. The information content of the training data is calculated by Shannon's entropy:

$$I(M_{C_1}, M_{C_2}, \ldots, M_{C_K}) = -\sum_{I=1}^{n} \frac{M_{C_I}}{M} \log_2\left(\frac{M_{C_I}}{M}\right). \tag{24}$$

Equation (24) provides a measure of the variety of classes contained in the data set. If all examples belong to the same class, then the entropy is zero. Smaller entropy implies less variety of classes (more order) in the data set. If the data set is split into groups, G_1 and G_2, with M_{C_I,G_J} being the number of examples belonging to class, C_I, that are present in group, G_J, then the total information content is

$$E(G_1, G_2) = \frac{M_{G_1}}{M} I(M_{C_1,G_1}, \ldots, M_{C_K,G_1}) + \frac{M_{G_2}}{M} I(M_{C_1,G_2}, \ldots, M_{C_K,G_2}). \tag{25}$$

Equations (24) and (25) are adequate for designing decision trees. The feature that minimizes the information content is selected as a node. This procedure is repeated for every leaf node until adequate classification is obtained. Techniques for preventing overfitting of training data, such as cross validation are then applied.

Induction via decision trees is a greedy procedure and does not guarantee optimality of the mapping, but works well in practice, as illustrated by successful applications in several areas. The attractive features of learning by decision trees are listed below.

1. The model is easy to understand, and interpret physically. Concise rules may be developed.
2. Only the most relevant features are selected as nodes. Redundant, or unnecessary features are clearly identified through the maximization of entropy change.
3. Both continuous and discrete-valued features can be handled in a uniform framework. This permits easy combination of quantitative and structural methods.
4. No a priori assumptions about the distribution of data or class probability are required.
5. The technique is robust to noisy examples.

The discriminant hypersurface is approximated in a piecewise constant manner. This may result in decision trees that are very large and complicated, and deriving meaningful rules may not be very easy. This problem is alleviated by the hierarchical learning procedure, since the number of features evaluated is increased gradually, only if necessary. Techniques for overcoming some of the disadvantages of decision trees are described by Saraiva and Stephanopoulos (1992).

C. Pattern Recognition with Single Input Variable

Consider the three distinct records (examples) of measurements for the same operating variable, shown in Fig. 23. Let the records $[x(t)]_1$ and $[x(t)]_2$ correspond to operating conditions of class C_A, while the third, $[x(t)]_3$ corresponds to operating conditions of class C_B. At the lowest scale, the high-frequency components of the three records make all of them look very different. The first scale at which the first two records have the same qualitative description, yields the following common trend: $AB - C - B - C$. On the other hand, the most stable representation of the third record is $AB - C - B - CD - AB - C$. In this particular example, syntactic representation of trends has been sufficient to provide complete classification of the three records. Using inductive learning through decision trees leads to the following two cases for classifying trends of class C_A

