TDA Progress Report 42-112

February 15, 1993

Failure Monitoring in Dynamic Systems: Model Construction Without Fault Training Data

P. Smyth and J. Mellstrom
Communications Systems Research Section

Advances in the use of autoregressive models, pattern recognition methods, and hidden Markov models for on-line health monitoring of dynamic systems (such as DSN antennas) have recently been reported. However, the algorithms described in previous work have the significant drawback that data acquired under fault conditions are assumed to be available in order to train the model used for monitoring the system under observation. This article reports that this assumption can be relaxed and that hidden Markov monitoring models can be constructed using only data acquired under normal conditions and prior knowledge of the system characteristics being measured. The method is described and evaluated on data from the DSS 13 34-m beam waveguide antenna. The primary conclusion from the experimental results is that the method is indeed practical and holds considerable promise for application at the 70-m antenna sites where acquisition of fault data under controlled conditions is not realistic.

I. Introduction and Background

In previous articles, the problem of on-line health monitoring of a dynamic system (in particular, a DSN 34-m beam waveguide [BWG] antenna) has been investigated [1–3]. The problem can be stated in the following simple manner: let the observed data be denoted by $\underline{X}(t) = \{\underline{x}(t), \underline{x}(t-\tau), \ldots, \underline{x}(0)\}$ where each of the $\underline{x}(t)$ is a k-dimensional vector measurement of sensor data sampled at discrete time intervals τ. Given $\underline{X}(t)$, the problem is to determine the most likely current state of the system at time t, where the system is assumed to be in one of m states $\{\omega_1, \ldots, \omega_m\}$. The states are unobservable directly, but can be inferred from the observable data $\underline{X}(t)$. In probabilistic terms, the *modelling* goal is to accurately model $p[\omega_i(t)|\underline{X}(t)]$ (either from prior knowledge, training data, or a combination of both), while for *prediction*, $p[\omega_i(t)|\underline{X}(t)]$ is used to predict the current state given a specific set of data $\underline{X}(t)$ for which the system state is unknown. Typically ω_1 corresponds to the normal operating state of the system, while the other states represent various system faults that may occur. The quality of a particular model for $p[\omega_i(t)|\underline{X}(t)]$ can be obtained by measuring an empirical estimate of the *prediction accuracy*, which is simply the percentage of time that the state predicted by the model agrees with the true state—the test is performed over a period of time where the system cycles through various states (not known to the model) using data that are independent of those on which the model was trained.

In [1] and [2], an autoregressive-exogenous (ARX) time series model coupled to a pattern recognition component was used as the basis for estimating $p[\omega_i|\underline{x}(t)]$. This is a relatively simple model providing state estimates based only on instantaneous measurements $\underline{x}(t)$ but ignoring

past data. This model resulted in prediction accuracies of about 90 percent on independent test data sets obtained at DSS 13. A significant improvement on this method was reported in [3] whereby the past data were used in the state estimates by embedding the problem in a hidden Markov model (HMM) framework. The key point of the HMM method is that prior knowledge regarding the temporal behavior of the states can be used to effectively model temporal correlations in the system at the state level. On-line tests of this method at DSS 13 in November 1991 resulted in no prediction errors during a 1-hr test with state estimates being provided by the model every 6 sec [3].

It should be pointed out that the autoregressive and hidden Markov modelling methods are not the only approach for the fault detection problem. In [4] a number of statistical change detection methods were investigated. It was found that change detection methods require significant prior knowledge of the behavior of parameter characteristics when the system enters a fault state. In practice this type of detailed prior knowledge is unlikely to be available, limiting the applicability of these methods in practice.

II. Limitations of Previously Reported Methods

While the models described in [1–3] display useful capabilities in terms of on-line fault detection, they suffer from two major limitations:

(1) The models assume that the known states are exhaustive, i.e., the set of states $\{\omega_1, \ldots, \omega_m\}$ covers all possible states in which the system may be.

(2) The models also require that labelled training data are available for each state, i.e., for each state ω_i there is a set of data $\{\underline{x}(t), \underline{x}(t-\tau), \ldots, \underline{x}(0)\}$ which was measured when the system was known to be in state ω_i.

