SYMBOLIC TIME SERIES ANALYSIS (STSA) FOR ANOMALY DETECTION

Devendra Tolani
Intelligent Automation Inc
December 6th 2005

National Institute of Standards and Technology (NIST)

Acknowledgements

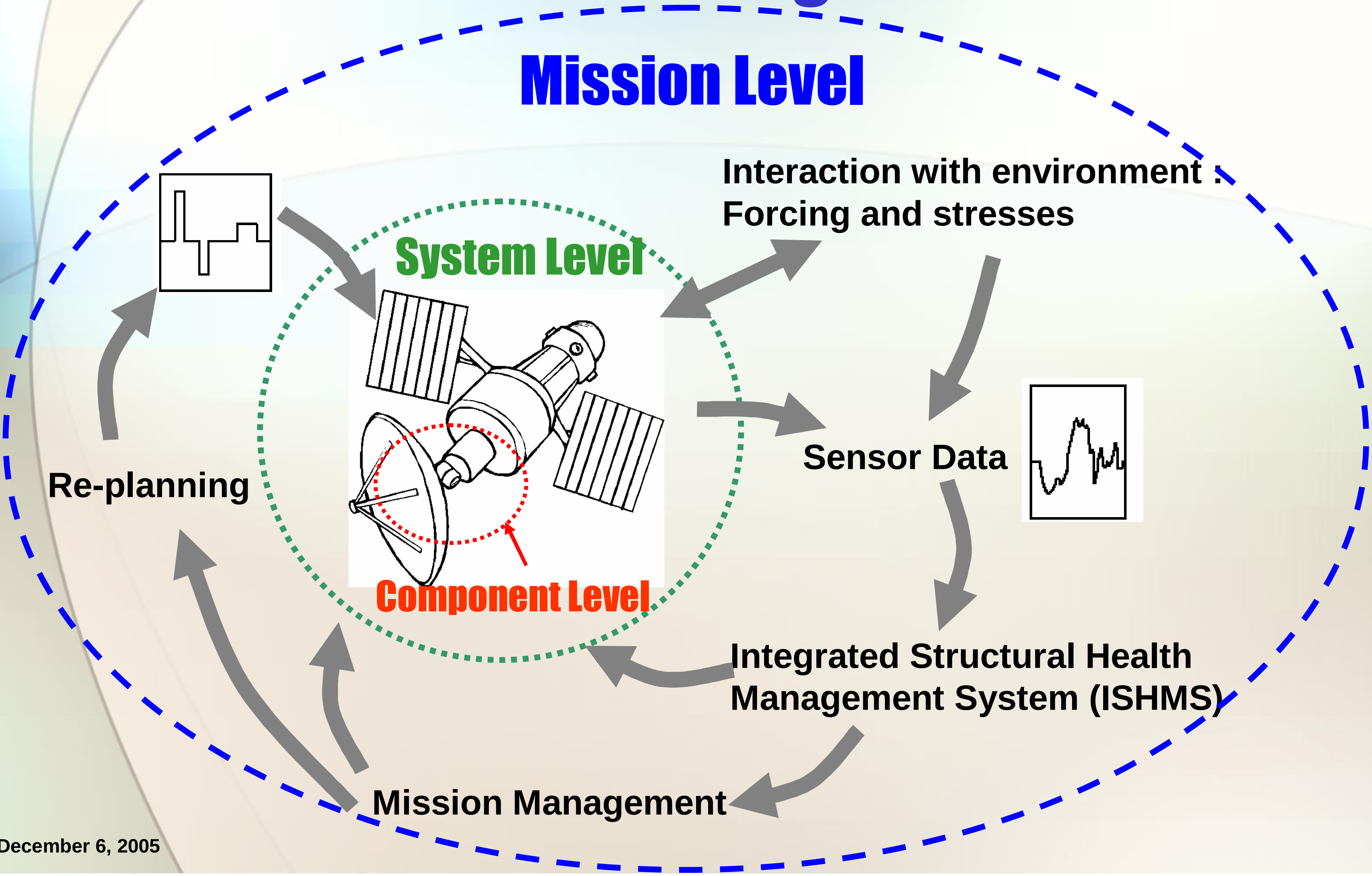
- Dr Asok Ray and my colleagues at Penn State
- This work has been supported in part by ARO under grant No DAAD19-01-1-0646

Outline of the talk

- Motivation
- Introduction and explanation of the STSA concept
- Traditional Approaches
- Practical Applications of the symbolic dynamics based Anomaly detection
  - Gas turbine Simulation
  - Hybrid electronic circuit
  - Fatigue Test Apparatus for Damage Sensing in Ductile Alloys
- Conclusions and Summary

Motivation

Mission Level Perspective for Health Management



Definition and Motivation

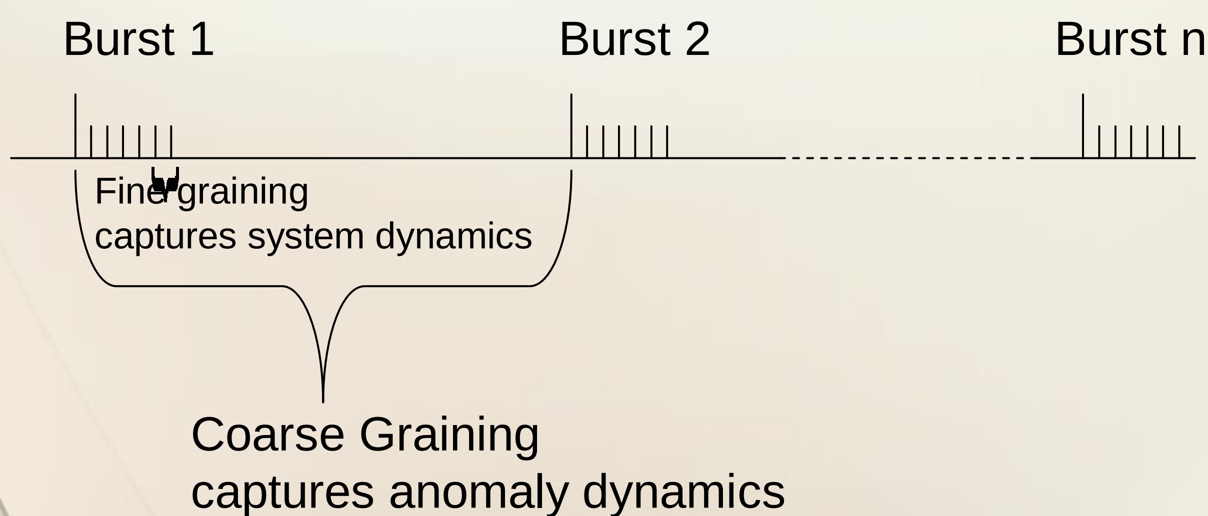
- **"Anomaly is deviation from nominal behavior"**
- Why early detection of Anomalies?
  - Prevention of cascading catastrophic failures
  - Enhancement of performance and availability
- What do we work with?
  - Modeling of complex dynamical systems solely based on fundamental principles of physics is often infeasible
  - Data driven approach : Time Series data generated from sensors

Physics of Anomalies

Two time scale approach

- Anomaly Propagation : slow
- System Response to inputs : fast

Collect bursts of data at fast sampling

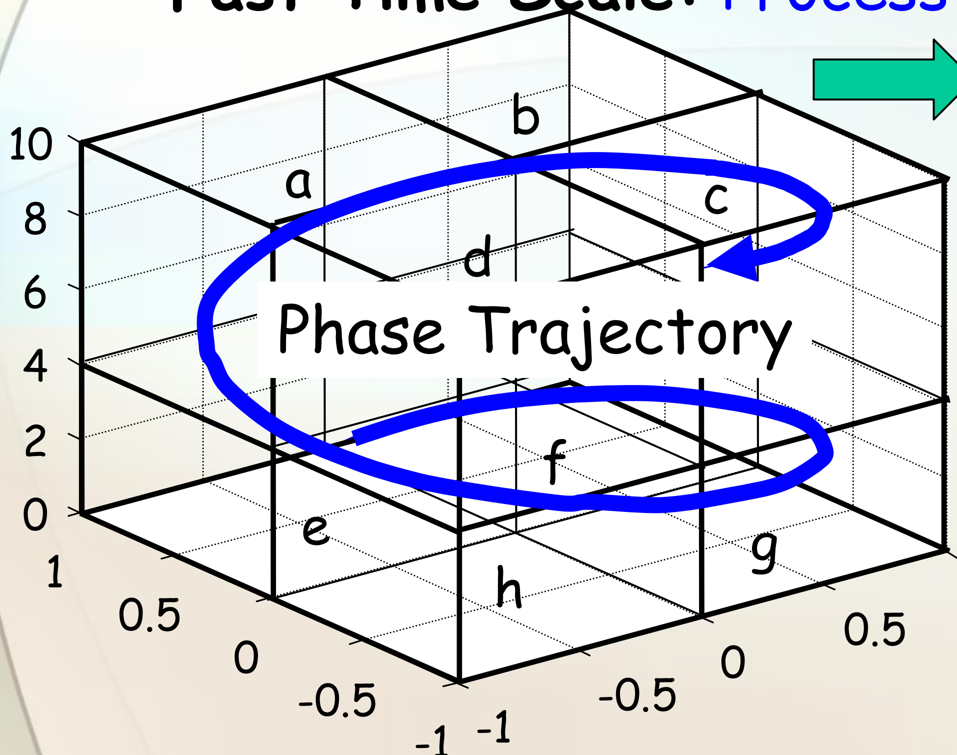


Notion of Symbolic Dynamics

Multi-Time-Scale Nonlinear Dynamics

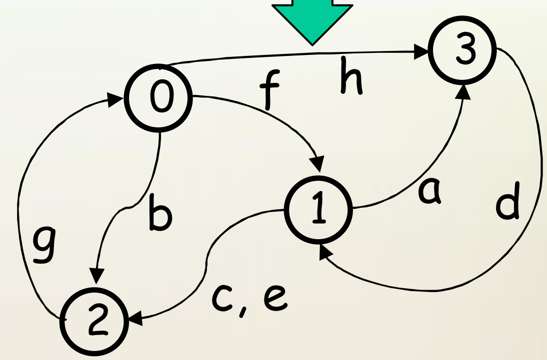
Slow Time Scale: Anomaly Propagation

Fast Time Scale: Process Response



Symbol Sequence

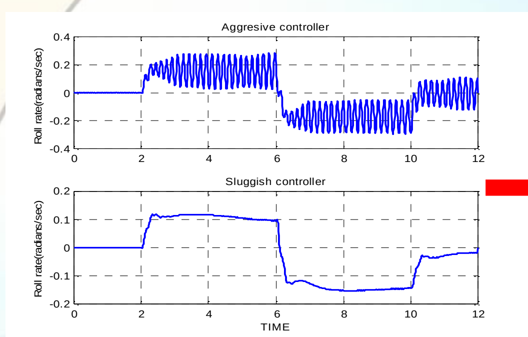
……f c g h d a d c……



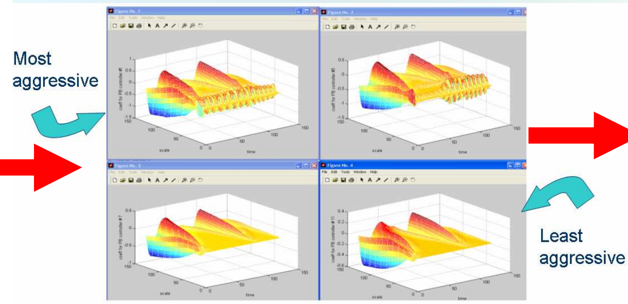
Finite State Machine

- Discretization of the Dynamical System in Space and Time
- Representation of Trajectories as Sequences of Symbols