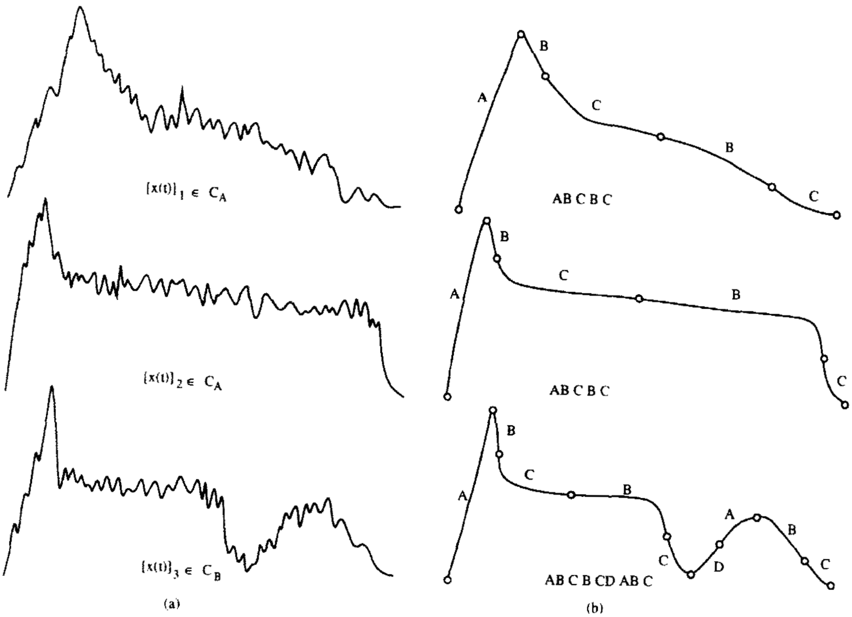


FIG. 23. (a) The raw data of three distinct records and (b) their corresponding syntactic generalizations. (Reprinted from Bakshi and Stephanopoulos, Representation of Process Trends, Part IV. *Computers and Chemical Engineering*, **18**(4), p. 303, Copyright (1994), with kind permission from Elsevier Science Ltd., The Boulevard, Langford Lane, Kidlington OX5 1GB, UK).

from those of class C_B:

Case 1. The syntactic difference is the string $(D-AB-C)$ at the end of the trend, since

$$\begin{array}{ll} AB - C - B - C & \text{(description of first two records)} \\ AB - C - B - C(D-AB-C) & \text{(description of third record)} \end{array}$$

Case 2. The syntactic difference is the string $(B-CD-A)$ in the middle of the trend, since

$$\begin{array}{lll} AB - C - & B - C & \text{(for first two records)} \\ AB - C - (B-CD-A) & B - C & \text{(for third record)} \end{array}$$

D. PATTERN RECOGNITION WITH MULTIPLE INPUT VARIABLES

The inductive classification of multiple-dimensional trends involves the mapping between the distinguishing features of several input-variables and the corresponding classes of a single output. It is carried out in two phases, as follows:

1. Phase 1: Generate Generalized Descriptions and Extract Distinguishing Features for Each Variable

The procedure for generating generalized trends was described earlier. The features of the generalized trends are given by the triangular episodes at any level of required detail, i.e., qualitative, semiquantitative, real-valued analytic.

2. Phase 2: Learn Mapping between Extracted Features and Output Classes

Once the distinguishing features for each input variable have been extracted through the generalization of descriptions of the available records, these features become the inputs of the inductive learning procedure through decision trees.

Bakshi and Stephanopoulos (1994b) have applied the above procedure to a fed-batch fermentation process. The problem involved 41 sets of batch records on 24 measured variables. Of these variables only very few were found by the decision tree to be relevant, and yield rules such as the following for guiding the diagnosis or control of a fermentor.

Diagnostic Rule-1 If the duration of the high bottom dissolved oxygen (BDO) phase is > 3.6 h and the level of CO_2 generation during the production phase is > 4.3 units, then the quality of the fermentation is excellent.

Control rule-1. To achieve fermention of excellent quality, keep (a) the agitation rate in the first scaling episode (growth phase) > 37 units and (b) the duration of the first episode in the air flowrate ≤ 25 h.

For the detailed discussion on the inductive learning of diagnostic and control rules around the fed-batch fermentor system, the reader should refer to the work of Bakshi and Stephanopoulos (1994b).

VI. Summary and Conclusions

The wavelet decomposition of measured data provides a natural framework for the extraction of temporal features, which characterize operating process variables and their trends. Such characterization, local in frequency and time, provides valuable insight as to what is going on in a process, and thus it can explicitly support a number of engineering methodologies, such as data compression, diagnosis of process upsets, and pattern recognition for quality control. Nevertheless, a number of interesting issues arise that can shape future developments in the area process operations and control.

