*For review for presentation in **ACC 2012**, Montreal, Canada*

Symbolic Transient Time-series Analysis for Fault Detection in Aircraft Gas Turbine Engines★

Soumalya Sarkar† svs5464@psu.edu
Kushal Mukherjee† kum162@psu.edu
Soumik Sarkar‡ sarkars@utrc.utc.com
Asok Ray† axr2@psu.edu

† *Department of Mechanical Engineering*
The Pennsylvania State University
University Park, PA 16802, USA

‡ *United Technologies Research Center, East Hartford, CT, USA*

Keywords: *Statistical Pattern Classification; Symbolic Dynamics; Probabilistic Finite State Automata*

***Abstract*— This paper focuses on data-driven detection of incipient fault in commercial aircraft gas turbine engines. Detection of incipient engine fault often manifest better in transient data. This paper extends recently reported literature in the areas of symbolic dynamic filtering, i.e., Markov model based analysis of steady state data, to model and analyze transient data generated during the take-off phase. The fault detection and classification algorithms are validated on the NASA C-MAPSS transient test case generator.**

I. Introduction

Condition-based maintenance of aircraft gas turbine engines is critical for aviation safety and reliability. Engine performance monitoring is typically performed on steady-state data collected during cruise conditions. However, in a time and safety-critical operation like this, understanding transient data from takeoff, climb or landing is extremely important. Furthermore, incipient fault detection during the transient operations can significantly reduce the costs related to in-flight shutdowns, unscheduled engine removals, or take-off aborts. Engines operate under much higher stress and temperature conditions under these circumstances compared to the low-stress cruise phase. Certain incipient engine fault (e.g., bearing faults, controller miss-scheduling, starter system faults) signatures magnify during the transient conditions [1]. Moreover, the effects of feedback control in suppressing sensor and component faults is minimal during transient conditions, increases the possibility of detecting any fault in the system [2]. Several model-based and data-driven fault diagnosis studies have been done using transient data. A neural network based fault diagnosis method was developed in [3] for automotive transient operations. In [4], adaptive Myriad filter has been used to improve the quality of transient data for gas turbine engines. Often model-based diagnostics appear to be difficult due to lack of reliability of the transient models. Data-driven diagnostics using Hidden Markov Models (HMMs) have been performed for transient gas turbine engine operation [5].

A recently developed data-driven technique, called the symbolic dynamic filtering (SDF) [6], have been shown to yield superior performance in terms of early detection of anomalies and robustness to measurement noise in comparison with other existing techniques such as Principal Component Analysis (PCA), Neural Networks (NN) and Bayesian techniques [7]. Recently, in a two-part paper [8][9], an SDF-based algorithm for detection and isolation of engine subsystem faults (specifically, faults that cause efficiency degradation in engine components) has been reported and an exten-

★This work has been supported in part by NASA under Cooperative Agreement No. NNX07AK49A, the U.S. Army Research Laboratory and the U.S. Army Research Office under Grant No. W911NF-07-1-0376, and by the U.S. Office of Naval Research under Grant No. N00014-09-1-0688. Any opinions, findings and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the sponsoring agencies.

sion of that work to estimate simultaneously occurring multiple component-level faults has been presented in [10]. Furthermore, an optimized feature extraction technique has been developed under the same semantic framework in [11]. However, all of the above studies were done on steady-state cruise flight data, that also conformed with the quasi-stationary data assumption made in SDF. Due to this assumption, SDF potentially could not handle transient data (and of limited length). The goal of this paper is to extend SDF's capability to be able to handle transient data. In this context, Dirichlet and multinomial distributions have been used to construct the *a priori* and *a posteriori* models of uncertainties, respectively. The algorithms are formulated by quantitatively incorporating the effects of finite-length symbol strings in both training and testing phases of fault detection. The resulting algorithm is validated using the *Transient Test-case Generator* [12] of the Commercial Modular Aero Propulsion System Simulation (C-MAPSS) test bed, developed by NASA.

The paper is organized in five sections including the present one. Section II explains the symbolic framework of transient time-series analysis along with necessary background information. Section III describes the C-MAPSS test bed [13] along with the sensors and the fault injection scheme. Section IV presents the results of case studies to validate the proposed method on the C-MAPSS test bed. Finally, the paper is summarized and concluded in Section V with recommendations of future work.

II. Symbolic Analysis of Time Series Data

A symbol string is obtained from the output of a dynamical system (a gas turbine engine) by partitioning (also called quantization) of the time-series data. Thereafter, a probabilistic finite state automaton (PFSA) is constructed from the (finite-length) symbol sequence via one of the construction algorithms (e.g., [6], [14], and [15]). Due to the quasi-stationarity assumption in SDF, it may not be feasible to obtain sufficiently long strings of symbols in both training and testing phases of classification. Therefore, the estimated parameters of the resulting PFSA model may not be precise.