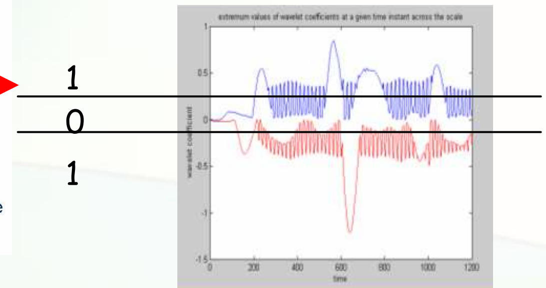
Symbolic Dynamics Methodology



1. Time series data

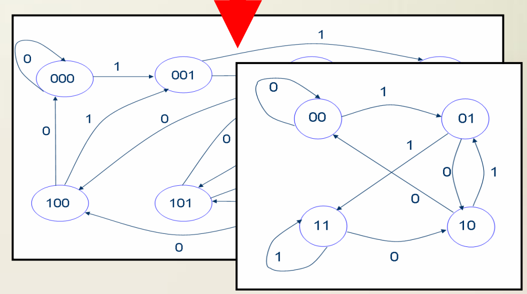


2. Wavelet transform



3. Partitioning

4. Generation of symbol string 000001111010101......

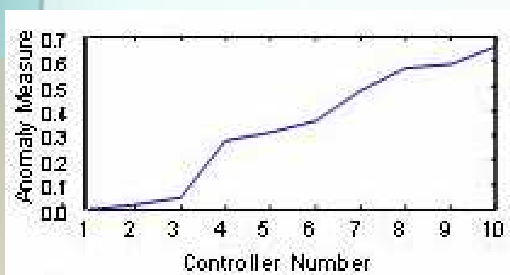


5. Construction of Finite State Machines

$$p_i = \frac{n_i}{\sum_{i=1}^{N} n_i}$$

$$a_i = \| p_i - p_1 \|$$

6. Computation of state probabilities p_i and the anomaly measure a_i

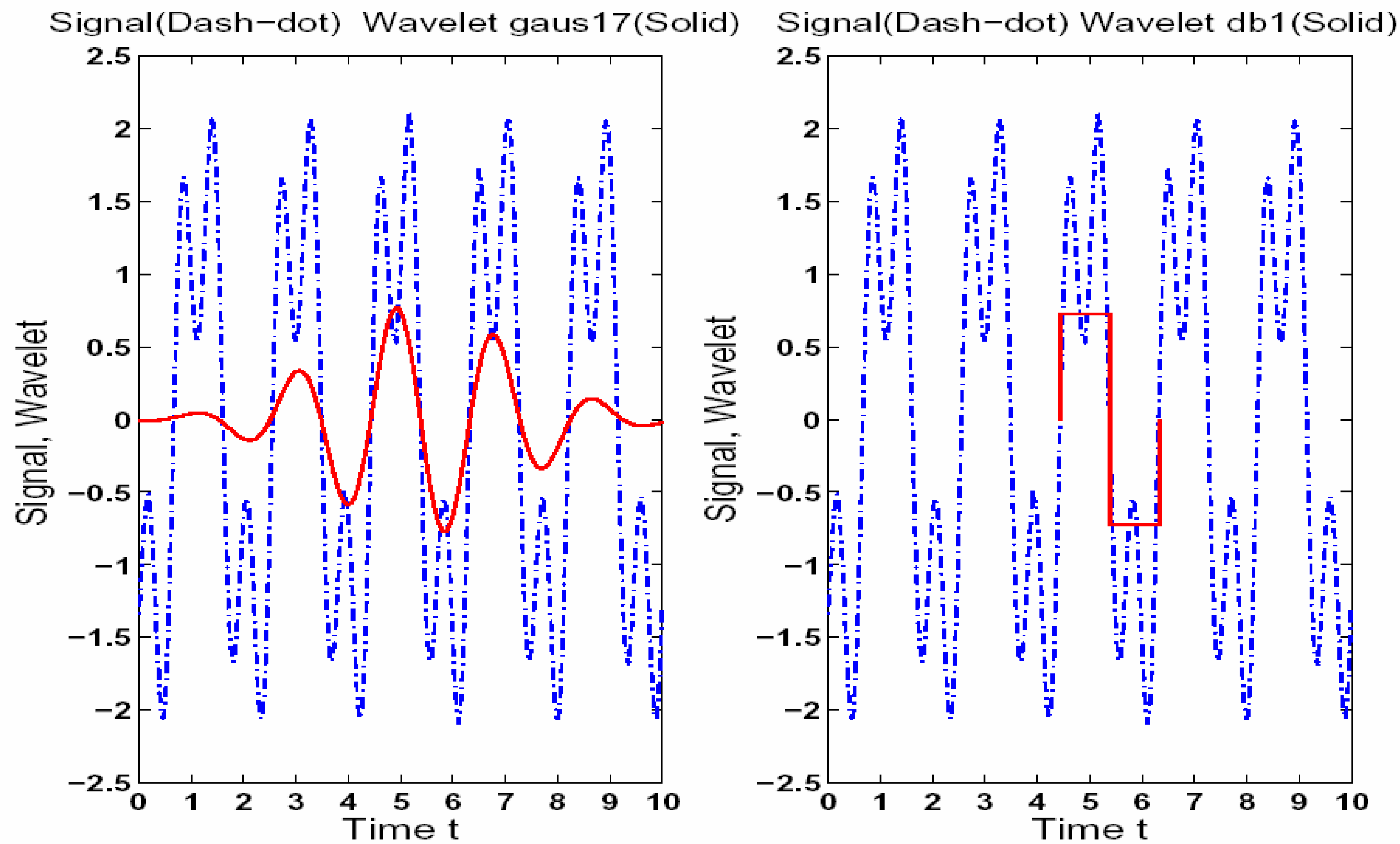


7. Deviation from nominal behavior captured using symbolic dynamics tools

Why Wavelets?

- Preprocessing Times series data necessary for extraction of pertinent information
- Fourier analysis is sufficient if the signal to be analyzed is stationary and if the time period is accurately known
- Wavelet analysis is needed for non-stationary characteristics such as drifts, abrupt asynchronous changes and frequency trends

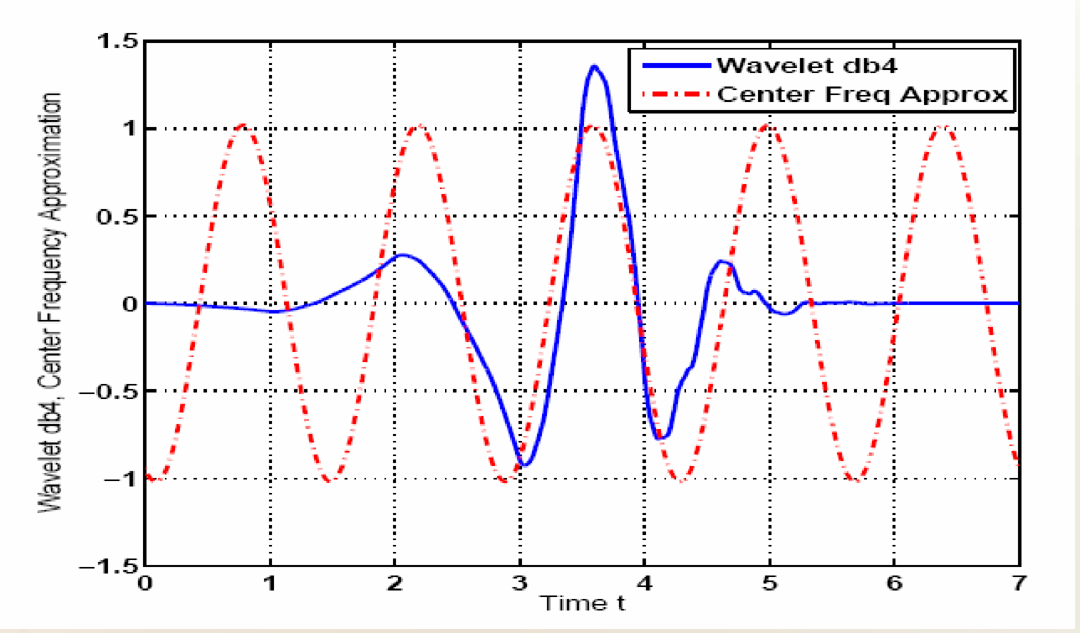
Choosing Appropriate Wavelet



Pseudo-frequency of a wavelet

For every wavelet, there exists a certain frequency called the center frequency F_c that has the maximum modulus in the Fourier transform of the wavelet .The pseudo-frequency f_p of the wavelet at a particular scale a and sampling interval Δt is

$$f_p = F_c / (a.\Delta t)$$



Why is pseudo-frequency important?