1. *Multiscale modeling of process operations.* The description of process variables at different scales of abstraction implies that one could create models at several scales of time in such a way that these models communicate with each other and thus are inherently consistent with each other. The development of multiscale models is extremely important and constitutes the pivotal issue that must be resolved before the long-sought integration of operational tasks (e.g., planning, scheduling, control) can be placed on a firm foundation.
2. *Multiscale process identification and control.* Most of the insightful analytical results in systems identification and control have been derived in the frequency domain. The design and implementation, though, of identification and control algorithms occurs in the time domain, where little of the analytical results in truly operational. The time-frequency decomposition of process models would seem to offer a natural bridge, which would allow the use of analytical results in the time-domain deployment of multiscale, model-based estimation and control.
3. *Integration of process operational tasks.* The industrial deployment of computer-aided systems that can integrate planning-scheduling-diagnosis-control rests on two pillars; multiscale process models (see paragraph 1, above) and multiscale depiction of process data. Although the wavelet decomposition offers the theoretical answer to the second, the industrial implementation requires the solution of the following problems: (a) integration of quantitative, qualitative and semiquantitative descriptions of process operations through the establishment of an "appropriate" language; and (b) integration of planning, scheduling, diagnosis, and control methodologies around the common language. These issues require creative modeling of process data around the wavelet decomposition and need to be addressed soon by researchers.

References

Bader, F. P., and Tucker, T. W., Data compression applied to a chemical plant using a distributed historian station. *ISA Trans.* **26**(4), 9–14 (1987a).

Bader, F. P., and Tucker, T. W., Real-time data compression improves plant performance assessment. *InTech.* **34,** 53–56 (1987b).

Bakshi, B. R., and Stephanopoulos, G., Wave-net: A multi resolution, hierarchical neural network with localized learning. *AIChE J.* **39**(1), 57–81 (1993).

Bakshi, B. R., and Stephanopoulos, G., Representation of process trends. Part III. Multi-scale extraction of trends from process data. *Comput. Chem. Eng.* **18**, 267 (1994a).

Bakshi, B. R., and Stephanopoulos, G., Representation of process trends. Part IV. Induction of real-time patterns from operating data for diagnosis and supervisory control. *Comput. Chem. Eng.* **18** 303 (1994b).

Bakshi, B. R., and Stephanopoulos, G., Compression of chemical process data through functional approximation and feature extraction. *AIChE J.*, accepted for publication (1995).

Bastl, W., and Fenkel, L., Disturbance analysis systems. *In* "Human Diagnosis of System Failures," (Rasmussen and Rouse, eds.) Nato Symp. Denmark, Plenum, New York, 1980.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., "Classification and Regression Trees." Wadsworth, Belmont, CA (1984).

Cheung, J. T.-Y., Representation and extraction of trends from process data. Sc.D. Thesis, Massachusetts Institute of Technology, Dept. Chem. Eng., Cambridge, MA (1992).

Cheung, J. T.-Y., and Stephanopoulos, G., Representation of process trends. Part I. A formal representation framework. *Comput. Chem. Eng.* **14**, 495–510 (1990).

Coifman, R. R., and Wickerhauser, M. V., Entropy-based algorithms for best basis selection. *IEEE Trans. Inf. Theory* **38**(2), 713–718 (1992).

Daubechies, I., Orthonormal Bases of Compactly Supported Wavelets, *Comm. Pure Appl. Math.*, **XLI**, 909–996 (1988).

Dvorak, D. L., "Expert Operations Systems," Tech. Rep. Dept. of Computer Science, University of Texas at Austin, 1987.

Feehs, R. J., and Arce, G. R., "Vector Quantization for Data Compression of Trend Recordings," Tech. Rep. 88-11-1, University of Delaware, Dept. Elect. Eng., Newark, 1988.

Fukunaga, K., "Introduction to Statistical Pattern Recognition." Academic Press, Boston, 1990.

Funahashi, K. I. On the approximate realization of continuous mappings by neural networks. *Neural Networks* **2** 183–192 (1989).

Goupillaud, P., Grossmann, A., and Morlet, J., Cycle-octave and related transforms in seismic signal analysis. *Geoexploration* **23**, 85–102 (1984).

Gray, R. M., Vector quantization. *IEEE ASSP Mag.*, April, pp. 4–29 (1984).