The goal of this section is to construct a Bayesian classifier for identification of the probability morph matrices of PFSA based on the transient data in both training and testing phases. The Dirichlet and multinomial distributions have been used to construct the *a priori* and *a posteriori* models of uncertainties, respectively. This formulation by quantitatively incorporates the effects of finite-length symbol strings in both training and testing phases of pattern classification.

A. Partitioning of time series data

The sensor time series is encoded by introducing partitions in the range of the signal. This step enables transformation of the sensor data from the continuous domain to the symbolic domain. In other words, the real valued sensor data (at each time step is replaced by a corresponding symbol from a set Σ (called the alphabet set)

B. Modeling via Probabilistic Finite State Automaton

The symbolic sequence is modeled as a probabilistic finite state automaton (PFSA). A PFSA is a tuple $G \triangleq (Q, \Sigma, \delta, \Pi)$. The alphabet Σ is a nonempty finite set of symbols. The set of states Q is nonempty and finite. As a simplifying assumption, this paper considers only a class of PFSAs know as D-Markov machines[6]. In D-Markov machines, the states are strings of the D past symbols. The number D is called the depth of the machine and the number of states $|Q| = |\Sigma|^D$. The state transition function $\delta : Q \times \Sigma \to Q$ indicates the new state given the previous state and an observed symbol. In addition, the morph function $\pi : Q \times \Sigma \to [0, 1]$ is an output mapping that satisfies the condition: $\sum_{\sigma \in \Sigma} \pi(q, \sigma) = 1$ for all $q \in Q$. The morph function π has a matrix representation Π, called the (probability) morph matrix $\Pi_{ij} \triangleq \pi(q_i, \sigma_j), \forall q_i \in Q$ and $\forall \sigma_j \in \Sigma$. Note that Π is a $(|Q| \times |\Sigma|)$ matrix where each element of Π is non-negative and each row sum of Π is equal to 1.

C. The Online Classification Problem

Let there be K symbolic systems of interest, denoted by $C_1, C_2, \ldots, C_K$, over the same alphabet Σ and each class C_i is modeled by an ergodic (i.e., irreducible) PFSA $G^i = (Q^i, \Sigma, \delta^i, \Pi^i, q_0^i)$, where $i = 1, 2 \ldots, K$.

During the training phase, a symbol string $S^i \triangleq s_1^i s_2^i \ldots s_{N_i}^i$ is generated from each class C_i. The state transition function δ of the D-Markov machine is fixed by choosing an appropriate depth D. Thus, Π^i's become the only unknowns and could be selected as the feature vectors for the purpose of classification. The distribution of the morph matrix Π^i is computed in the training phase from the finite length symbol sequences for each class.

In the testing phase, let another symbol string $\widetilde{S}$ be obtained from a sensor time series data. Then, the task is to determine which class this observed symbol string $\widetilde{S}$ belongs to. While the previous work [6][14][15] has aimed at identification of a PFSA from a given symbol string, the objective of this paper is to imbed the uncertainties due to the finite length of the symbol string in the identification algorithm that would influence the final classification decision.

In the training phase, each row of Π^i is treated as a random vector. Let the m^{th} row of Π^i be denoted as Π^i_m and the n^{th} element of the m^{th} row as $\Pi^i_{mn} \geq 0$ and $\sum_{n=1}^{|\Sigma|} \Pi^i_{mn} = 1$. The *a priori* probability density function $f_{\Pi^i_m|S^i}$ of the random row-vector Π^i_m, conditioned on a symbol string S^i, follows the Dirichlet distribution [16] [17] as described below.

$$f_{\Pi^i_m|S^i}(\boldsymbol{\theta}^i_m|S^i) = \frac{1}{B(\boldsymbol{\alpha}^i_m)} \prod_{n=1}^{|\Sigma|} (\theta^i_{mn})^{\alpha^i_{mn}-1} \quad (1)$$

where $\boldsymbol{\theta}^i_m$ is a realization of the random vector Π^i_m, namely,

$$\boldsymbol{\theta}^i_m = \begin{bmatrix} \theta^i_{m1} & \theta^i_{m2} & \dots & \theta^i_{m|\Sigma|} \end{bmatrix}$$

and the normalizing constant is

$$B(\boldsymbol{\alpha}^i_m) \triangleq \frac{\prod_{n=1}^{|\Sigma|} \Gamma(\alpha^i_{mn})}{\Gamma(\sum_{n=1}^{|\Sigma|} \alpha^i_{mn})} \quad (2)$$

where $\Gamma(\bullet)$ is the standard gamma function, and $\boldsymbol{\alpha}^i_m = \begin{bmatrix} \boldsymbol{\alpha}^i_{m1} & \boldsymbol{\alpha}^i_{m2} & \dots & \boldsymbol{\alpha}^i_{m|\Sigma|} \end{bmatrix}$ with

$$\alpha^i_{mn} = N^i_{mn} + 1 \quad (3)$$

where N^i_{mn} is the number of times the symbol σ_n in S^i is emanated from the state q^i_m, i.e.,

$$N^i_{mn} \triangleq \left|\{(s^i_k, v^i_k) : s^i_k = \sigma_n, v^i_k = q^i_m\}\right| \quad (4)$$