Clearly both of these requirements cannot be satisfied in most real-world fault detection applications. For fault detection, the assumption that all system states due to faults can be specified in advance is clearly inappropriate except for the simplest of systems—real-world systems (such as DSN antenna pointing systems) often contain large numbers of interacting components with feedback and nonlinearities, making prior prediction of all possible system behaviors under fault conditions unrealistic. However, it should be pointed out that it is usually possible to model system behavior under a small set of likely system faults—this point will be expanded upon later in this article.

The second requirement, that training data are available for each possible system state, is coupled to the first assumption: if all possible states cannot be described in advance, then the notion of having training data for such states is moot. However, even if the first assumption were satisfied and all fault states could be described in advance, the requirement that data can be recorded when the system is in each of these states is often unrealistic. A good example is a DSN 70-m antenna where hardware simulation of fault conditions is not a practical option due to operational considerations (as compared to the DSN 34-m antenna at DSS 13).

Hence, there is considerable practical motivation to develop methods that relax the assumptions on which the earlier-reported models are based, while still retaining the accurate prediction capabilities of these models. This article describes a relatively simple yet effective method that can detect the presence of states for which no training data sets were available, i.e., states about which the model has no knowledge. It is assumed that training data (or else a strong prior model) for at least one state is available—this is not restrictive since data under normal conditions are almost always available. The proposed method is based on the use of prior knowledge to constrain the possible distribution of system parameters, which when coupled with the model derived from the training data, allows detection of both known states and a generic, unknown state category.

This article outlines the general model, illustrates its use and effectiveness on data collected from the elevation axis of the DSS 13 BWG antenna pointing servomechanism, and describes the limitations of the current approach.

III. Notation and Assumptions

For the purposes of this article, the distinction is made between the observable data at time t which is $\underline{x}(t)$ and the estimated parameters at time t, denoted by the vector $\underline{\theta}(t)$. Typically $\underline{x}(t)$ is the original sensor data or time series (such as the motor current in an antenna pointing system), whereas the values of $\underline{\theta}(t)$ are typically statistical estimates of some characteristics of the time series such as the mean, variance, or autoregressive (AR) coefficients. In this article, attention will be limited to *block* estimation methods whereby $\underline{\theta}(t) = f[\underline{x}(t), \underline{x}(t-\tau), \ldots, \underline{x}(t-N\tau)]$, etc. Hence, each of the parameter estimates is derived from disjoint windows or blocks of the original data, where N is chosen to be large enough to enable reasonably reliable statistical estimates.

Let $\Phi_t = \{\underline{\theta}(t), \underline{\theta}(t-N\tau), \ldots, \underline{\theta}(0)\}$. In effect, Φ_t is then viewed as the observable data sequence and the prob-

lem can be treated as that of recovering the likely system states given the estimates Φ_t, i.e., find $p[\omega_i(t)|\Phi_t]$. Issues such as choosing appropriate estimators, block sizes, etc., will not be dealt with in this article. For the experimental results reported later in this article, values of $\tau = 20$ msec and $N = 200$ are used. However, for the purposes of simplification of notation, it will be assumed without loss of generality that $N\tau = 1$ during the development of the probabilistic models that follow.

It is also assumed that there are $m-1$ states for which prior information is available either in the form of: (1) specific parametric models for the dependence of the states on the observable data, or (2) training data. An additional mth state is used in the model as a single state which accounts for all other possible behaviors of the system that are qualitatively different from the known states. This state will be referred to as the unknown fault state. Hence, in the simplest case, for example, if prior information is only available for the normal state, then the model has two states: normal and the unknown fault state.

IV. The General Model

The goal of the modelling process is to provide a means of estimating the posterior state probabilities

$$p[\omega_i(t)|\Phi_t] = p[\omega_i(t)|\underline{\theta}(t), \underline{\theta}(t-1), \ldots, \underline{\theta}(0)] \qquad 1 \leq i \leq m \tag{1}$$

which are required for prediction. In the Appendix it is shown how the hidden Markov framework can be used such that the full number of conditioning terms in Eq. (1) is not necessary if the appropriate assumptions are met. The hidden Markov model leads to recursive estimates of the form

$$p[\omega_i(t)|\underline{\theta}(t), \underline{\theta}(t-1), \ldots, \underline{\theta}(0)] \approx p[\underline{\theta}(t)|\omega_i(t)] \times \sum_{j=1}^{m} f\Big(p[\omega_j(t-1)|\underline{\theta}(t-1), \ldots, \underline{\theta}(0)]\Big) \tag{2}$$

so that knowledge of the likelihood $p[\underline{\theta}(t)|\omega_i(t)]$ at each time t (in addition to the Markov transition matrix $\mathbf{A}$) is *sufficient* to calculate the posterior estimates.

