2005 American Control Conference
June 8-10, 2005. Portland, OR, USA

Incipient Fault Detection in Mechanical Power Transmission Systems†

Saurabh Bhatnagar sbhat@psu.edu
Venkatesh Rajagopalan vxr139@psu.edu
Asok Ray axr2@psu.edu
The Pennsylvania State University
University Park, PA 16802

***Abstract*— This paper presents a novel method for anomaly detection in a helical gear box, where the objective is to predict incipient faults before they become catastrophic. The anomaly detection algorithm relies on symbolic time series analysis and is built upon concepts from automata theory, information theory, and pattern recognition. Early detection of slow time-scale anomalous behavior is achieved by observing time series data at the fast time-scale of machine operation.**

I. INTRODUCTION

Most of the systems that we encounter in the real world, like aircraft engines and nuclear power plants are inherently complex. An exact dynamic model for such systems cannot be developed easily and this makes the study of fault-propagation in such systems, non-trivial. Catastrophic failures in complex systems can often be averted, if anomalies are detected sufficiently early in time before the actual failure. Anomalies are defined as deviations from the nominal behavior. In addition to safety and reliability, early detection of anomalies can also enhance performance and availability of the system.

Anomaly detection methodology adopted in this paper was introduced in [7]. This method is not model based, but relies on the time-series data obtained from the system. The detection methodology is formulated as a two time-scale problem:

1) The fast time scale i.e. the time scale of the process dynamics. The phase trajectories of the process evolve in the fast time scale.
2) The slow time scale: Anomaly, if any, propagates in the slow time scale.

The objective is to infer the occurrence of anomalies in slow time scale based on the changes in patterns in signals at the fast time scale. This paper validates the symbolic dynamics approach to anomaly detection for failure analysis in a helical gear box. The test apparatus is a helical gearbox which carries a load and is run by a 30 HP motor. Nonlinearities and partially correlated interactions between the components make the gear box, a complex system. An exact dynamic model does not exist for the gear-box system and hence it serves as a suitable test-bed for validating the proposed anomaly detection methodology.

This paper is organized into five sections including the present one. Section II provides a description of the test apparatus and the experiments. Section III presents the concept of anomaly detection in complex systems. Section IV discusses the experimental results. Section V summarizes the paper with recommendations for future research.

II. DESCRIPTION OF THE EXPERIMENT

The test bed consists of a helical gearbox, a 30 HP drive motor, and a 75 HP load motor as depicted in Figure 1. The gearbox is instrumented with accelerometers, thermocouples, torque cells and a microphone to monitor the dynamic behavior of the assembly. The location of the sensors is shown in Figure 2. Specifications of the gearbox are presented in Table II. In this paper, two test runs of the apparatus with varying loading conditions are presented. The gearbox was run at 100% output torque and HP for 96 hours and then overloaded and run on overload until failure. Time-series data was collected in 10 second windows at instances that were triggered by accelerometer rms thresholds. The sampling frequency was fixed at 20 KHz. The time-series data from the accelerometers has been used for analysis in this paper. For the first 96 hours, with gearbox running at 100% load, the samples were recorded every few hours. After the gearbox was overloaded, samples were taken at least once every 20 minutes or more frequently, since the response of the system changed rapidly and significantly. The gearbox ratio was fixed at $1 : 1.15$ reduction for all the experiments.

Fig. 1. The Experimental Setup

†This work has been supported in part by Army Research Office (ARO) under Grant No. DAAD19-01-1-0646.