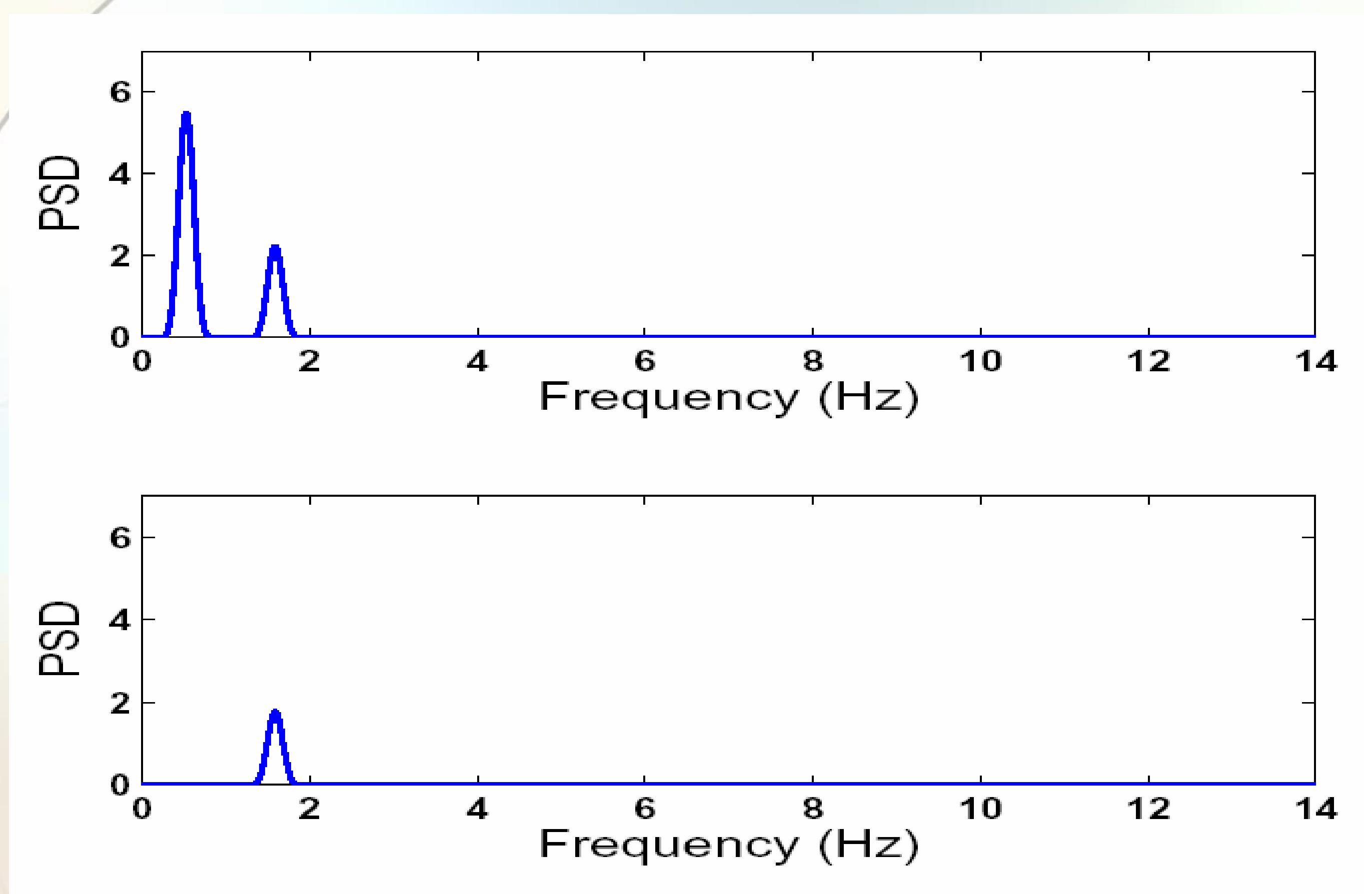
The wavelet coefficients of the signal are significantly large when the pseudo-frequency f_p of the wavelet corresponds to the locally dominant frequencies in the underlying signal

Choosing appropriate scales

- Perform PSD (Power Spectral Density) analysis on the time series data to find the frequencies of interest

- Substitute the above frequencies in place of f_p in the equation to obtain the respective scale in terms of the known parameters

Example



Power Spectrum plots for Nominal and Anomalous Condition :
The Wavelet coefficients at scales corresponding to pseudo-frequency of 0.54 Hz would be smaller in magnitude for anomalous condition

Constructing Scale Series data

- Once the wavelet and the scales are chosen, the wavelet coefficients are evaluated for each scale.

- The graphs of wavelet coefficients versus scale, at selected time shifts, are stacked starting with the smallest value of scale and ending with its largest value and then back from the largest value to the smallest value of the scale at the next instant of time shift.

- The arrangement of the resulting *scale series* data in the wavelet space is similar to that of the time series data in the phase space.

- The wavelet space is partitioned into segments of coefficients on the ordinate separated by horizontal lines

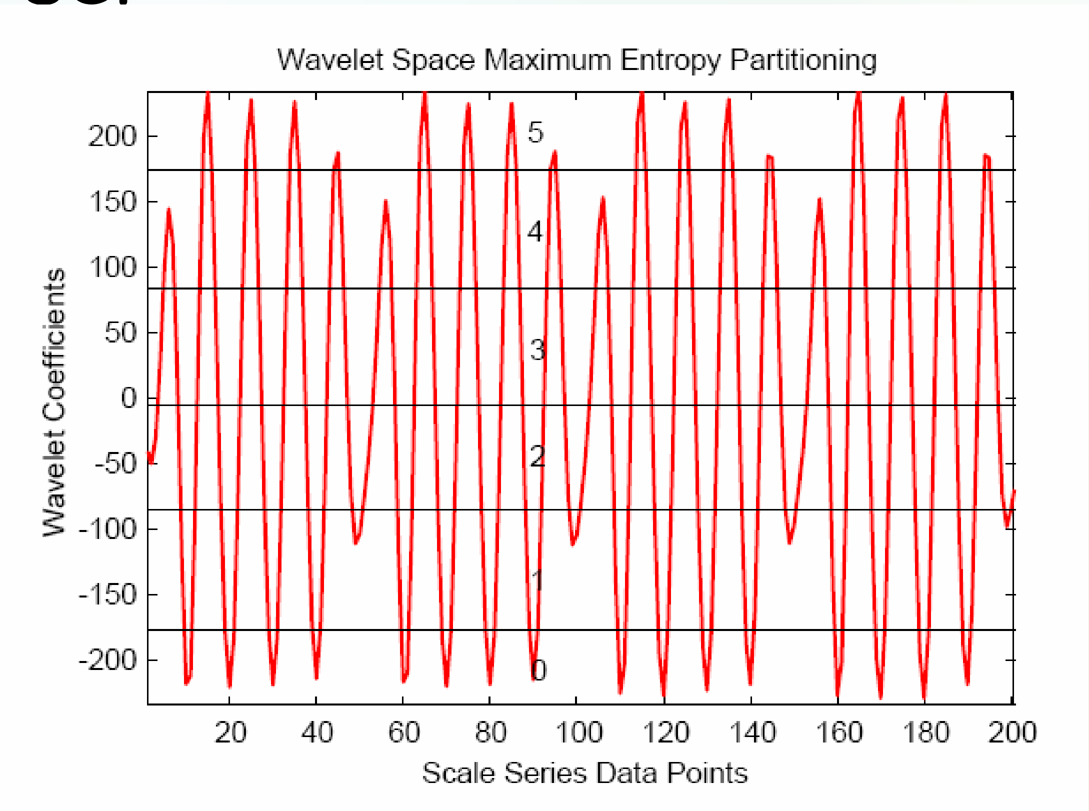
Partitioning

Uniform Partitioning

- the maximum and minimum of the scale series are evaluated. and the ordinates between the maximum and minimum are divided into equal sized regions.
- These regions are mutually disjoint and thus form a partition.
- Each region is then labeled with one symbol from the alphabet. If the data point lies in a particular region, it is coded with the symbol associated with that region.
- Thus, a sequence of symbols is created from a given sequence of scale series data.

Maximum Entropy Partitioning

- regions with more information are partitioned finer and those with sparse information are partitioned coarser



Maximum Entropy Partitioning

Maximum entropy is achieved by the partition that induces uniform probability distribution of the symbols in the alphabet

- N=length of scale series data
- $|\Sigma|$=size of the alphabet
- Sort the scale series data in ascending order
- Every consecutive segment of length int(N/$|\Sigma|$) is a distinct element of the partition. (where Int(x)=is the greatest integer less than or equal to x)

Choice of Alphabet Size

- Entropy Rate based approach
- H(k) denotes the entropy of the symbol sequence obtained by partitioning the data with k symbols