Note that it will be assumed that the statistics of interest are time-invariant, hence reference to a specific time t can be dropped at this point.

In previous work, direct forward models of $p(\omega_i|\underline{\theta})$ were estimated and then $p(\underline{\theta}|\omega_i)$ was estimated by the use of Bayes' rule in Eq. (2) [3]. In this article, it is proposed to use models of the form $p(\underline{\theta}|\omega_i)$ as the direct basis for the model. The rationale behind this approach is simple: based on prior knowledge alone it is impossible except in simple cases to specify the form of $p(\omega_i|\underline{\theta})$. However, it is much more likely that one can model the dependence of the data on the state, i.e., a prior density can be assigned to the likelihood $p(\underline{\theta}|\omega_i)$ based on prior knowledge. In particular, for state ω_m, which is the state that covers all possible states not included in the set $\{\omega_1, \ldots, \omega_{m-1}\}$, one can typically specify a noninformative uniform prior density over the set of possible parameter values for $\underline{\theta}$. In addition, one must also supply models for $p(\underline{\theta}|\omega_i)$, $1 \leq i \leq m-1$, which are typically estimated from training data.

The key difference between this method and those methods proposed in previous literature is that the model works with likelihoods (the probability of the observable data given the states), rather than directly with the posterior state probabilities (the probabilities of the states given the data). This approach rules out the use of many discriminant-based methods that only provide estimates of the posterior probabilities, but do not provide estimates of the likelihoods (for example, logistic regression, feedforward neural networks, decision trees, etc.). Methods that provide the required estimates include (naturally) both parametric methods (such as maximum likelihood classifiers based on a specific parametric form for $p(\underline{\theta}|\omega_i)$ and nonparametric methods such as kernel density estimators. For a more extensive general discussion of the differences between such models, see [5–7].

The proposed method can be summarized as follows:

(1) Specify or estimate prior density models, $p(\underline{\theta}|\omega_i)$, for the *known* classes, $\omega_1, \ldots, \omega_{m-1}$. As mentioned above, this requires the use of either a parametric model (such as a multivariate Gaussian) or a nonparametric density estimation method.

(2) Specify a prior density for $p(\underline{\theta}|\omega_m)$ where ω_m is the special unknown state. This is typically done by establishing bounds or constraints on each of the parameters in $\underline{\theta}$ and then (in the absence of any other information) specifying a uniform density over the bounded parameter space.

(3) The remainder of the method is the same as before: simply estimate the hidden Markov model parameters from reliability data (as described in the Appendix) and run the model for prediction.

Note in step (2) it is important that the derived parameters can be bounded in some manner. The stronger the

constraints that can be placed on the parameters, the better will be the detection performance of the model. These constraints could be due to the basic physics of the system, such as energy limits, or a function of the particular representation being used, such as spectral or autoregressive estimates (a specific example is provided in Section V). If there are no constraints at all, then it is still possible to specify a prior model, such as a Gaussian model, although the choice of model may now be somewhat more problematic since it will inevitably reflect a prior bias which may not be appropriate. A better alternative (in the case of no constraints) would be not to use a prior model at all for ω_m and just detect data which appear to be outliers from the other $m-1$ models. However, outlier detection can be problematic—it is a central theme of this article that prior constraints can usually be placed on the parameter space of interest and that this provides the natural avenue for detecting data from ω_m. In essence, it is argued that if such prior constraints exist, this information should be used in the model, and should in principle provide better detection capabilities than any outlier detection method.

V. Applying the Likelihood Method In Practice

One significant difference between modelling the likelihoods and posterior probabilities is the issue of dimensionality, namely, that a high-dimensional parameter space will be potentially more problematic for the likelihood modelling method than for the posterior (or discriminant) modelling method. In a discriminant model (which calculates posterior probabilities), input dimensions can be ignored in the model if they are irrelevant to the state, allowing more efficient estimation at small training sample sizes. However, in the likelihood model, all input dimensions must be included in the model. If there are a significant number of irrelevant or redundant inputs, this can lead to a poor model, particularly as the ratio of sample size to input dimensions gets small. Hence, parsimony in parameter choice is recommended.