TABLE I
THE GEARBOX SPECIFICATIONS

Make	Dodge APG
Model	R86001
Rated Input Speed	1750 rpm
Rated Output Torque	528 in-lb
Max. rated input HP	10
Gear Ratio	1.533
Contact Ratio	2.388

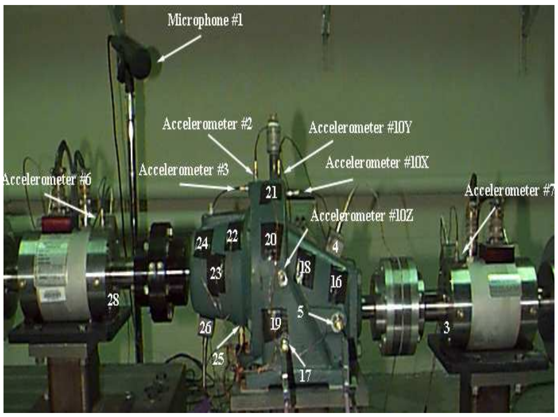

Fig. 2. Sensor locations

III. SYMBOLIC DYNAMICS BASED ANOMALY DETECTION

Anomaly detection in complex dynamical systems may be formulated as a two-time-scale problem, in which the system trajectories evolve in the fast time scale and anomalies evolve in the slow time scale. Anomalies may be parametric or non-parametric variations in the system response. The goal is to capture these variations by recognizing the patterns in the time series data as early as possible. Symbolic dynamics presents a useful tool for early detection of anomalies.

An important practical advantage of working with symbols is that the efficiency of numerical computations is greatly increased over what it would be for the original data. In some cases, efficiency may be mainly of value for reducing the need for computational resources, but it can also imply speed as well. The latter may be important for real-time monitoring and control applications. Also, analysis of symbolic data is often less sensitive to measurement noise. In some cases, symbolization can be accomplished directly in the instrument by appropriate design of the sensing elements. Such low-resolution (even disposable) sensors combined with appropriate analysis can significantly reduce instrumentation cost and complexity. Applications of symbolic methods are thus favored in circumstances where robustness to noise, speed, and/or cost are paramount [2].

A. Modeling of Dynamical Systems

Continuously varying physical processes are often modeled as a finite-dimensional dynamical system in the setting of an initial value problem as:

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = f(x(t,\theta)); \;\; x(0) = x_0 \tag{1}$$

where $t \in [0, \infty]$ is time; $x \in \mathbb{R}^n$ is the state vector in the phase space; and $\theta \in \mathbb{R}^m$ is the (possibly slowly varying) parameter vector. The vector field $f(x)$ is, in general, a locally Lipschtiz nonlinear operator and hence, a unique local solution to (1) exists [8].

Formally, such a solution can be expressed as a continuous function of the initial value x_0 as $x(t) = \Phi_t x_0$, where Φ_t represents a one-parameter family of maps of the phase space into itself. This evolution in time t can be viewed as a flow of points in the phase space. For a dissipative system, the flow may contract onto sets of lower dimension, known as attractors. The attractors represent the asymptotic behavior of the dynamical system, characterized by several phenomena (e.g.,fixed point,limit cycle, chaos and bifurcation) [6]. As mentioned previously, modeling a physical process with the mathematical structure of (1), may not always be feasible. A convenient way of learning the dynamical behavior is to rely on the (sensor-based) observed behavior. The need to extract relevant physical information about the observed dynamics has lead to development of nonlinear time series analysis (NTSA) techniques.

B. Encoding Time Series Data

Identification of nonlinear systems can be achieved using Formal Languages. In this paper, a discrete symbolic description is derived from the smooth dynamics in (1) [1]. Continuous time is first discretized based on an appropriate Poincare section P [5]. The resulting phase space of the map f associated with the flow Φ_t through P is divided into cells, so as to obtain a coordinate grid for the dynamics. For simplicity, a compact set $\Omega \in \mathbb{R}^n$ within which the motion is circumscribed,is identified with the phase space itself. The encoding of Ω is accomplished by introducing a partition $\mathbf{B} = (B_0, B_1, B_2, \ldots, B_{n-1})$ consisting of n mutually exclusive and exhaustive subsets, i.e., $B_j \bigcap B_k = \phi, \forall j \neq k$, $\bigcup_{j=1}^{n} B_j$ = Ω. The dynamical system describes an orbit $O = (x_0, x_1, x_2, \ldots, x_n, \ldots)$ which passes through or touches various elements of the partition B. Let us denote the index of domain $B_i \in \mathbf{B}$ visited at the time instant i as the symbol $\sigma_i \in \Sigma$. The set of symbols $\Sigma = (\alpha, \beta, \ldots, \phi, \chi)$ labeling the partition elements is called the alphabet. Each initial state x_0 generates a (infinite) sequence of symbols defined by a mapping as:

$$x_0 \longmapsto \alpha\beta\phi\chi\ldots \tag{2}$$

from the phase space to the space of symbols. Such a mapping is called Symbolic Dynamics as it attributes a legal (i.e., physically admissible) symbol sequence to the system dynamics starting from an initial state x_0. As the size of each cell is finite and the cardinality of the alphabet Σ is finite, any such symbol sequence represents, through iteration, a phase trajectory that has the compact support Ω.

The partition of Ω represented by the symbol alphabet Σ is called "generating partition" if an infinite symbol sequence $\alpha\beta\gamma\ldots$ determines a unique initial condition. That is, the mapping from the phase space to the space of symbol sequences is invertible.

In general, a dynamical system would generate only a subset of all possible sequences of symbols as there are many illegal (i.e., physically inadmissible) sequences. The grammar (i.e., set of rules) that determines legality of symbol strings in the alphabet Σ may change with the parameter vector θ. Symbolic dynamics can be viewed as coarse graining of the phase space and hence subjected to (possible) loss of information. However, the essential robust features (e.g., periodic behavior or chaotic behavior of an orbit) are expected to be preserved in the symbol sequences through an appropriate partitioning of the phase space [1].

Several methods of phase-space partitioning have been suggested in literature [3] [4]. There does not exist a standard procedure for phase-space partitioning of complex dynamical systems; this is a subject of active research. A scheme, used in this paper, for obtaining partitions is based on wavelet transform of time series data. This method was reported in [7]. The wavelet coefficients are a function of scale and time. After the wavelet transform is applied to the data, we partition the space of wavelet coefficients. These coefficients are stacked from end to end starting with the smallest value of scale and ending with the largest value. For example, the wavelet coefficients versus scale at time shift t_k are stacked after the ones at time shift t_{k-1}, to obtain the so-called scale series data in the wavelet space. The scale series data is analogous to the time series data in the phase space. The wavelet space is partitioned into horizontal slabs. The number of blocks in a partition is equal to the size of the alphabet and each block of the partition is associated with a symbol in the alphabet. The concept of proposed wavelet-space partitioning would require significant theoretical research before its acceptance for application to a general class of dynamical systems. Partitions are obtained with data collected at steady state nominal (no failure) conditions. The partitions remain invariant for all time epochs. As the fault propagates, small changes occur in the observed time-series data which in turn are reflected in the symbol sequences obtained from the data. The changes in patterns of the symbol sequences can be modeled with finite state machines.

C. Finite State Machine Construction and Anomaly Measure

Finite state machines, generated from the symbol sequences may be used for representing behaviorial patterns. This section presents the concept of finite state machine construction.

A probabilistic finite state machine is constructed from each symbol sequence, where the states of the machine are defined corresponding to the given alphabet Σ and window length D. The states are joined by edges labeled by a symbol in the alphabet. The state machine moves from one state to another upon occurrence of an event as a new symbol in the symbol sequence is received.