Recalling that s^i_k is the k-th symbol in S^i, and denoting the number of occurrence of the state q^i_m in the state sequence $\mathbb{V}^i \setminus \{v^i_{N^i}\}$ as $N^i_m \triangleq \sum_{n=1}^{|\Sigma|} N^i_{mn}$, it follows from Eqs. (2) and (3) that

$$B(\boldsymbol{\alpha}^i_m) = \frac{\prod_{n=1}^{|\Sigma|} \Gamma(N^i_{mn}+1)}{\Gamma(\sum_{n=1}^{|\Sigma|} N^i_{mn} + |\Sigma|)} = \frac{\prod_{n=1}^{|\Sigma|} (N^i_{mn})!}{(N^i_m + |\Sigma| - 1)!} \quad (5)$$

by use of the relation $\Gamma(n) = (n-1)! \ \ \forall n \in \mathbb{N}_1$.

By the Markov property of the PFSA G^i, the $(1 \times |\Sigma|)$ row-vectors, $\{\Pi^i_m\}, m = 1, \dots |Q|$, are statistically independent of each other. Therefore, it follows from Eqs. (1) and (5) that the *a priori* joint density $f_{\Pi^i|S^i}$ of the probability morph matrix Π^i, conditioned on the symbol string S^i, is given as

$$\begin{aligned} f_{\Pi^i|S^i}(\boldsymbol{\theta}^i|S^i) &= \prod_{m=1}^{|Q^i|} f_{\Pi^i_m|S^i}\left(\boldsymbol{\theta}^i_m|S^i\right) \\ &= \prod_{m=1}^{|Q^i|} \left(N^i_m + |\Sigma| - 1\right)! \prod_{n=1}^{|\Sigma|} \frac{(\boldsymbol{\theta}^i_m)^{N^i_{mn}}}{(N^i_{mn})!} \end{aligned} \quad (6)$$

where $\boldsymbol{\theta}^i = \begin{bmatrix} (\boldsymbol{\theta}^i_1)^T & (\boldsymbol{\theta}^i_2)^T & \dots & (\boldsymbol{\theta}^i_{|Q|})^T \end{bmatrix}^T \in [0,1]^{|Q| \times |\Sigma|}$

In the testing phase, the probability of observing a symbol string $\widetilde{S}$ belonging to a particular class of PFSA, $(Q^i, \Sigma, \delta^i, \Pi^i)$ is a product of independent multinomial distribution [18] given that the exact morph matrix Π^i is known.

$$\begin{aligned} &\Pr\left(\widetilde{S}|Q^i, \delta^i, \Pi^i\right) \\ &= \prod_{m=1}^{|Q|^i} (\widetilde{N}^i_m)! \prod_{n=1}^{|\Sigma|} \frac{\left(\Pi^i_{mn}\right)^{\widetilde{N}^i_{mn}}}{(\widetilde{N}^i_{mn})!} \quad (7) \\ &\triangleq \Pr\left(\widetilde{S}|\Pi^i\right) \quad \text{when } Q^i \text{ and } \delta^i \text{ are kept invariant} \quad (8) \end{aligned}$$

Similar to N^i_{mn} defined earlier for S^i, $\widetilde{N}^i_{mn}$ is the number of times the symbol σ_n is emanated from the state $q^i_m \in Q^i$ in the symbol string $\widetilde{S}$ in the testing phase, i.e.,

$$\widetilde{N}^i_{mn} \triangleq |\{\tilde{s}_k : \tilde{s}_k = \sigma_n, \ (\delta^i)^\star(q^i_o, \tilde{s}_1 \dots \tilde{s}_{k-1}) = q^i_m\}| \quad (9)$$

where $\tilde{s}_k$ is the k-th symbol in the observed string $\widetilde{S}$. It is noted that $\widetilde{N}^i_m \triangleq \sum_{n=1}^{|\Sigma|} \widetilde{N}^i_{mn}$.