From previous work with the DSS-13 BWG-antenna pointing system, it has been found that autoregressive coefficients and standard deviation estimates are both particularly useful characteristics of the motor current for the purpose of detecting abnormal events [2,3]. In this article, three such characteristics as estimated from the motor current signal will be chosen: the two coefficients of a second-order autoregressive model [AR(2)], ϕ_1 and ϕ_2, and the standard deviation, σ. Hence, $\underline{\theta} = (\phi_1, \phi_2, \sigma)$. In [2] and [3], an eighth-order ARX model was used to model the motor current signal, using the rate command as the forcing term. However, in the interests of keeping the input dimensionality relatively low, a simpler AR(2) model was used for the purposes of this experiment. While the simpler model is not appropriate for complete system identification, it is sufficient for the purposes at hand to extract useful signal characteristics that can be used to discriminate between normal and abnormal operating conditions.

The next step is to specify a prior density over the AR parameters ϕ_1 and ϕ_2. In accordance with standard time series theory, if the estimated process (as represented by the two coefficients) is to be stationary, then the coefficients must obey the following restrictions [8]:

$$\phi_1 + \phi_2 < 1$$

$$\phi_2 - \phi_1 < 1$$

$$-1 < \phi_1 < 1$$

It will be assumed that the estimated coefficients are in fact stationary, thus providing bounds on the possible parameter values (see Fig. 1). A uniform density is specified over all such allowable values of ϕ_1 and ϕ_2. Of course, this does not allow for the fact that in practice (and in particular for fault conditions) there is no guarantee that the estimated coefficients will obey these bounds. The following approach is adopted: if the estimated coefficients lie outside the bounds of the stationary region, then the probability of the normal state $p(\omega_1|\underline{\theta})$ is set to zero.

The third parameter, the standard deviation of the voltage from the Hall effect sensor, which measures motor current, is about 20 mV under normal conditions. Based on experience from observing the motor current signal under a variety of conditions, it is estimated that under any fault condition the standard deviation should not exceed 1 V. Hence, in the absence of any other prior information, a uniform density is placed on the standard deviation over this range 0 to 1 V for σ. This density is assumed independent of the AR(2) coefficient density. This completes the specification of the likelihood model for ω_m.

For the other $m-1$ states, normal and any known fault conditions, likelihood models can be found via the use of Gaussian assumptions with maximum likelihood parameter estimation or nonparametric density estimation.

VI. Experimental Results

In [2] the acquisition of data at DSS 13 was described. Specifically, sensor data were measured under controlled conditions from the elevation axis servomechanism of the 34-m BWG antenna. Data are available for two different

days of antenna operation, referred to as day 42 and day 53. Data were recorded for about 30 min with four different fault conditions present. The faults are: tachometer noise, tachometer failure, compensation loss in the amplifier, and encoder failure. The fourth fault, encoder failure, was a real fault that was subsequently repaired. It shows up in the data as being intermittent in nature. The other three faults were purposely introduced into the hardware in a controlled manner.

The same model for the prior likelihood $p(\underline{\theta}|\omega_m)$ as described in Section V was used in all experiments. The Markov transition matrix was set to have probability of 0.99998 of remaining in the normal state, which corresponds to a mean time between failure of about 2 days. The probability of transiting to any particular fault state was set uniformly, and the probability of remaining in a fault state was set to 0.95 (corresponding to a mean failure duration of 1 min before shutdown occurred).

A model was trained on normal data and on one of the known faults (the compensation loss fault), giving a three-state model (normal, known fault, and unknown fault). The normal and known fault likelihood models were constructed using a multivariate Gaussian density where the mean and covariance parameters were estimated from the data using maximum likelihood estimators.

Two models were constructed in this manner (one on each of the day 42 and day 53 data sets) and then tested on the independent data from the other day. The goal of the experiment was to see if the model could correctly identify data as being either normal, a known fault, or an unknown fault. The ability to classify data into the third unknown category was of particular relevance, since, as described earlier in this article, previously developed models did not have this capability, i.e., all data were classified into one of the known states.

The state sequence in each test data set was as follows: normal, unknown fault (tachometer failure); known fault (compensation loss); normal conditions, unknown fault (tachometer noise); and finally an unknown intermittent fault (encoder failure). Each state lasted roughly 5 min in duration. Figure 2 shows how one of the AR(2) coefficients changed as a function of the underlying state (for day 53). Note how noisy the estimates are, due in part to the fact that an AR(2) model is too simple to capture the full dynamics of the data.