$$H(k) = -\sum_{i=1}^{i=k} p_i log_2 p_i$$

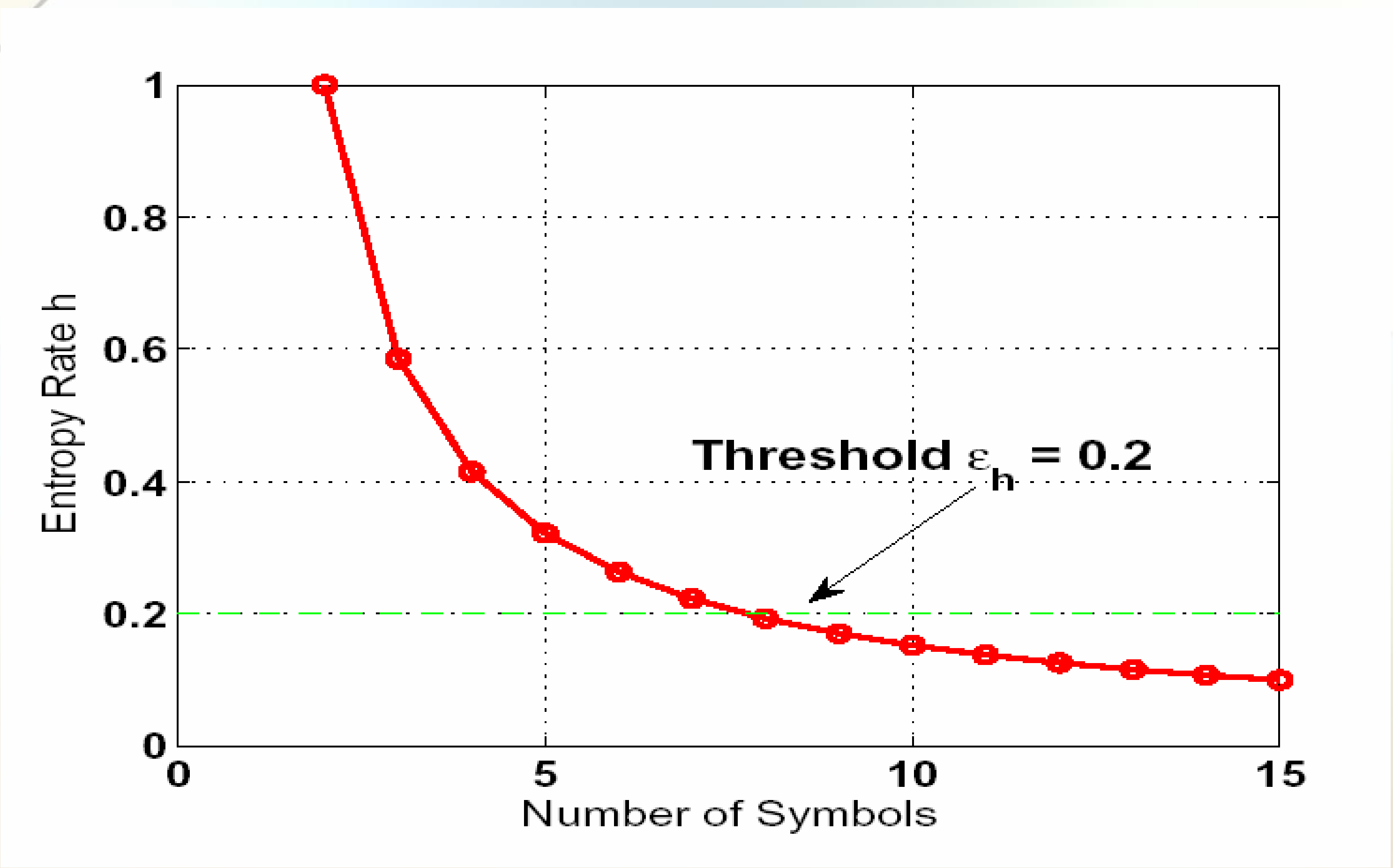
- Entropy Rate is given by

$$h(k) = H(k) - H(k-1) \quad \forall k \geq 2$$

Algorithm for choosing Alphabet Size

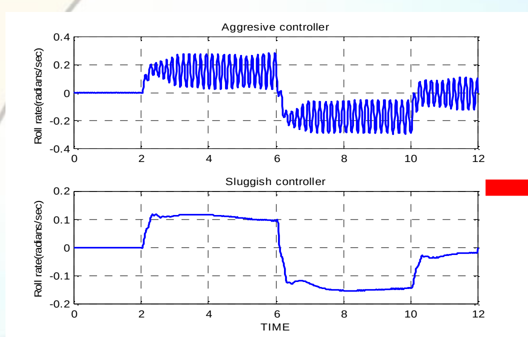
1. Set $k = 2$. Choose a threshold ϵ_h, where $0 < \epsilon_h << 1$.
2. Sort the scale series data set (of length N) in the ascending order.
3. Every consecutive segment of length $\lfloor \frac{N}{|\Sigma|} \rfloor$ in the sorted data set (of length N)forms a distinct element of the partition.
4. Convert the scale series sequence to a symbol sequence with the partitions obtained in Step 3. If the data point lies within or on the lower bound of a partition, it is coded with the symbol associated with that partition.
5. Compute the symbol probabilities p_i, i=1,2,...k.
6. Compute the entropy $H(k) = -\sum_{i=1}^{i=k} p_i log_2 p_i$ and the entropy rate $h(k) = H(k) - H(k-1)$
7. If $h(k) < \epsilon_h$, then exit; else increment k by 1 and go to Step 3.

Example

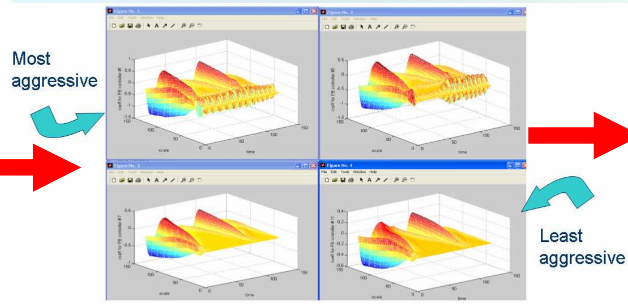


Selection of number of symbols from Entropy Rate

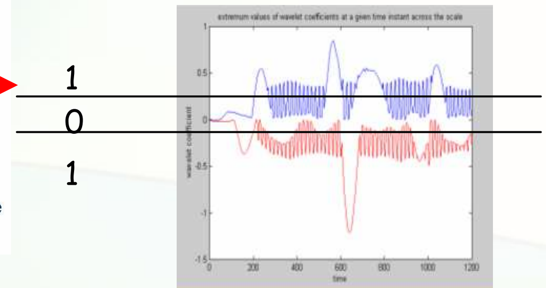
Symbolic Dynamics Methodology



1. Time series data

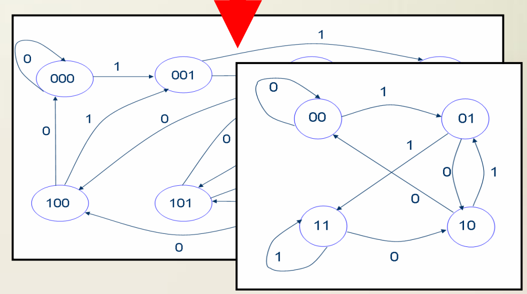


2. Wavelet transform



3. Partitioning

4. Generation of symbol string 000001111010101……



5. Construction of Finite State Machines

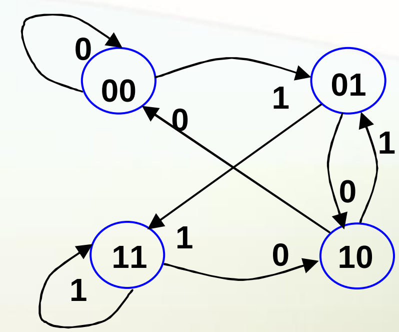
State Machine Construction

D-Markov Machine

State to state transitions from a symbol sequence

- alphabet size = $|A| \in \mathbb{N}$
- window size = D $\in \{0\} \cup \mathbb{N}$
- kth word (state)= $W^k = w_0^k w_1^k \ldots w_{D-1}^k$
- kth word value = $W^k = \sum_{i=0}^{D-1} (w_i^k) A^{D-1-i}$

Example



$$\begin{bmatrix} p_{00} & 1-p_{00} & 0 & 0 \\ 0 & 0 & p_{01} & 1-p_{01} \\ p_{10} & 1-p_{10} & 0 & 0 \\ 0 & 0 & p_{11} & 1-p_{11} \end{bmatrix}$$

|A|=2; D=2; A^D = 4

$$w_0^k w_1^k \ldots\ldots\ldots\ldots w_{D-1}^k$$

$$w_0^k \leftarrow w_0^{k+1} w_1^{k+1} \ldots\ldots\ldots w_{D-1}^{k+1} \leftarrow w_{D-1}^{k+1}$$

Out **In**

Most ***least***

significance

- (k+1)th word = $W^{k+1} = w_0^{k+1} w_1^{k+1} \ldots w_{D-1}^{k+1}$
- (k+1)th word value = $W^{k+1} = (W^k - w_0^k A^{D-1}) A + w_{D-1}^{k+1}$

State Space Construction via D-Markov Machine

- Computationally efficient
- Fixed depth D and alphabet size A
  Number of states $N=A^D$
- Only the state transition probabilities to be determined based on symbol strings derived from time series data or wavelet-transformed data
- States represented by an equivalence class of strings whose D most recent symbols are identical

Anomaly Detection Procedure

Problem can be split into two parts

(1)Forward Problem (Analysis): Anomalies apriori known. Objective is to find their signature and create a databank

(2)Inverse problem (Synthesis): Anomaly classification. Objective is to match the signatures with the processed real-time data

Anomaly Detection Procedure

Forward (Analysis) Problem:

1. Selection of an appropriate set of input stimuli.
2. Signal–noise separation, time interval selection, and phase-space construction.
3. Choice of a phase space partitioning to generate symbol alphabet and symbol sequences.
4. State Machine construction using generated symbol sequence(s)
5. Selection of an appropriate metric for the anomaly measure
6. Formulation and calibration of a relation between the computed anomaly measure and known physical anomaly under which the time-series data were collected at different (slow-time) epochs.

Anomaly Detection Procedure

Inverse (Synthesis) Problem:

- Excitation with known input stimuli selected in the forward problem.
- Generation of the stationary behavior as time-series data for each input stimulus at different (slow-time) epochs.
- Embedding the time-series data in the phase space determined for the corresponding input stimuli in Step 2 of the forward problem.
- Generation of the symbol sequence using the same phase-space partition as in Step 3 of the Forward problem.
- State Machine construction using the symbol sequence and determining the anomaly measure.
- Detection and identification of an anomaly (if any) based on the computed anomaly measure and the relation derived in Step 6 of the forward problem.

Comparison of Epsilon Machine and D-Markov Machine

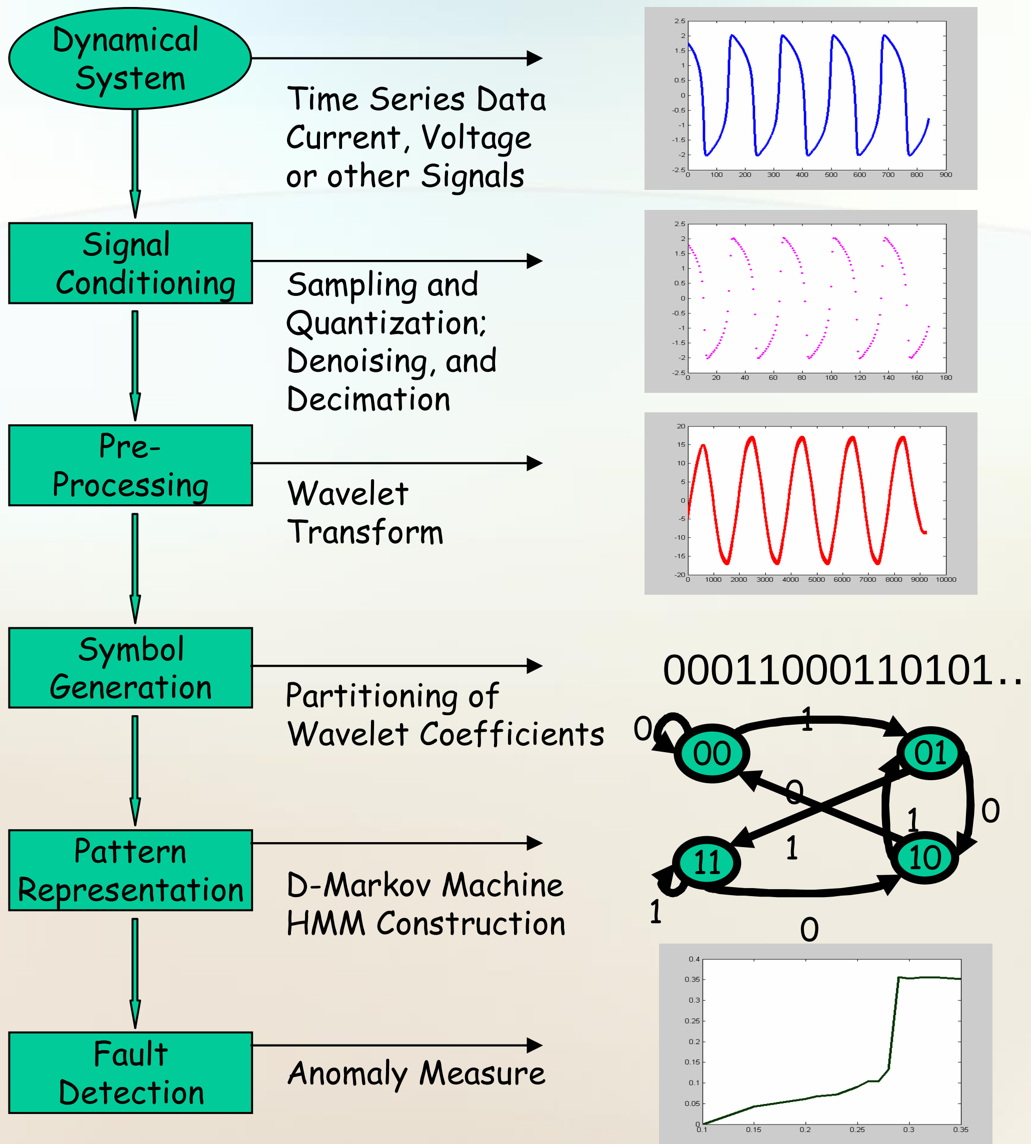
Epsilon Machine [Santa Fe Institute]