The results, derived in the training phase and the testing phase, are now combined. Given a symbol string S^i in the training phase, the probability of observing a symbol string $\widetilde{S}$ in the testing phase is obtained as

follows.

$$\begin{aligned}\Pr(\widetilde{S}|S^i) &= \int\cdots\int \Pr\left(\widetilde{S}|\Pi^i=\boldsymbol{\theta}^i\right) f^i_{\Pi^i|S^i}(\boldsymbol{\theta}^i|S^i)d\boldsymbol{\theta}^i \\ &= \int\cdots\int \left[\prod_{m=1}^{|Q^i|}(\widetilde{N}^i_m)!\prod_{n=1}^{|\Sigma|}\frac{(\theta^i_{mn})^{\widetilde{N}^i_{mn}}}{(\widetilde{N}^i_{mn})!}\right] \\ &\times\prod_{m=1}^{|Q^i|}\left[(N^i_m+|\Sigma|-1)!\prod_{n=1}^{|\Sigma|}\frac{(\theta^i_{mn})^{N^i_{mn}}}{(N^i_{mn})!}\,d\theta^i_{mn}\right] \\ &= \prod_{m=1}^{|Q^i|}(\widetilde{N}^i_m)!\left(N^i_m+|\Sigma|-1\right)! \\ &\times\frac{\int\cdots\int\prod_{n=1}^{|\Sigma|}(\theta^i_{mn})^{\widetilde{N}^i_{mn}+N^i_{mn}}d\theta^i_{mn}}{\prod_{n=1}^{|\Sigma|}(\widetilde{N}^i_{mn})!(N^i_{mn})!}\end{aligned} \tag{10}$$

The integrand in Eq. (10) is the density function for the Dirichlet distribution up to the multiplication of a constant. Hence, it follows from Eq. (5) that

$$\begin{aligned}&\int\cdots\int\prod_{n=1}^{|\Sigma|}(\theta^i_{mn})^{\widetilde{N}^i_{mn}+N^i_{mn}}\,d(\theta^i_{mn}) \\ &= \frac{\prod_{n=1}^{|\Sigma|}(\widetilde{N}^i_{mn}+N^i_{mn})!}{\left(\widetilde{N}^i_m+N^i_m+|\Sigma|-1\right)!}\end{aligned}$$

Then, it follows from Eq. (10) that

$$\begin{aligned}\Pr(\widetilde{S}|S^i) &= \prod_{m=1}^{|Q^i|}\frac{(\widetilde{N}^i_m)!\left(N^i_m+|\Sigma|-1\right)!}{\left(\widetilde{N}^i_m+N^i_m+|\Sigma|-1\right)!} \\ &\times\prod_{n=1}^{|\Sigma|}\frac{(\widetilde{N}^i_{mn}+N^i_{mn})!}{(\widetilde{N}^i_{mn})!(N^i_{mn})!}\end{aligned} \tag{11}$$

In practice, it might be easier to compute the logarithm of $\Pr(\widetilde{S}|S^i)$ by virtue of Stirling's approximation formula $\log(n!)\approx n\log(n)-n$ [19] because, in most cases, both N^i and $\widetilde{N}$ would consist of large numbers.

The posterior probability of the observed symbol string S belonging to the class C_i is denoted as $\Pr(C_i|\widetilde{S})$ and is given as

$$\Pr(C_i|\widetilde{S}) = \frac{\Pr(\widetilde{S}|S^i)\Pr(C_i)}{\sum_{j=1}^{K}\Pr(\widetilde{S}|S^j)\Pr(C_j)},\quad i=1,2,\cdots,K \tag{12}$$

where $\Pr(C_i)$ is the known prior distribution of the class C_i. Then, the classification decision is made as follows.

$$\begin{aligned}D_{class} &= \arg\max_i \Pr(C_i|\widetilde{S}) \\ &= \arg\max_i\left(\Pr(\widetilde{S}|S^i)\Pr(C_i)\right)\end{aligned} \tag{13}$$

III. Fault Injection in C-MAPSS Test Bed

This section presents the C-MAPSS test bed along with the fault injection scheme. The C-MAPSS simulation test bed [13] was developed at NASA for a typical commercial-scale two-spool turbofan engine and its control system. Figure 1 shows the schematic diagram of a commercial aircraft gas turbine engine used in the C-MAPSS simulation test bed.

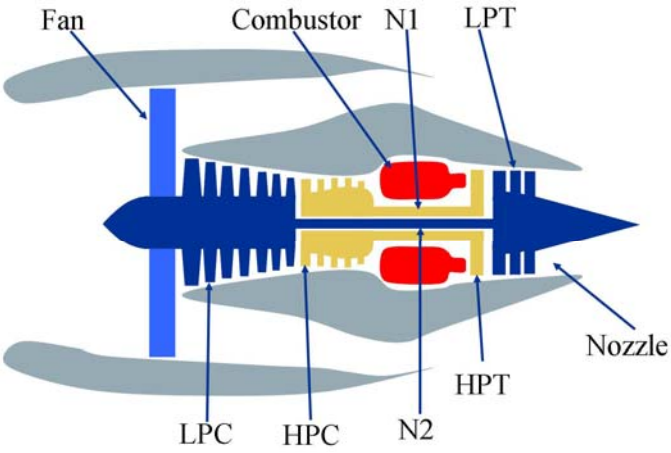

Fig. 1. Gas turbine engine schematic [13]