Hale, J. C., and Sellars, H. L., Historical data recording for process computers. *Chem. Eng. Prog.*, November, pp. 38–43 (1981).

Hummel, R., and Moniot, R., Reconstructions from zero crossings in scale space. *IEEE Trans. Acoust., Speech, Signal Processing* **ASSP 37**(12), 2111–2130 (1989).

Jantke, K., "Analogical and Inductive Inference." Springer-Verlag, Berlin, 1989.

Kanal, L. N., and Dattatreya, G. R., Problem-solving methods for pattern recognition. *In* Handbook of Pattern Recognition and Image Processing," (T. Y. Young, and K.-S. Fu, eds.) Academic Press, New York, 1985.

Kramer, M. A., Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.* **37**, 233–243 (1991).

Long, A. B., and Kanazava, R. M., "Summary and Evaluation of Scoping and Feasibility Studies for Disturbance Analysis and Surveillance Systems," EPRI Report NP-1684. Electr. Power Res. Inst., Palo Alto, CA, 1980.

MacGregor, J. F., Marlin, T. E., Kresta, J. V., and Skagerberg, B., Multivariate statistics methods in process analysis and control. *In* "Chemical Process Control, CPCIV," (Y. Arkun and W.H. Ray, eds.). CACHE, AIChE Publishers, New York, 1991.

Mallat, S. G., A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-11**(7), 674–693 (1989).

Mallat, S. G., Zero crossing of a wavelet transform. *IEEE Trans. Inf. Theory* **IT-37**(4), 1019–1033 (1991).

Mallat, S., and Zhong, S., Characterization of signals from multiscale edges, *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-14**(7), 710–732 (1992).

Marr, D., and Hildreth, E., Theory of edge detection. *Proc. R. Soc. London B Ser.* **207**, 187–217 (1980).

Meyer, Y., Principle d'incertitude, bases hilbertiennes et algebres d'operateurs. *Bourbaki Sem.* No. 662 (1985–1986).

O'Shima, E., Computer-aided plant operation. *Comput. Chem. Eng.* **7**, 311 (1983).

Papoulis, A., "Signal analysis." McGraw-Hill, New York, 1977.

Quinlan, J. R., Induction of decision trees. *Mach. Learn.* **1**(1), 81–106 (1986).

Rumelhart, D. E., McClelland, J. L., "Parallel Distributed Processing," Vol. 1. MIT Press, Cambridge, MA, 1986.

Saraiva, P., and Stephanopoulos, G., Continuous process improvement through inductive and analogical learning. *AIChE J* **38**(2), 161–183 (1992).

Silverman, B. W., "Density Estimation for Statistics and Data Analysis." Chapman & Hall, New York, 1986.

Stephanopoulos, G., Artificial intelligence...'What will its contributions be to process control?' *In* "The Second Shell Process Control Workshop," (D.M. Prett, C. E. Garcia, and B. L. Ramaker, eds.) Butterworth, Stoneham, MA, 1990.

Takei, S., Multiresolution analysis of data in process operations and control. M.S. Thesis, Massachusetts Institute of Technology, Dept. Chem. Eng., Cambridge, MA, 1991.

Vaidyanathan, R., and Venkatasubramanian, V., Process fault detection and diagnosis using neural networks: II. Dynamic processes. *AIChE Ann. Meet.*, Chicago, IL (1990).

Van Trees, H., "Detection, Estimation and Modulation Theory." Wiley, New York, 1968.

Wickerhauser, M. V., "INRIA Lectures on Wavelet Packet Algorithms." Yale University, New Haven, CT, 1991.

Witkin, A. P., Scale space filtering: A new approach to multi-scale description. *In* "Image Understanding" (S. Ullman and W. Richard, eds.), pp. 79–95. Ablex, Norwood, NJ, 1983.

Yuille, A.L., and Poggio, T., Fingerprints theorems. *Proc. Natl. Conf. on Artif. Intell.*, pp. 362–365 (1984).