- A priori unknown machine structure
- Optimal prediction of the symbol process
- Maximization of mutual information
  (i.e., minimization of conditional entropy)
  $I[X;Y] = H[X] - H[X|Y]$

D-Markov Machine

- A priori known machine structure
  (Fixed order fixed structure with given **|A|** and **D**)
- Suboptimal prediction of the symbol process

Summary of Anomaly Detection Procedure



Traditional Approaches

Traditional method 1: RBFNN

- Neural networks (NN) provide a new suite of nonlinear algorithms for feature extraction and classification.
- A major class of NN model is the radial basis function (RBF) neural network (NN)
  - the activation of a hidden unit is determined by the distance between the input vector and the prototype vector
  - essentially a nearest neighbor type of classifier
  - The anomaly measure is defined as the distance function

$$\mathcal{M} = d(f_{nom}, f_k)$$

- Advantages:
  - Unified approaches for feature extraction and classification
  - flexible procedures for finding good, moderately nonlinear solution

Traditional method 2: PCA

- The best known linear feature extractor is the Principal Component Analysis (PCA)
- Makes use of Karhunen-Loéve expansion to compute the *m* largest eigenvectors of the covariance matrix of the *N* *d*-dimensional patterns.
- PCA uses the most expressive features (eigenvectors with the largest eigenvalues) to effectively approximate the data by a linear subspace using the mean squared error criterion.

Practical Applications of the symbolic dynamics based Anomaly detection

1. Gas turbine Simulation
2. Hybrid electronic circuit
3. Fatigue Test Apparatus for Damage Sensing in Ductile Alloys

Application 1: Gas Turbine Project

To identify slow time scale anomalies for health management of aircraft gas turbine

For this purpose comparison study of different pattern recognition algorithms

- Based on new theory developed in Symbolic Dynamics and Information Theory
- More Traditional methods
  - Principal Component Analysis (PCA)
  - Artificial Neural Network (ANN)

Proposed Scheme 1 : SFNN

- Finding the dimensionality of the phase space of relevant system dynamics is difficult especially if the time series data are noise-corrupted
- Kennel and Buhl have formulated a phase-space partitioning method that is built upon the concept of Symbolic False Nearest Neighbors (SFNN)
- The partitions are defined with respect to a set of radial-basis influence functions
- Major advantage: The partitioning is entirely by the algorithm, based on the time series data

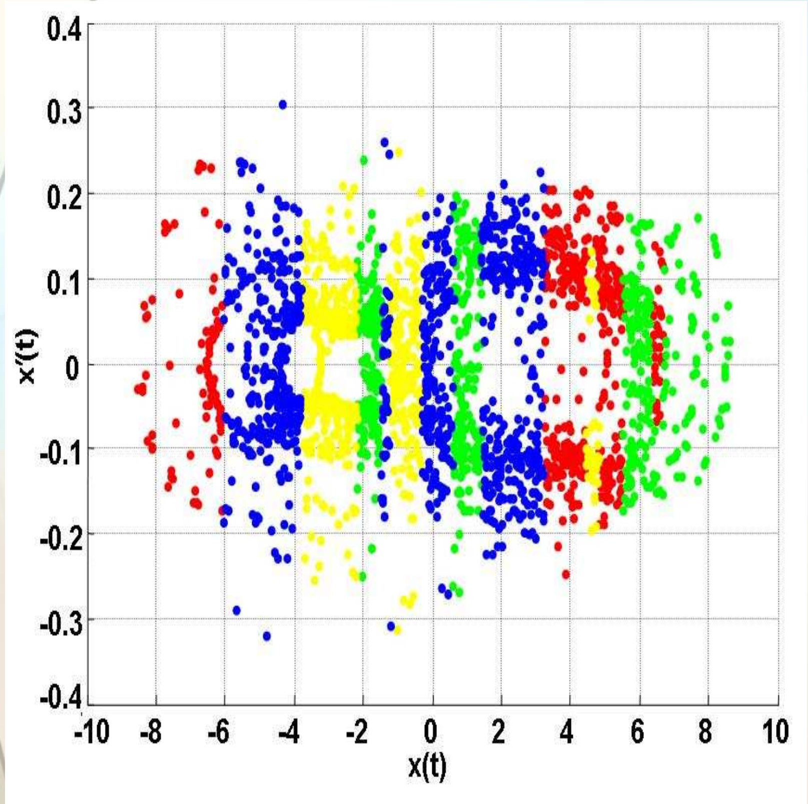
Proposed Scheme 1 : SFNN

The partitions are defined with respect to a set of radial-basis influence functions,

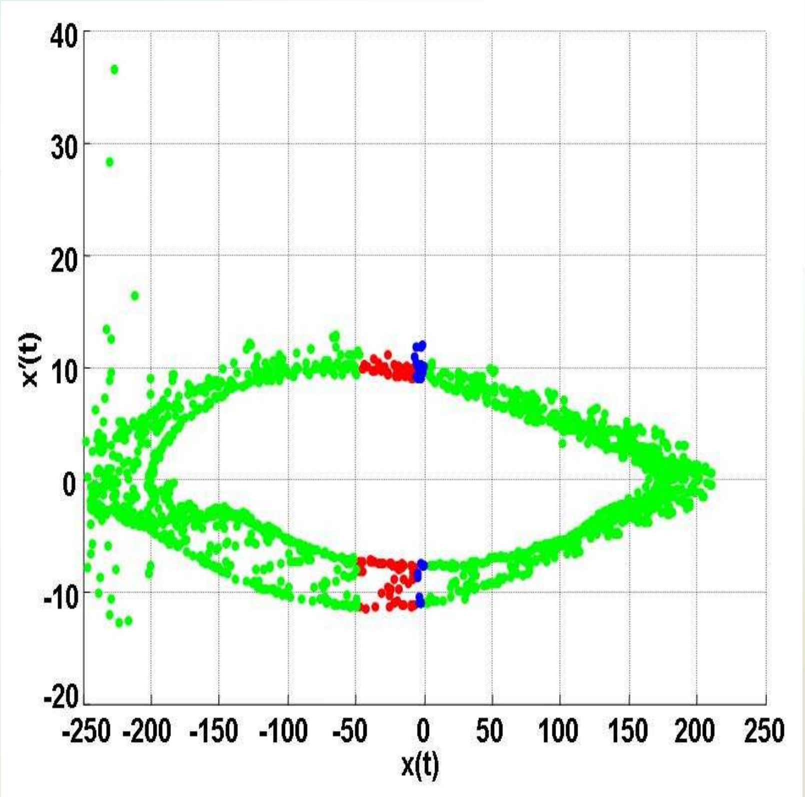
$$f_k(x) = \frac{\alpha_k}{\|x - z_k\|^2}$$

Each associated with a symbol s_k with the center z_k and weight α_k . For each element "x" of the time series data set, one $f_m(x)$ is generally expected to be greater than other $f_k(x)$ with $k \neq m$. Then, the data point "x" in the phase space is transformed to a symbol "s" in the symbol space. The parameters z_k and α_k are the free optimization variables, with the constraint $\alpha_k \geqslant 0$.

Partitioning using SFNN



Phase space of the nominal time series data



Phase space of the anomalous time series data

Methodology

- GE XTE-46 Gas Turbine Engine model
- To identify anomalous condition : measure the deviation of the efficiency values from nominal state (brand new engine).
- To replicate hundreds of hours of engine operation : accelerated testing approach
- The time series data used for analysis is the combustor outlet temperature.

Results

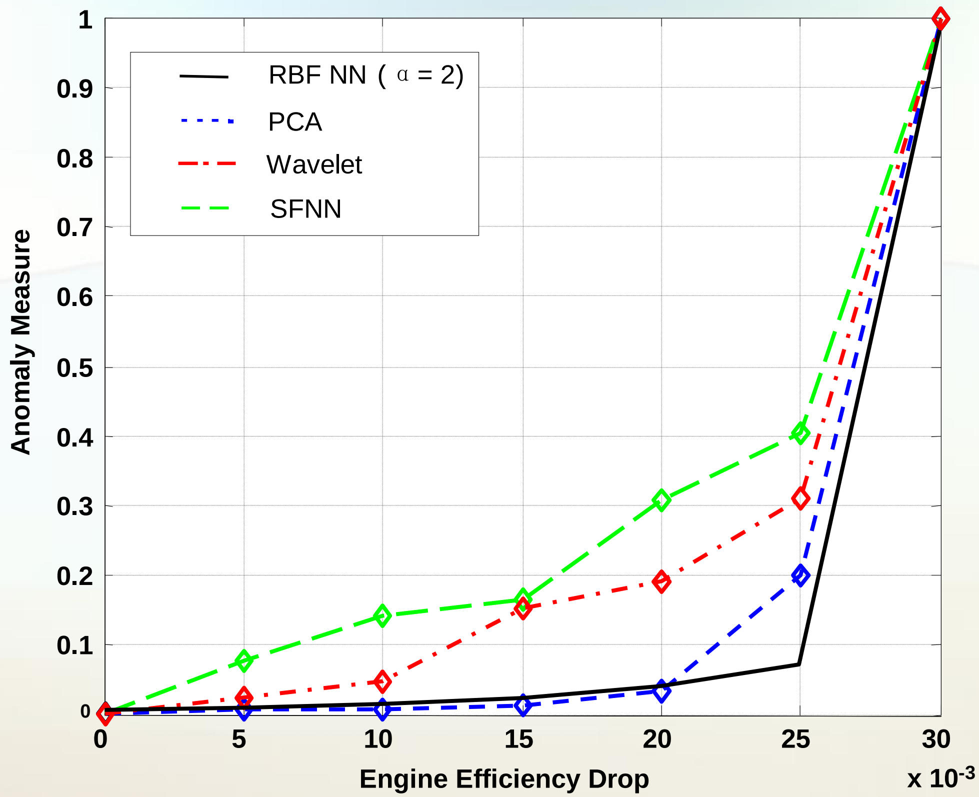


Figure compares the four different methods of anomaly detection

RBFNN and PCA based methods are comparatively inferior to the symbolic dynamics based methods in terms of early detection of anomalies.