Figures 3(a) and 3(b) show the state estimation results in terms of estimated state probabilities [as in Eq. (2)] for each of the three states in the model. The results clearly indicate that the likelihood model has the ability to infer the correct state of the system from the observable data. As in [3], the Markov model adds stability to the estimates, reducing false alarms while still allowing a rapid transition when the underlying state changes.

The important aspect of this new model is its ability to identify data as being of the unknown category, namely, between minutes 5 and 10, minutes 20 and 25, and the intermittent fault that occurred between minutes 25 and 30. The response of the model is not entirely perfect. For example, during the test of day 42 data [Fig. 3(a)], in the first 5 min of normal operations, there appear to be at least two short false alarms, i.e., where the probability of normal conditions drops significantly below 1 even though the system is supposed to be in the normal state. This can be attributed to one of two possible causes: either the model is not quite accurate, or, more interestingly, although the system is assumed to be normal it is in fact in some other transient state. Closer examination of the original sensor data revealed that the second explanation was more likely to be true: the model detected the possibility of an unknown transient state that had not been noticed when these data were originally recorded. While this is a relatively simple example, it nonetheless demonstrates the basic concept of a model which can detect subtle changes and abnormalities in the behavior of a dynamic system—changes that are not noticeable to the human observer.

It is also worth noting that the present model assigns a relatively low probability to short a priori states. An obvious extension to the model proposed here would be to further refine the unknown state into substates based on their temporal characteristics, i.e., intermittent or transient, or permanent.

VII. Discussion

This article has described the basic principles behind the construction of dynamic system monitoring models that can classify system states into an "unknown" category. Although the basic idea is quite simple, it has some very useful properties. In addition, it is worth noting that all previous fault monitoring methods described in the literature (of which the authors of this article are aware) implicitly assume that all system states of interest are known in advance. For large-scale complex systems, this is clearly an undesirable and unrealistic assumption.

A possible criticism of the proposed method is the possibly arbitrary nature by which the prior density for the unknown state is assigned. Certainly it must be admitted that this can never be a purely objective choice and requires the careful judgment of the modeller. However, any

model is by nature the result of various implicit biases and subjective judgments, and, hence the standard argument of the Bayesian school of statistical modelling can be appealed to: if any reasonable prior information exists, then it is judicious to include it in the model. The astute reader will have noted that by simply changing the boundaries or constraints on the parameters of interest, the modeller can in effect control the detection to false alarm trade-off characteristic of the model [also known as the receiver operating characteristic (ROC) in signal detection theory]. The use of decision theoretic methods to minimize the relevant loss function (in the context of choosing the prior density) would seem the appropriate avenue by which to control this aspect of the model.

The ability to detect new system states does not come without a cost. As alluded to earlier, the mapping describing how the observed data depend on the system states (the likelihood) is generally more difficult to estimate than the mapping describing how the states depend on the observed data. Hence, for example, in the case where one has three known faults, a model of the type which is proposed here may not be as accurate in terms of discriminating among these faults as the types of discrimination models that focus exclusively on these faults but which ignore the possibility of an unknown fault. One way to avoid this problem is to improve the quality of the likelihood modelling process. For example, a Gaussian assumption is often not appropriate: nonparametric density estimation methods, if used correctly, may provide more accurate models for the known states.

Another possibility would be to use both likelihood models and discriminative models as part of one overall model. Letting $p(\omega_m)$ be the posterior probability that the data are from an unknown state (as calculated by a likelihood model of the type described in this article), and letting the symbol $\omega_{\{1,\ldots,m-1\}}$ denote the event that the true system state is one of the known states, one can estimate the true posterior probability of individual known states as

$$p_d(\omega_i|\Phi,\omega_{\{1,\ldots,m-1\}}) \times [1 - p(\omega_m)], \quad 1 \leq i \leq m-1$$

where $p_d(\omega_i|\Phi,\omega_{\{1,\ldots,m-1\}})$ is the posterior probability estimate of the known states as provided by a discriminative model such as described in [2] and [3]. Note that this method does not in any way help to improve the ability of the overall model to detect unknown states since that estimate remains unchanged; however, in principle, it should improve the ability of the model to distinguish between specific known states. The possibility of improving the model described in this article by using this particular technique has not been tested in an experimental manner at this point.