The states are chosen as words of length D from the symbol sequence, thereby making the total number of states to be equal to $|\Sigma|^D$. For machine construction, the window of length D is shifted to the right by one symbol upon receiving a new symbol , such that it retains the last $(D-1)$ symbols of the previous state and appends it with the new symbol in the end. The symbolic permutation in the current window gives rise to a new state that might be a different one or the same as the previous one (i.e., forming a self loop on that state). The entire state machine is constructed in this way. As an example, let us choose $D=2$ and $\Sigma=(0,1)$. Figure 3 elucidates the state machine construction for this specific example. The state transition matrix for this example will be a 4×4 matrix as shown below:

$$\left(\begin{array}{c|cccc} & 00 & 01 & 10 & 11 \\ \hline 00 & \pi_{11} & \pi_{12} & 0 & 0 \\ 01 & 0 & 0 & \pi_{23} & \pi_{24} \\ 10 & \pi_{31} & \pi_{32} & 0 & 0 \\ 11 & 0 & 0 & \pi_{43} & \pi_{44} \end{array}\right)$$

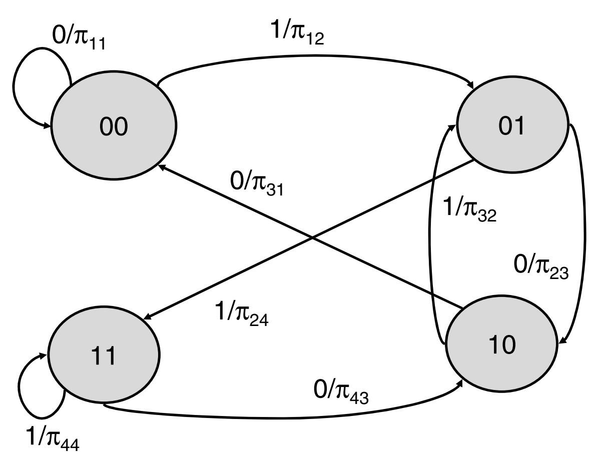

Fig. 3. D-Markov machine

The machines described above are capable of representing patterns in the behavior of a dynamical system under anomalous conditions. In order to quantify changes in the patterns, we induce a measure on these machines, denoted as an anomaly measure $\mathbf{M}$.

The state transition probabilities are dependent on the dynamics of the complex system as reflected in the symbol sequence. This is a factor in detecting an anomaly because perturbations in the system dynamics causes significant changes in the state transition probabilities, though such changes may also be dependent on the partitioning. The induced norm of the difference between the state transition matrices Π_0 and Π, respectively for the nominal and anomalous state machines G_0 and G, may be used as a measure of anomaly, i.e. $\mathbf{M}(G_0, G) = \|\Pi_0 - \Pi\|$. Such a measure, adopted in [7], was found to be very effective. Alternatively, a measure of anomaly may be derived directly from state probability vector $\mathbf{p}$. The state probability vector $\mathbf{p}$ is the vector, whose elements denote the probability of occurrence of states of the D-Markov machine. Since D-Markov machines have a fixed state structure, the state probability vector associated with all state machines will have the same dimension. Hence they may be used for vector representation of the anomaly, leading to the anomaly measure $\mathbf{M}(G_0, G) = \|\mathbf{p_0} - \mathbf{p}\|$. Another likely anomaly measure would be the angle between the probability vectors under nominal and anomalous conditions. This measure is defined as

$$\mathbf{M}(G_0, G) = \arccos\left(\frac{<\mathbf{p_0}, \mathbf{p}>}{\|\mathbf{p_0}\|_2 \, \|\mathbf{p}\|_2}\right), \tag{3}$$

where $< x, y >$ represents the inner product between the vectors x and y and $\|x\|_2$ represents the ℓ_2-norm of x [9].

IV. Results and Discussion

This section describes the details of the two test runs, lists and compares the observations made, and validates the anomaly detection technique in both the runs.

A. *Experimental Observations*

One test run of the system consists of fixing the reduction ratio, loading with appropriate output load, and running the system until failure occurs in the bearing or the gear. Reduction ratio being a constant across all test runs, the system undergoes the following steps in a single test run:

1) *Condition 1*: For the first 96 hours, the gearbox is run at 100% output i.e. 540 in-lb. Input speed is 1750 rpm and input power is 9.8 HP(98% of the maximum). This is called the break-in condition for the gear-box. The system runs at its rated output, and hence the propagation of failure is minimal.
2) *Condition 2*: After the first 96 hours, the output torque is increased to a higher value. This accelerates the failure of the system. The gear-box is run in this condition till a failure occurs in the gear or the bearing.
3) *Failure*: Allowing the system to run in condition two leads to failure. The gears and bearing are then analyzed to infer the site and mode of failure.