Application 2: Electromechanical Systems

Externally Stimulated Duffing Equation with a single slowly varying parametric anomaly

Governing Equations:

$$\frac{d^2y(t)}{dt^2} + \theta(t_s)\frac{dy(t)}{dt} + y(t) + a\, y^3(t) = A\mathrm{Cos}(w\, t) \qquad t \in [t_o, \infty)$$

Random Initial Conditions

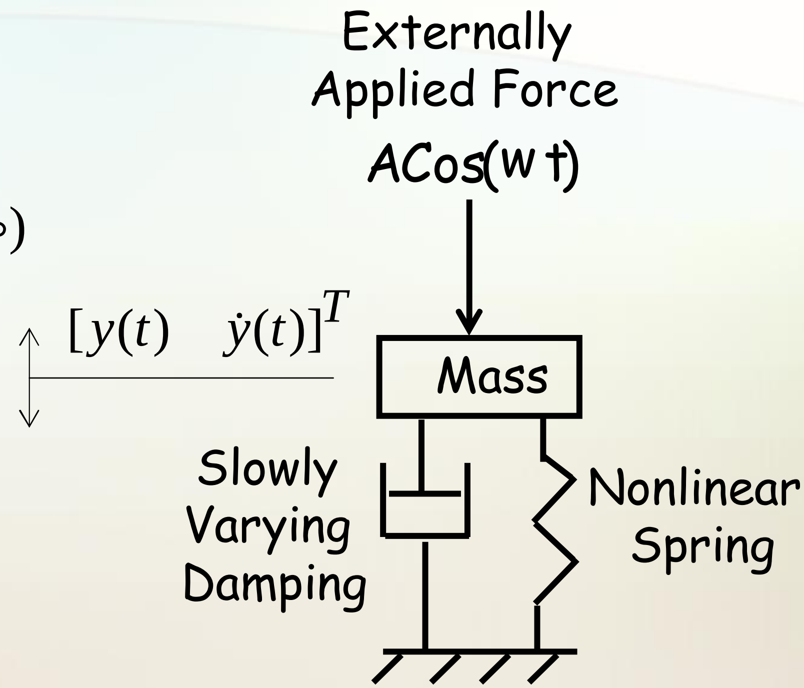
$$[y(t_o) \quad \dot{y}(t_o)]^T \in B_\delta(\underline{0})$$

t fast time $t \in [t_o, \infty)$

t_s slow time

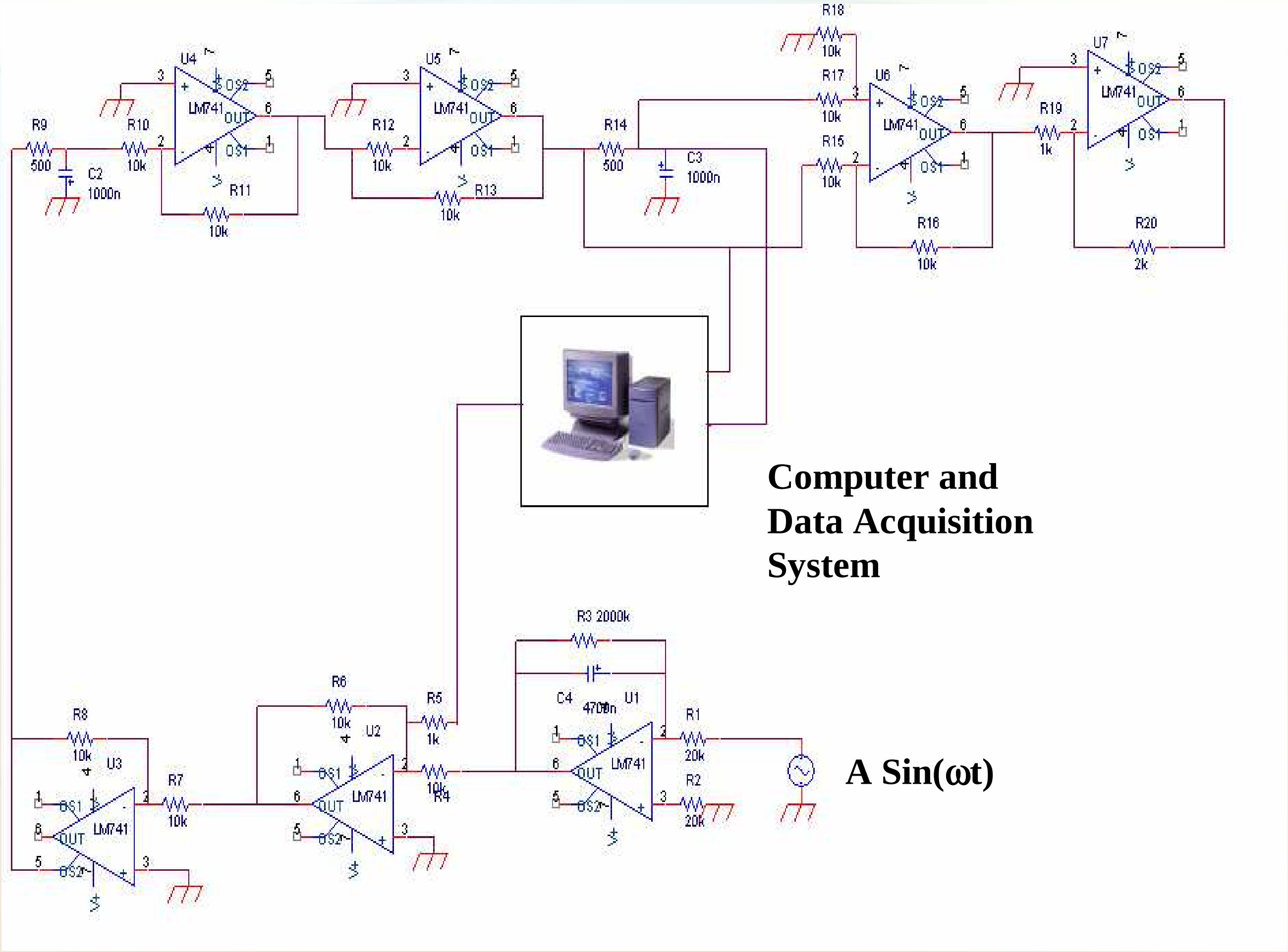
Parameters:

$a = 1;\ g = 0;\ A = 22;\ w = 5.0$



Anomaly Detection Apparatus for Hybrid Electronic Circuits

Externally Stimulated Duffing Equation

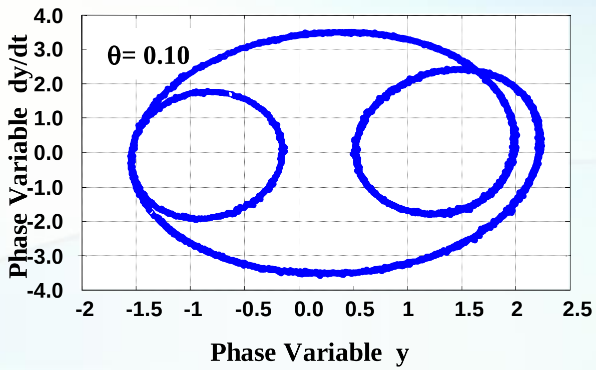


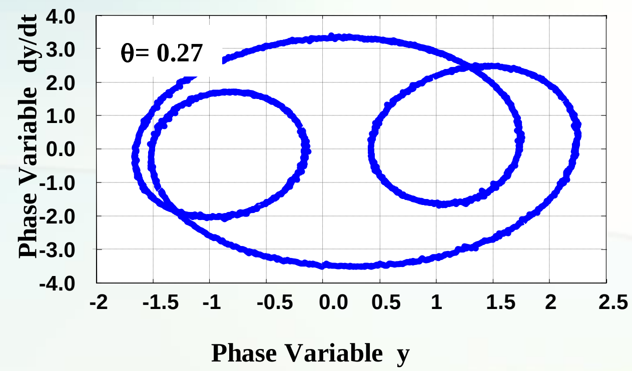
Computer and Data Acquisition System

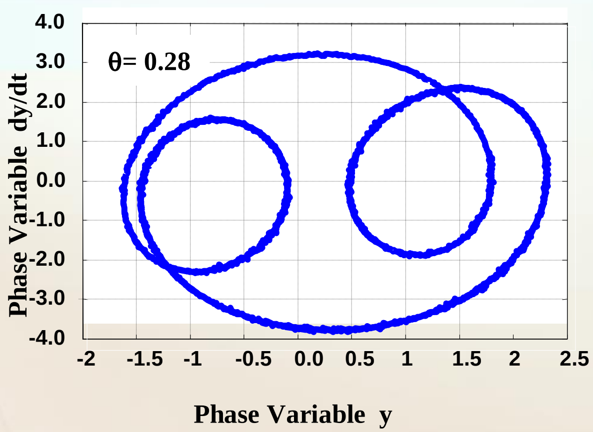
A Sin(ωt)

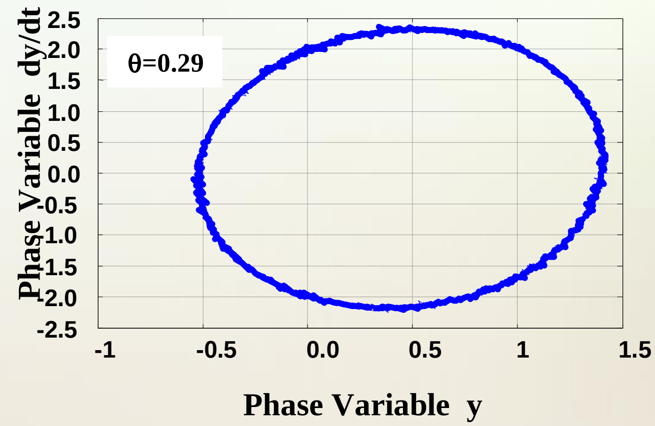
Electromechanical Systems Laboratory

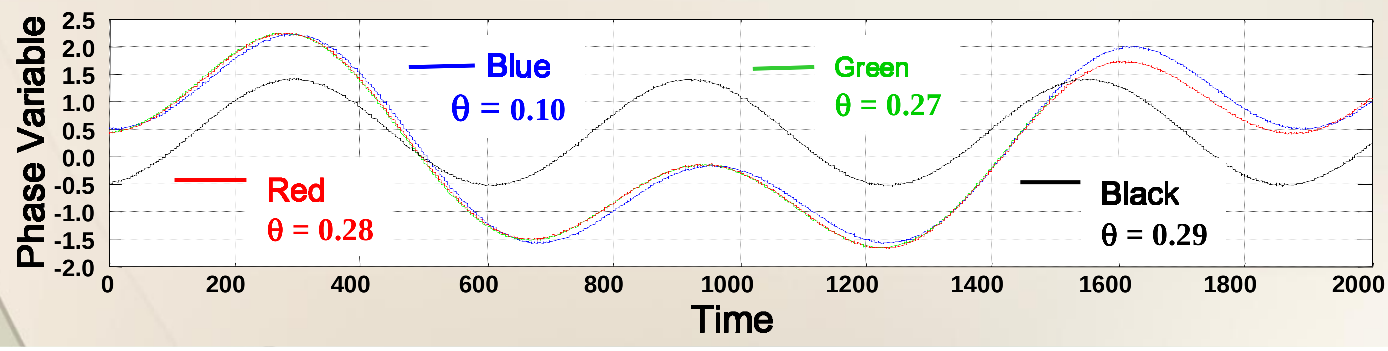
Phase-plane Plots under Nominal and Anomalous Conditions