A final comment is that the ability of the likelihood model to detect unknown states is necessarily limited by the information in the observable data. For example, although the simple AR models reported here have given very useful information in terms of discriminating between normal and various fault states, it is quite possible that a fault state may not be well modelled by a simple linear AR model, i.e., that the AR coefficients will not yield any useful information. Hence, in general, the use of more robust signal characteristics should improve the model performance.

VIII. Summary

A new method was proposed that allows the construction of HMM monitoring algorithms without the requirement that training data for each of a prescribed set of faults be made available. Naturally, if such data (or equivalent prior knowledge) are available, then these data can also be incorporated into the new model. The proposed method was validated on data from the DSS-13 BWG-antenna pointing system. In particular, the model was able to detect system states that could not have been detected using previously reported methods. While there is still room for improvement in terms of the performance of this class of models, the results are nonetheless quite accurate and of significant practical importance in the context of monitoring 70-m antenna data where fault training data are unlikely to be available.

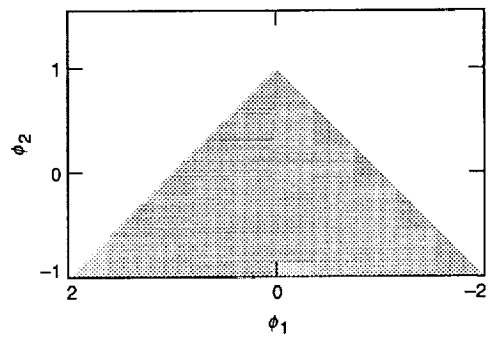

Fig. 1. Admissible region for AR(2) parameters.

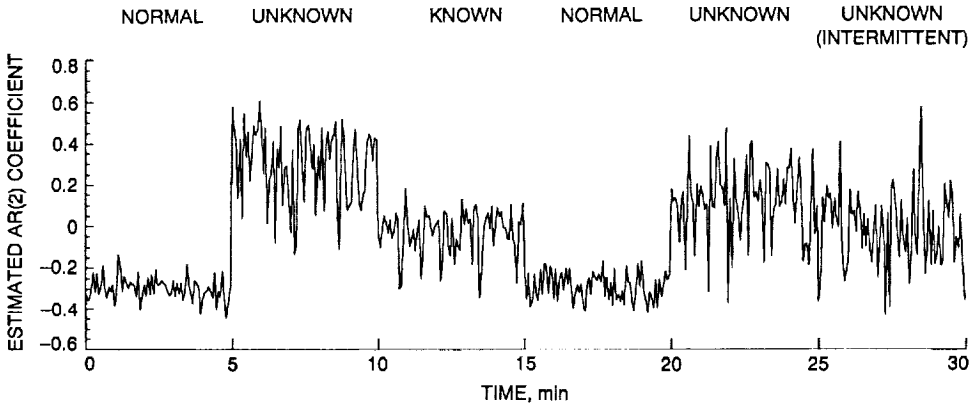

Fig. 2. Estimates of coefficient ϕ_2 as a function of the system state (day 53 data).