The engine under consideration produces a thrust of approximately 400,000 N and is designed for operation at altitude (A) from the sea level (i.e., 0 m) up to 12,200 m, Mach number (M) from 0 to 0.90, and temperatures from approximately -50°C to 50°C. The throttle resolving angle (TRA) can be set to any value in the range between 0° at the minimum power level and 100° at the maximum power level. The gas turbine engine system consists of five major rotating components, namely, fan (F), low pressure compressor (LPC), high pressure compressor (HPC), high pressure turbine (HPT), and low pressure turbine (LPT), as seen in Figure 1. Apart from the rotating components, three actuators are modeled in the simulation test bed, namely, Variable Stator Vane (VSV), Variable Bleed Valve (VBV), and Fuel Pump that controls the fuel flow rate (Wf).

Given the inputs of TRA, A and M, the interactively controlled component models in the simulation test

bed compute nonlinear dynamics of real-time turbofan engine operation. A gain-scheduled control system is incorporated in the engine system, which consists of speed controllers and limit regulators for engine components.

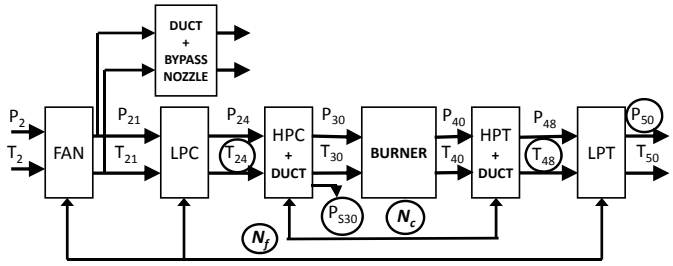

Fig. 2. Schematic diagram of the C-MAPSS engine model with Sensors

TABLE I

SENSOR SUITE FOR THE ENGINE SYSTEM

Sensors	Description
T_{24}	LPC exit/ HPC inlet temperature
Ps_{30}	HPC exit static pressure
T_{48}	HPT exit temperature
P_{50}	LPT exit pressure
N_f	Fan spool speed
N_c	Core spool speed

Out of the different types of sensors (e.g., pressure, temperature, and shaft speed) used in the C-MAPSS simulation test bed, Table I lists those sensors that are commonly adopted in the Instrumentation & Control system of commercial aircraft engines, as seen in Figure 2.

In the current configuration of the C-MAPSS simulation test bed, there are 13 component level health parameter inputs, namely, efficiency parameters (ψ), flow parameters (ζ) and pressure ratio modifiers, that simulate the effects of faults and/or degradation in the engine components. Ten, out of these 13 health parameters, are selected to modify efficiency (η) and flow (ϕ) that are defined [20] as:

- $\eta \triangleq$ Ratio of actual enthalpy and ideal enthalpy changes.
- $\phi \triangleq$ Ratio of rotor tip and axial fluid flow velocities.

For the engine's five rotating components F, LPC, HPC, LPT, and HPT, the ten respective efficiency and flow health parameters are: (ψ_F, ζ_F), (ψ_{LPC}, ζ_{LPC}), (ψ_{HPC}, ζ_{HPC}), ((ψ_{HPT}, ζ_{HPT}), and (ψ_{LPT}, ζ_{LPT}). An engine component C is considered to be in nominal condition if both ψ_C and ζ_C are equal to 1 and fault can be injected in the component C by reducing the values of ψ_C and/or ζ_C. For example, $\psi_{HPC} = 0.98$ signifies a 2% relative loss in efficiency of HPC.

A stochastic damage model has been developed and incorporated in the C-MAPSS *Transient Test-case Generator* [12] (developed by NASA), based on the experimental data for trending the natural deterioration of the engine components. Although injection of faults is described in the transient test-case generator code [12], it is explained in this paper for completeness. For all five rotating components, faults exhibit random magnitudes (F_m), and a random health parameter ratio (HPR). While F_m and HPR directly determines the change in efficiency health parameter $\psi(C)$ of a component C, a change in the flow health parameter ζ_C is determined by HPR for a given perturbation in ψ_C. Formally, the following two relations are used.

$$\delta_{\psi_C} = -\frac{F_m}{\sqrt{1+HPR^2}} \quad \text{and} \quad \delta_{\zeta_C} = \delta_{\psi_C} \cdot HPR \quad (14)$$

where δ_{ψ_C} and δ_{ζ_C} denote the changes in ψ_C and ζ_C respectively.

In the case study, fault magnitude (F_m) follows a random uniform distribution ranging from 1 to 7. Health parameter ratios (HPR) for Fan, LPC, and HPC are uniformly distributed between 1.0 and 2.0, whereas HPR s for HPT and LPT are uniformly distributed between 0.5 to 1.0. The changes in health parameters occur from certain base values of ψ_C and ζ_C.