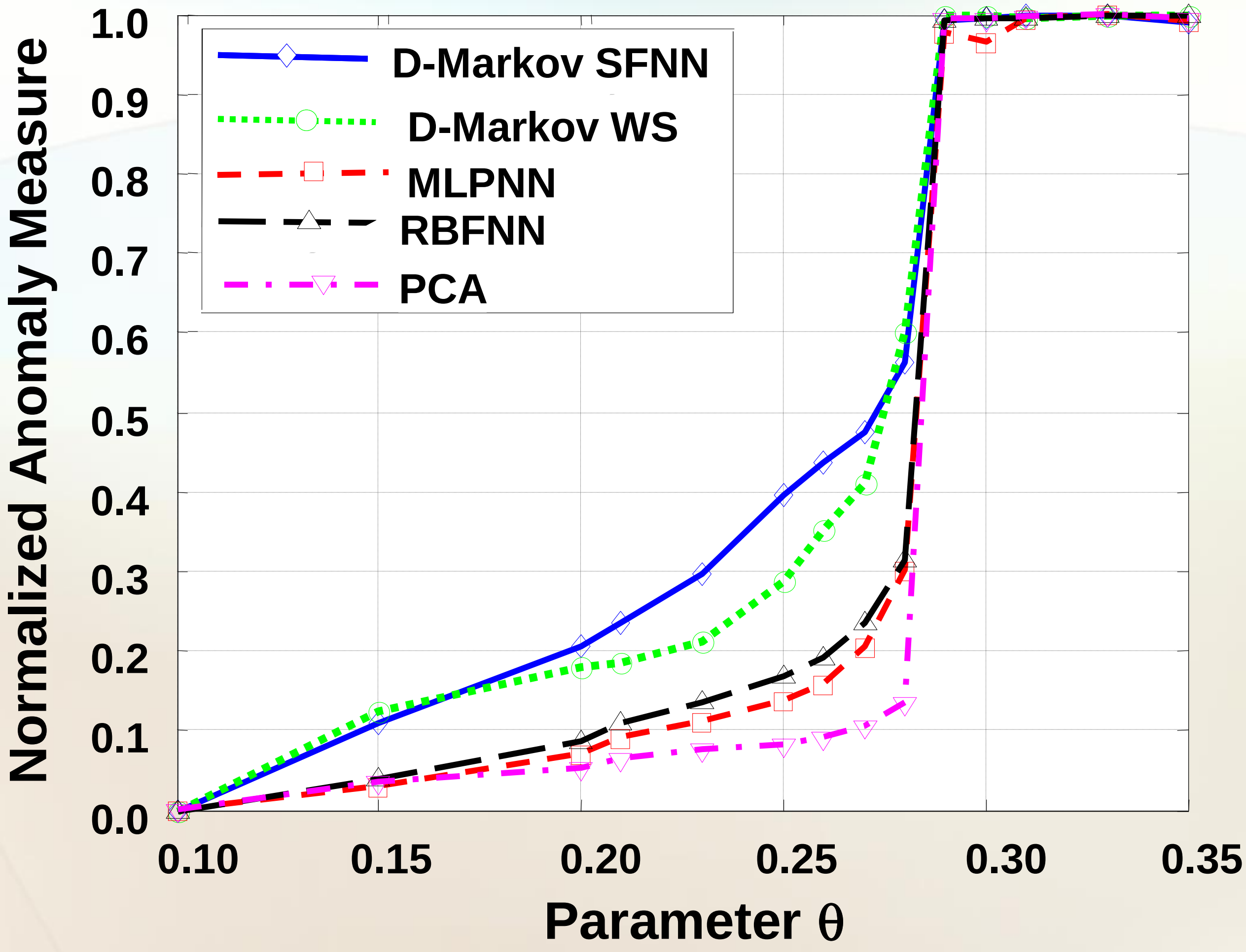




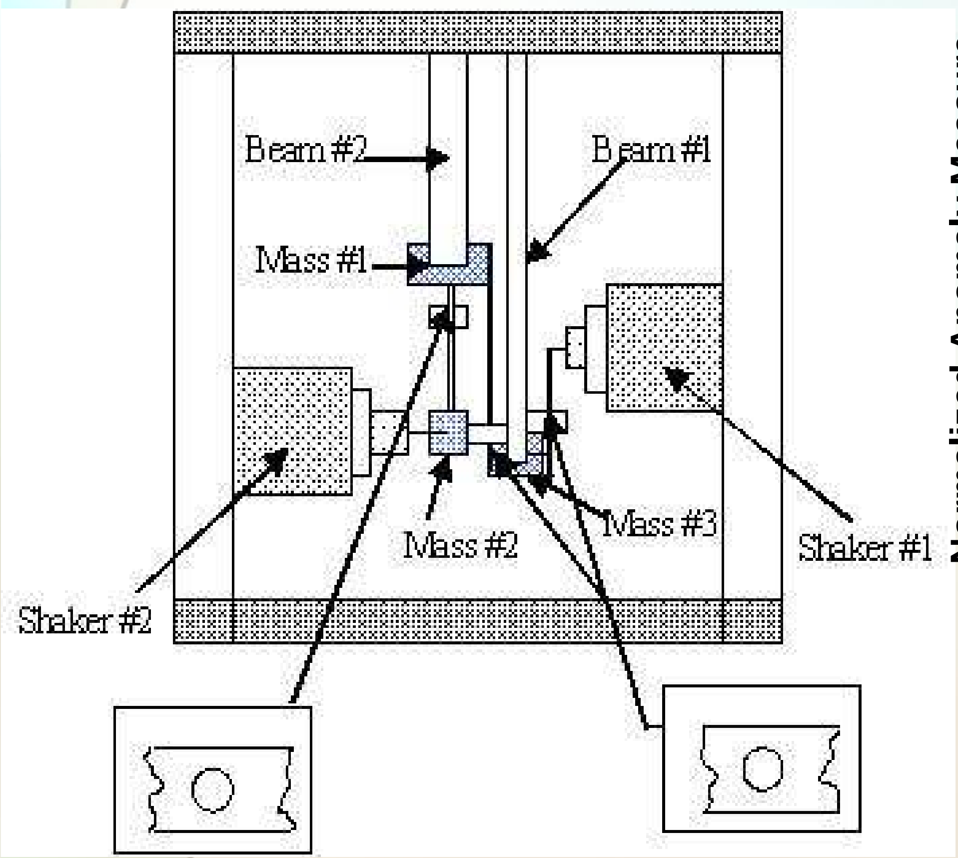


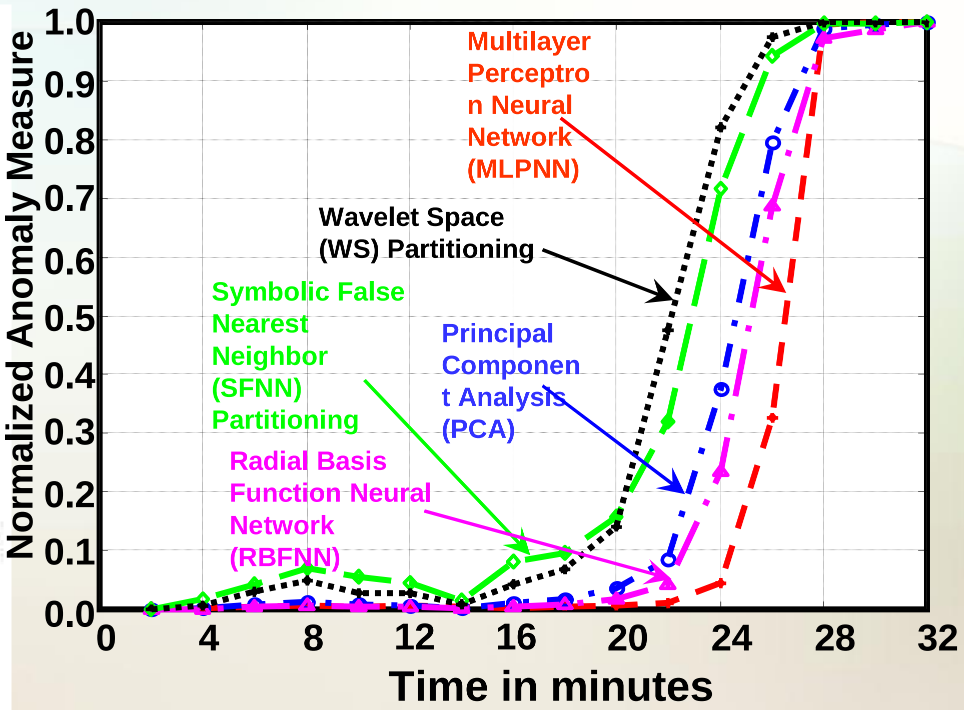
Electronic Circuit Apparatus

Sensitivity of the Detection Algorithm to the Anomalous Parameter θ



Application 3 (a) : Anomaly Detection Apparatus for Mechanical Vibration Systems





Application 3 (b) : Fatigue Test Apparatus for Damage Sensing in Ductile Alloys

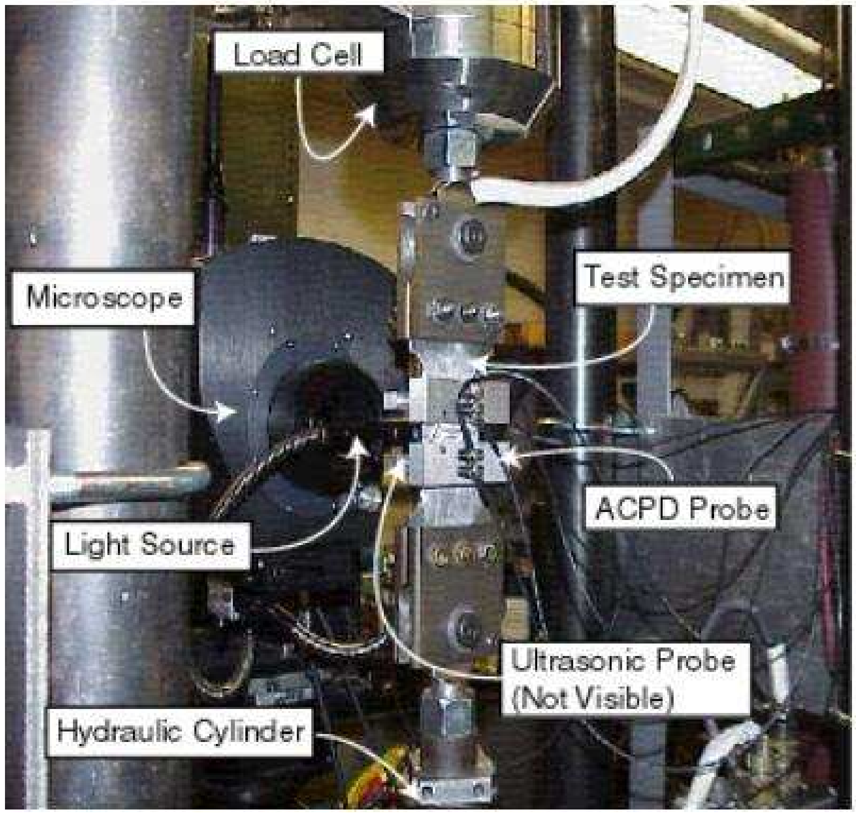
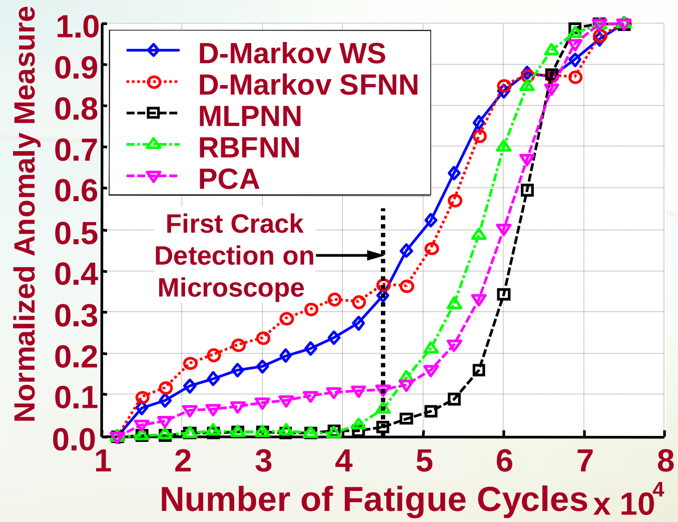
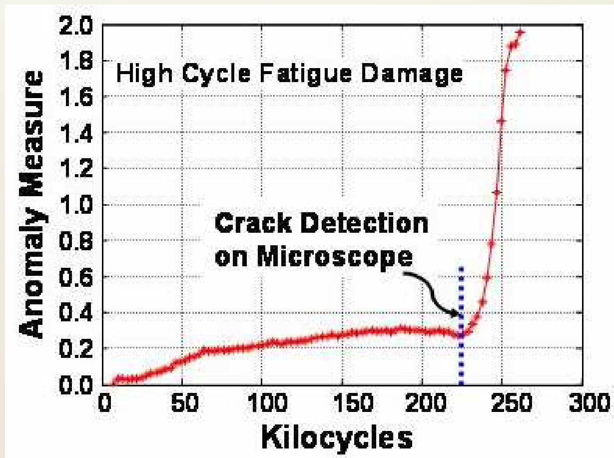
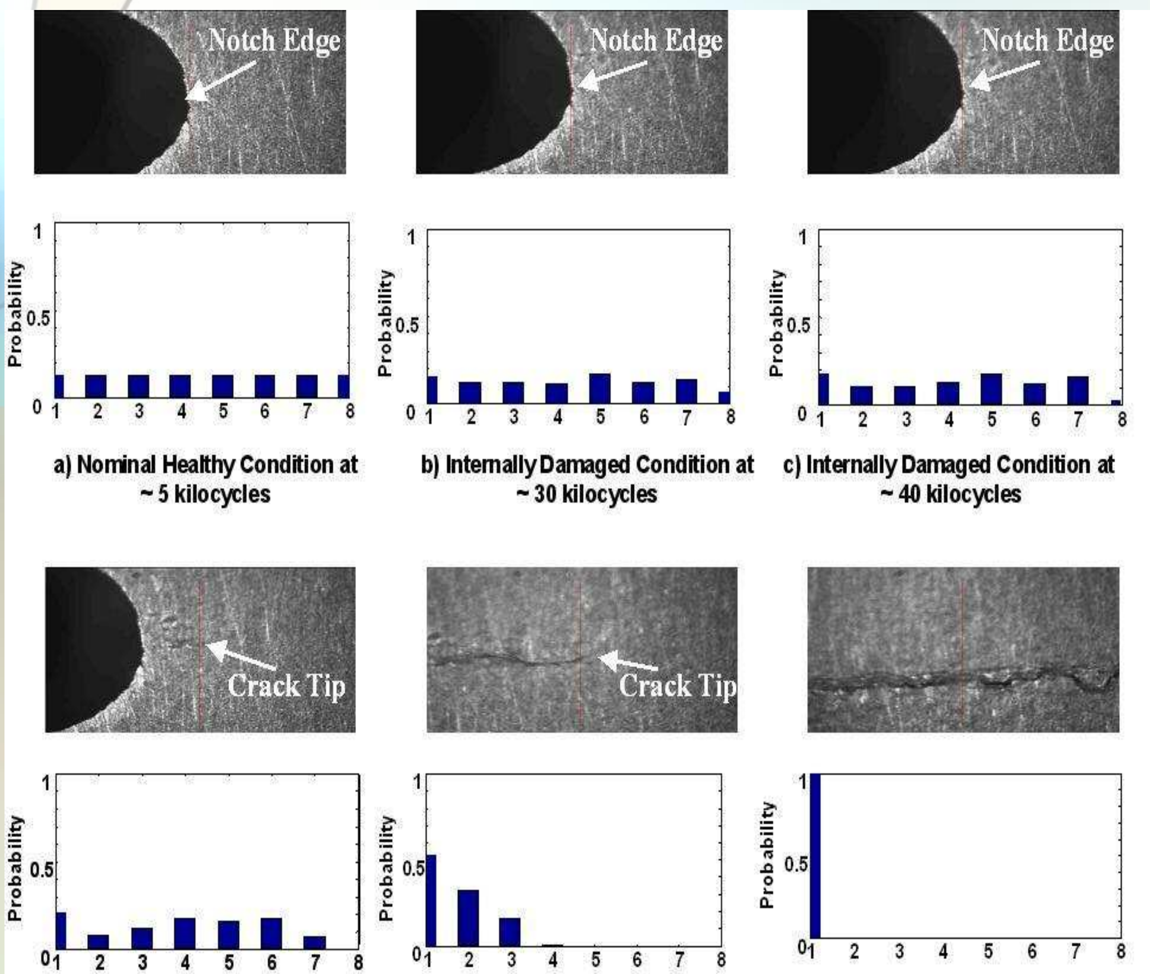


Figure A-1 Fatigue Damage Apparatus





Fatigue Crack Damage Sensing



a) Nominal Healthy Condition at ~ 5 kilocycles

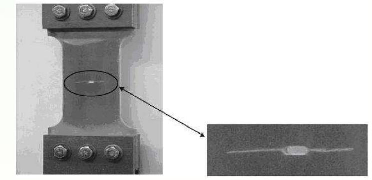
b) Internally Damaged Condition at ~ 30 kilocycles

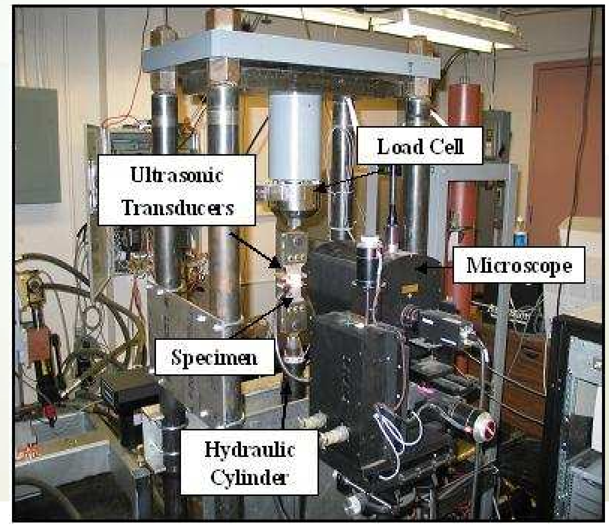