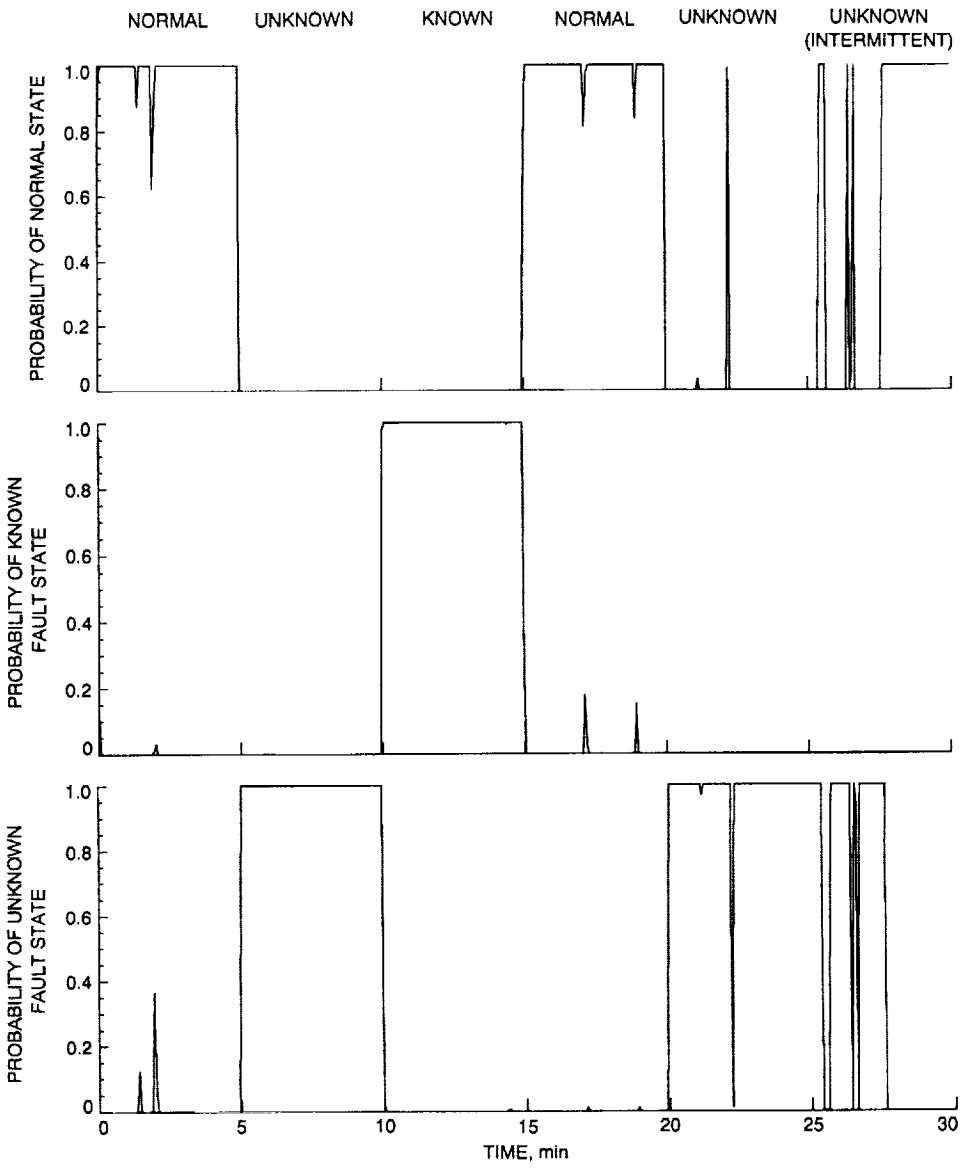

Fig. 3. Estimates of posterior state probabilities as provided by the likelihood and hidden Markov models: (a) training on day 53 and testing on day 42 data and (b) vice versa.