The details of the test runs have been listed below.

1) Test Run 1:
    a) *Condition 1*: Input speed $= 1750$ rpm, output torque $= 540$ in-lbs (100% rated), Power $=$ 9.8 HP(98% rated)
    b) *Condition 2*: Input speed $= 1750$ rpm, output torque $= 1080$ in-lbs (200% rated), Power $=$ 20 HP (200% rated)
    c) Time of run $= 96$ hours (Condition 1) and 51.5 hours (Condition 2)
2) Test Run 2:
    a) *Condition 1*: Input speed $= 1750$ rpm, output torque $= 540$ in-lbs (100% rated), Power $=$ 9.8 HP (98% rated)
    b) *Condition 2*: Input speed $= 1750$ rpm, output torque $= 1620$ in-lbs (300% rated), Power $=$ 30 HP (300% rated)
    c) Time of run= 96 hours (Condition 1) and 31.4 hours (Condition 2)

Fig. 4. Failure in the drive gear for test run 1

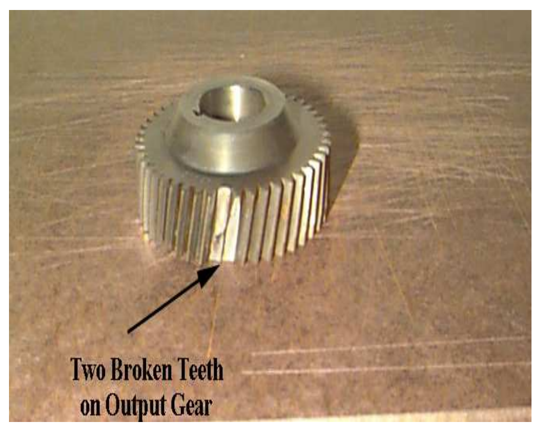

Fig. 5. Failure in the drive gear for test run 2

B. *Failure Analysis in two test runs*

1) Test Run 1 : As can be clearly seen in the Figure 4, the drive gear suffered two broken teeth. However the pinion gear and the bearing suffered no visible damage.
2) Test Run 2 : As shown in Figure 5, the drive gear suffered two broken teeth. However, as expected the damage was greater in the test run 2, as a root crack was also visible in the drive gear. There were again no visible marks on the pinion gear or the bearing.

C. Interpretation of Anomaly Measure Curve

As required by the proposed technique, this experiment clearly has two time-scales. The fast time scale is defined on time intervals between data snapshots and is of the order of 10 seconds. The slow time scale is defined on time intervals between successive data collections and is of the order of 20 minutes to few hours. The fault or anomaly as required propagates in the order of minutes, at the slower time scale.

The experiment is a rich source of time-series data as the apparatus is equipped with more than 40 sensors. For the purpose of validation of the proposed technique, one of the sensors is chosen at a time as the data source. Presented in this section are the results using data from one of the accelerometers. Results using data from other sensors are similar in nature, and hence not presented here.

The first step in the proposed technique is to determine the partitioning, to convert the time-series data into a symbol sequence. The partitioning procedure explained in section III was used to obtain the symbol sequence with desired alphabet size. After choosing an alphabet size of 12 and accordingly generating the symbol sequence, the D-Markov machine is constructed. The depth, D is chosen as 1, and hence the D-Markov machine constructed has $A^D = 12$ states. Nominal condition is defined with the system started at the 100% load. Anomaly measure $\mathbf{M}(G_0, G) = \arccos(\frac{<\mathbf{p_0},\mathbf{p}>}{\|\mathbf{p_0}\|_2 \|\mathbf{p}\|_2})$ was used to quantify the fault. Plots of the anomaly measure for the two test runs are shown in Figure 6 and Figure 7 respectively. As an example, the state probability vectors for test run one at nominal condition, i.e., t = 0, and anomalous condition at $t = 100$ hours are presented here.