IV. VALIDATION EXPERIMENTS AND RESULTS

This section discusses the validation experiments and results of the symbolic transient time series analysis and incipient fault detection on the C-MAPSS test bed. It is important to identify the fault quickly in time-critical operations such as takeoff, climb and landing. To find the relation between detection time and detection accuracy, numerous experiments are performed during the 'takeoff' operating condition, where Mach number varies from 0 to 0.24 in 60 seconds keeping Altitude zero and TRA at 80%. Faulty operation in three different components (Fan, LPC and HPT) along with the ideal engine condition are considered in this analysis. Hence, from the fault classification point of view, this is a four-class problem. T_{48} sensor is chosen to provide the transient response as it is able to capture the above failure signatures in the gas path model. The rationale behind choosing this sensor can be attributed

to the physical fact that it is placed between HPT and LPT, and LPT is mechanically connected to the Fan and LPC via the outer shaft. The sampling frequency of the T_{48} sensor data is 66.7 Hz. For the purpose of training, the duration for each experiment run is chosen to be 60 sec (i.e., length of the time series 4002) and transient responses from 50 runs of experiments for each fault class are concatenated.

The next step is to partition the data sets to yield respective symbol strings. The range of the time series is partitioned into 5 intervals (the size of the alphabet Σ is 5, i.e., $|\Sigma| = 5$) , each of which corresponds to a distinct symbol $\sigma_n \in \Sigma$, $n = 1, 2, \cdots, |\Sigma|$. The conversion to symbol strings is achieved by substituting each real-valued data point in the discrete time series by a symbol corresponding to the interval within which the data point lies. The training phase commences after the symbol strings are obtained for each of the four classes i.e., three component fault conditions and one nominal condition. In the D-Markov construction [6], the depth D is chosen to be 1, which implies that the probability of generation of a future symbol depends only on the last symbol, and hence the set of states is isomorphic to the symbol alphabet (i.e., $Q \equiv \Sigma$). For every class C_i, the parameters N^i_{mn} are obtained by counting the number of times the symbol σ_n is emitted from state q_m.

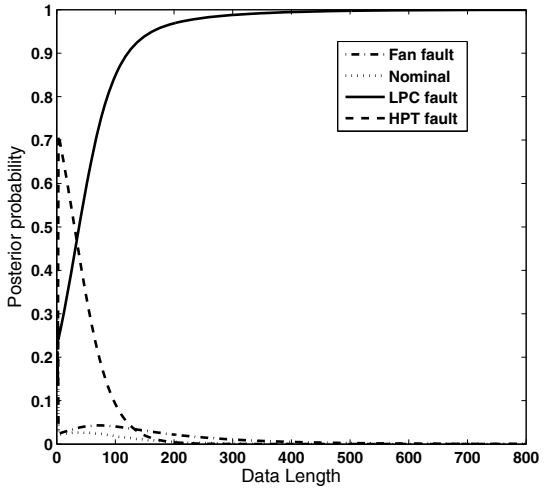

Fig. 3. Fault detection in multi-fault framework (The ground truth data corresponds to a faulty LPC)

The testing phase starts with partitioning a new time series from one of the classes and obtaining the symbol sequence by using the same alphabet and partitioning in the training phase. Following Eq. (12), the posterior probability of each class is calculated as a function of the length of the testing data set. Figure 3 shows the posterior probability of each class as a function of the length of the observed test data. It is seen that the observed sequence is correctly identified to belong to the class of LPC fault as the posterior probability of the corresponding class approaches one, while for each of the remaining classes (i.e., for other two component faults and nominal condition) it approaches zero. By repeating the same classification technique on 50 test runs of LPC fault, it is observed that 400 data length (i.e., 6 seconds) is enough to detect the fault with reasonable confidence. For other two component faults the posterior probability for the correct fault class approaches unity within data length 50. This is because the fault signatures for fan and HPT are very dominant in T_{48} response, even if they are incipient.

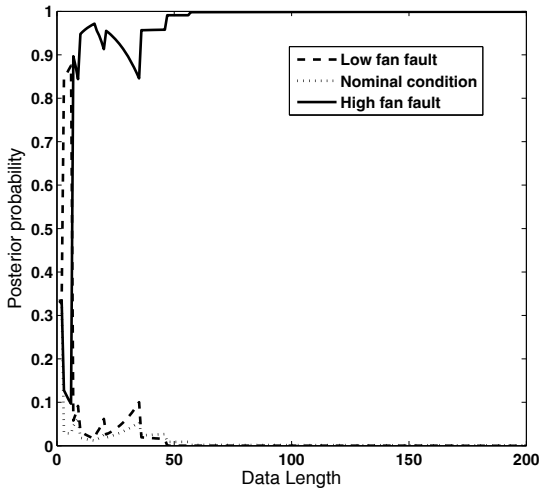

Fig. 4. High level fan fault detection (The ground truth data corresponds to high level of fault in fan)