c) Internally Damaged Condition at ~ 40 kilocycles

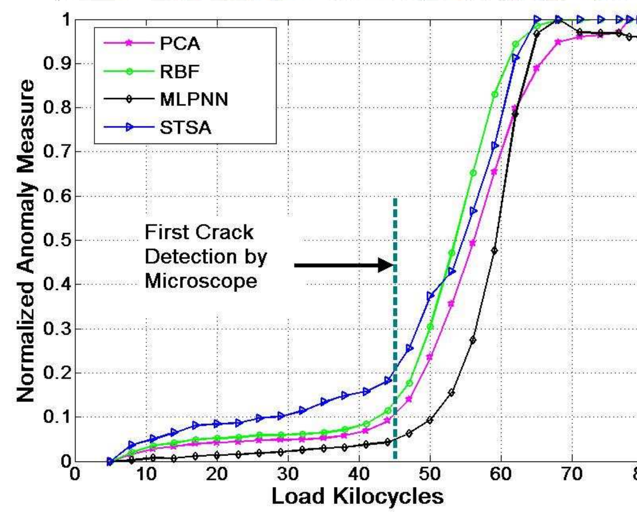
d) First detected appearance of surface crack at ~ 45 kilocycles

e) Fully Developed Crack at ~ 60 kilocycles

f) Completely Broken Specimen at ~ 78 kilocycles







Summary : Anomaly Detection
Symbolic Time Series Analysis

Advantages

- Foundations on basic principles of physics and mathematics
- Quantitative measure as opposed to qualitative measure
- Robustness to measurement noise and spurious signal distortion
- Adaptability to low-resolution sensing
- Applicability to real-time anomaly detection

(Near-term) Disadvantages

- Need for much theoretical and experimental research (especially in the area of optimal phase space partitioning)
- Seemingly counter-intuitive to inadequately trained technical personnel

Thank You