$\mathbf{p}(t = 0)$	$\mathbf{p}(t = 100\,\text{hours})$
0.066000	0.103530
0.081333	0.076000
0.084933	0.066400
0.082933	0.064870
0.090733	0.072733
0.086667	0.083400
0.099067	0.095400
0.092400	0.091800
0.091133	0.079200
0.085000	0.074267
0.075333	0.081000
0.064467	0.111400

TABLE II
PROBABILITY VECTORS AT $t = 0, 100$ HOURS

As can be easily seen, the two vectors are significantly different and the anomaly measure at $t = 100$ hours is calculated to be $\mathbf{M} \approx 0.25$.

As can be easily seen from both the anomaly curves, the anomaly measure is a non-decreasing function of time. This is expected for a system proceeding towards failure. An important observation can be made at $t = 96$ hours in both the curves. As stated earlier, the gearbox was being run at 100% load and HP for first 96 hours and beyond that the

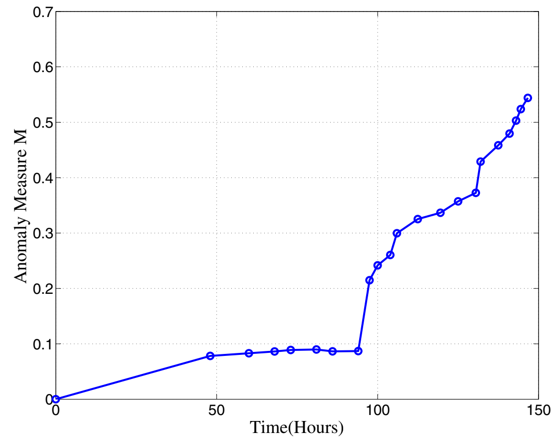

Fig. 6. Anomaly Measure Plot for test run 1

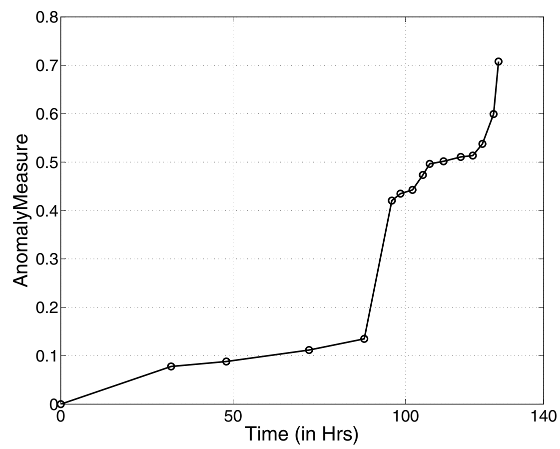

Fig. 7. Anomaly Measure Plot for test run 2

load was increased to accelerate failure. As seen from the graph, till $t = 96$ hours, the anomaly measure is around 0.1 (i.e. $< 20\%$ of the value around failure). This indicates that the system is far from failure. At around 96 hours, there is a steep rise in both the anomaly measure curves (the anomaly measure more than doubles its value before $t = 96$ hours). This can be attributed to the abrupt change in the output load (200% in the run 1 and 300% in the run 2). If we compare the two graphs at $t = 96$ hours, it is seen that the rise in anomaly measure is higher for the run 2, where a greater change in the loading conditions was introduced. Thus, the proposed algorithm is able to detect an abrupt change in the loading conditions immediately.