The symbolic transient fault detection technique can also be extended to classify different levels of fault in a single component of the gas-turbine engine. To verify that, some samples of nominal data are injected with low level fan fault, where fault magnitude (F_m) follows a random uniform distribution ranging from 1 to 3. The remaining samples of nominal data are injected with high level fan fault with fault magnitude (F_m) within the range of 5 to 7. In this case, the alphabet size is 6 to obtain better class separability. Figure 4 shows that posterior probability for high fan fault reaches one quickly within 70 samples which is equivalent to one

second. But when the test case is low fan fault, the posterior probability of the high fan fault is dominant till 400 data length resulting in a false classification and the posterior probability for true class reaches 1 slowly at around 1000 data length (i.e., 15 seconds) as shown in fig. 5. This result agrees well with the intuition, that with decrease in the fault level, the detection time increases.

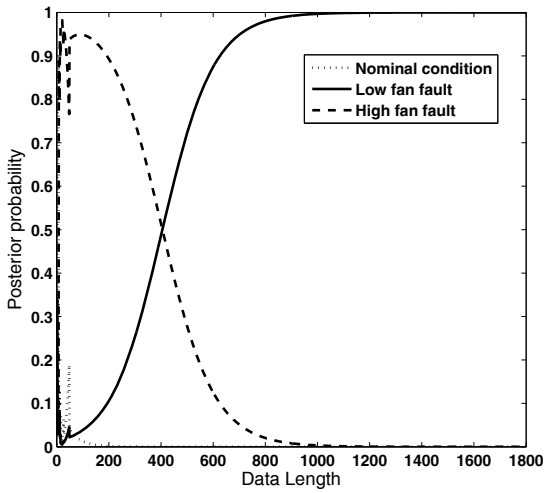

Fig. 5. Low level fan fault detection (The ground truth data corresponds to low level of fault in fan)

To examine the performance of the symbolic transient fault detection technique, a family of receiver operating characteristic (ROC) is constructed for different test data length. A binary classification scenario is considered which consists of two classes, namely, nominal engine condition belonging to the class C_1, and faulty fan condition belonging to the class C_2. The training data length is same as described in the previous experiments. The general classification rule [21] in a symbol string $\widetilde{S}$ is given by

$$\frac{\Pr(\widetilde{S}|C_1)}{\Pr(\widetilde{S}|C_2)} \underset{C_2}{\overset{C_1}{\gtrless}} \lambda \tag{15}$$

where the threshold λ is varied to generate the ROC curve. For the binary classification problem at hand, the ROC curve provides the trade-off between the probability of detection $P_D = \Pr\{\text{decide } C_2|C_2 \text{ is true}\}$ and the false alarm rate $P_F = \Pr\{\text{decide } C_2|C_1 \text{ is true}\}$. Figure 6 exhibits a family of ROC curves for the proposed fault detection technique with varying lengths of test data. It is observed that the ROC curve improves (i.e., moves toward the top left corner) considerably as

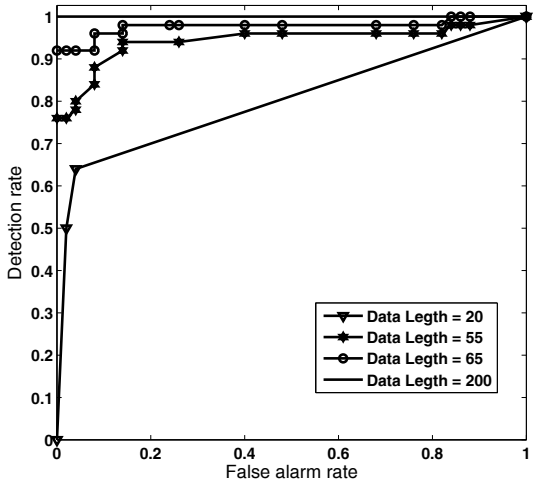

Fig. 6. ROC curves for fan fault identification with different test data lengths

the test data length is increased from $N_{test} = 20$ to $N_{test} = 200$. Based on a family of such ROC curves, it is possible to select a best combination of P_D and N_{test} for a given P_F, which would lead to a choice of the parameter λ.

V. Summary, Conclusions and Future Work

This paper addresses the analysis of transient time series data, obtained from sensors, for detection of incipient faults in aircraft gas turbine engines. A probabilistic finite state automaton is used to model the process and subsequent fault identification is performed in the context of symbolic dynamic filtering (SDF). In this method, the Dirichlet distribution and the multinomial distribution are used to model the uncertainties resulting from the finite length of symbol strings in both the training and testing phases respectively. The relation of the detection accuracy with data-length is also reported.

Although the method presented in the paper can be implemented in real time, various operating conditions need to be investigated for its on-board (in-flight) application. In addition, the following research areas are currently being pursued.

- Single sensor may not capture small faults in components or actuators. As a result, for accurate fault identification the sensor data may have to be observed for a longer period of time. Sensor fusion using cross relation among sensors in SDF framework may reduce the detection time.

- comparison with other methods, model-based and data-driven
- extension of the methodology to accommodate other types of sensor degradation, such as bias and drifting

REFERENCES

[1] Y. G. Li, "A gas turbine diagnostic approach with transient measurements," *Proceedings of IMechE, Part A: Journal of Power and Energy*, vol. 217, no. 2, pp. 169–177, 2003.

[2] G. Merrington, O. Kwon, G. Goodwin, and B. Carlsson, "Fault detection and diagnosis in gas turbines," *J. Eng. Gas Turbines Power*, vol. 113, pp. 276–282, 1991.

[3] X. Wang, N. McDowell, U. Kruger, G. McCullough, and G. W. Irwin, "Semi-physical neural network model in detecting engine transient faults using the local approach," in *Proceedings of the 17th World Congress The International Federation of Automatic Control*, July 6-11, 2008.

[4] V. P. Surender and R. Ganguli, "Adaptive myriad filter for improved gas turbine condition monitoring using transient data," *J. Eng. Gas Turbines Power*, vol. 127, pp. 329–339, 2005.

[5] S. Menon, O. Uluyol, K. Kim, and E. O. Nwadiogbu, "Incipient fault detection and diagnosis in turbine engines using hidden markov models," *ASME Conference Proceedings*, vol. 2003, no. 36843, pp. 493–500, 2003.

[6] A. Ray, "Symbolic dynamic analysis of complex systems for anomaly detection," *Signal Processing*, vol. 84, no. 7, pp. 1115–1130, 2004.

[7] C. Rao, A. Ray, S. Sarkar, and M. Yasar, "Review and comparative evaluation of symbolic dynamic filtering for detection of anomaly patterns," *Signal, Image and Video Processing*, vol. 3, no. 2, pp. 101–114, 2009.

[8] S. Gupta, A. Ray, S. Sarkar, and M. Yasar, "Fault detection and isolation in aircraft gas turbine engines: Part i - underlying concept," *Proceedings of the I Mech E Part G: Journal of Aerospace Engineering*, vol. 222, no. 3, pp. 307–318, May 2008.

[9] S. Sarkar, M. Yasar, S. Gupta, A. Ray, and K. Mukherjee, "Fault detection and isolation in aircraft gas turbine engines: Part ii - validation on a simulation test bed," *Proceedings of the I Mech E Part G: Journal of Aerospace Engineering*, vol. 222, no. 3, pp. 319–330, May 2008.

[10] S. Sarkar, C. Rao, and A. Ray, "Statistical estimation of multiple faults in aircraft gas turbine engines," *Proceedings of the I Mech E Part G: Journal of Aerospace Engineering*, vol. 223, no. 4, pp. 415–424, 2009.

[11] S. Sarkar, X. Jin, and A. Ray, "Data-driven fault detection in aircraft engines with noisy sensor measurements," *Journal of Engineering for Gas Turbines and Power-Transactions of the ASME*, vol. 133, no. 8, pp. 081602–081611, August, 2011.

[12] J. Armstrong, "User′s guide for the transient test case generator," September 2009. NASA GRC Internal Report.

[13] D. K. Frederick, J. A. DeCastro, and J. S. Litt, "User′s guide for the commercial modular aero-propulsion system simulation (C-MAPSS)," October 2007. NASA/TM2007-215026.

[14] C. R. Shalizi and K. L. Shalizi, "Blind construction of optimal nonlinear recursive predictors for discrete sequences," in *AUAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence*, (Arlington, Virginia, United States), pp. 504–511, AUAI Press, 2004.

[15] I. Chattopadhyay, Y. Wen, A. Ray, and S. Phoha, "Unsupervised inductive learning in symbolic sequences via recursive identification of self-similar semantics," in *Preprints Proceedings of American Control Conference, San Francisco, CA, USA*, June-July 2011.

[16] T. Ferguson, "Ea bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, no. 2, 1973.

[17] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, 1994.

[18] S. Wilks, *Mathematical Statistics*. John Wiley, New York, NY, USA, 1963.

[19] R. Pathria, *Statistical Mechanics*. Oxford, UK: Butterworth-Heinemann, 2nd ed., 1996.

[20] T. Kobayashi and D. L. Simon, "A hybrid neural network-genetic algorithm technique for aircraft engine performance diagnostics," in *37th Joint Propulsion Conference and Exhibit cosponsored by the AIAA, ASME, SAE, and ASEE*, (Salt Lake City, Utah), 2001.

[21] V. Poor, *An Introduction to Signal Detection and Estimation*. New York, NY, USA: Springer-Verlag, 2nd ed., 1988.