Comparing the two curves beyond $t = 96$ hours, it is seen that the system failure occurs at $t = 127.4$ hours for test run two and $t = 149.5$ hours for the test run one. This was as expected as the gearbox was loaded to 300% of the rated load in the second run compared to 200% in the first. The failure in the second run is thus expected to occur earlier with more damage to the system. This can be easily verified from the anomaly plots. The rise of the anomaly curve for second run in Figure 7 is much steeper than that for the second run in Figure 6. Thus, the rise of anomaly measure curve is proportional to the rate of damage in the system, i.e., the greater the rate of damage to the system, the greater the slope of the curve. As seen above in the analysis of the faults occurring in the gears, test run 2 resulted in a greater damage to the system. This can again be verified by checking the final value of the anomaly measure in the two cases. Indeed, the final value of anomaly measure for test run 2 (0.71) is higher than that for test run 1 (0.55). Thus, for both the cases, the proposed algorithm for anomaly detection is able to identify the impending failure of the system at an early stage. The anomaly measure curve also gives a clear indication of the rate of damage and the total damage to the system on failure.

V. Conclusion and future work

This paper presented an approach for early detection of anomalies that occur in slow time scale by observing the process dynamics in the fast time scale. To verify the technique, sensor data from a helical gear box was used. The time series data from accelerometer attached to the gear-box was collected to create respective strings of symbols using the optimal phase space partitioning technique. A fixed structure finite state machine with varying state probabilities was generated from these symbol strings at each stage. The state probability vector was considered a representation of the stationary behavior of the system at various times. The angle between the vector at any time with the nominal state probability vector gave a measure of the anomaly at that time. The main conclusion of this research work is that symbolic dynamics based anomaly detection is an effective tool for early detection. It can also detect abrupt changes in loading conditions so that mitigating action can be taken in time to avoid failure. The slope of the anomaly measure curve indicates the damage rate inflicted upon the system. It was verified that the greater the final value of the anomaly measure, the greater is the total damage, hence we conclude that anomaly measure is indeed a measure of the extent of damage of the system.

Future work would involve the comparison of this technique with other techniques like the principal component analysis (PCA) and the neural network multi-layer perceptron (NNMLP). One criterion for comparison would be how early each algorithm can detect the anomaly. Future work will also include implementation of this technique in real time and developing control techniques for taking mitigating action to extend the life of the system with minimal loss in performance. Another important task would be to use the data from more than one sensors at a time. A combination of data from various sensors could also be used to differentiate between the various kinds of failures in the system. This may ultimately lead to the development of a self-healing system.

VI. Acknowledgements

The authors wish to acknowledge the Applied Research Laboratory at Pennsylvania State University for providing the experimental data for analysis.

References

[1] R. Badii and A. Politi, *Complexity hierarchical structures and scaling in physics*, Cambridge University Press, United Kingdom, 1997.

[2] C.S.Daw, C.E.A. Finney, and E.R. Tracy, *A review of symbolic analysis of experimental data*, http://www-chaos.engr.utk.edu/pap/crg-rsi2002.pdf, July 2002.

[3] R.L. Davidchack, Y.C. Lai, E.M. Bolt, and H. Dhamala, *Estimating generating partitions of chaotic systems by unstable periodic orbits*, no. 75, Physical Review, 2000.

[4] M.B. Kennel and M. Buhl, *Estimating good discerte partitions form observed data: symbolic false nearest neighbors*, http://arxiv.org/PS cache/nlin/pdf/0304/0304054.pdf, 2003.

[5] H.K Khalil, *Nonlinear systems*, Prentice-Hall Inc, Englewood Cliffs, N.J., 1996.

[6] E. Ott, *Chaos in dynamical systems*, Cambridge University Press, United Kingdom, 1993.

[7] A. Ray, *Symbolic dynamic analysis of complex systems for anomaly detection*, Signal Processing, Vol. 84 No.7, pp 1115-1130, 2004.

[8] M. Vidyasagar, *Nonlinear systems analysis*, Prentice Hall, Englewood Cliffs, N.J., 1993.

[9] V.Rajagopalan, R.Samsi, A.Ray, J.Mayer, and C.Lagoa, *A symbolic dynamics approach for early detection of slowly evolving faults in nonlinear systems*, IASTED-CSS, Nov-Dec 2004.