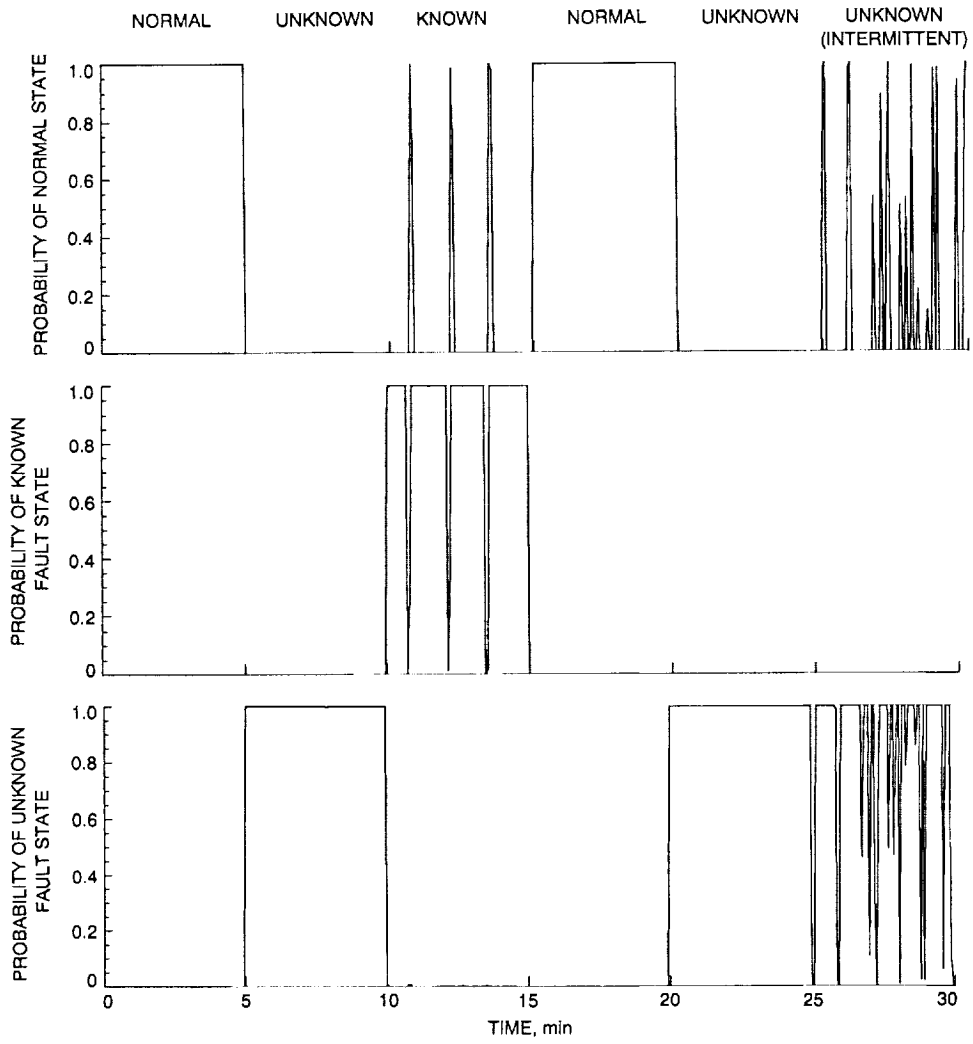
NORMAL
UNKNOWN
KNOWN
NORMAL
UNKNOWN
UNKNOWN
(INTERMITTENT)
PROBABILITY OF NORMAL STATE
PROBABILITY OF KNOWN FAULT STATE
PROBABILITY OF UNKNOWN FAULT STATE
1.0
0.8
0.6
0.4
0.2
0
0
5
10
15
20
25
30
TIME, min

Fig. 3. (contd)

Appendix

Hidden Markov Model Description

Let Ω denote the discrete-valued state variable taking values in the set $\{\omega_1, \ldots, \omega_m\}$. A first-order discrete-time Markov model is characterized by the assumption that

$$p[\Omega(t)|\Omega(t-1), \ldots, \Omega(0)] = p[\Omega(t)|\Omega(t-1)] \qquad \text{(A-1)}$$

That is, that the conditional probability of any current state given knowledge of all previous states is the same as the conditional probability of the current state given knowledge of the system state at time $t-1$. This is equivalent to the well-known assumption of "memorylessness" in that the evolution of the system depends only on the present state and not on the past state. A direct consequence is the fact that the number of consecutive time steps that the system spends in any given state will be a discrete random variable with a geometric distribution.

In a standard, *nonhidden* Markov model, to calculate the probability that the system is in a given state at time t, one needs only to know the initial state probabilities $\pi = p[\omega_1(0), \ldots, \omega_m(0)]$ and the values $a_{ij} = p[\omega_i(t)|\omega_j(t-1)]$, $1 \leq i, j \leq m$. The $m \times m$ matrix $\mathbf{A}$ is known as the transition matrix and characterizes the Markov model. The first-order Markov assumption governing state evolution in time may appear restrictive at first glance, but has been found in practice to be an extremely robust model for many real-world applications. In principle, the theory for higher-order Markov models can be developed, but at the cost of increased complexity in terms of specifying the model and of increased computational complexity in terms of on-line calculation of the posterior probabilities.

Under the first-order Markov assumption, it can easily be shown that the probability in which the system remains in the normal state from one instant to the next can be expressed as

$$a_{11} = 1 - \frac{\tau}{\text{MTBF}} \qquad \text{(A-2)}$$

where the MTBF is the mean time between failures of the system and τ is the time between states (both expressed in the same units). Similarly, the other elements in the transition matrix $\mathbf{A}$ can be estimated from information concerning the general nature of system faults, which may be available from an existing database or can be estimated based on known physical properties of the system. Augmented models may have a wide variety of additional states. For example, it may be useful to include a state to account for the transient behavior of the system. Similarly, states which account for known operational modes of the system, such as powered off and brakes on, may be necessary in practice. The specification of the Markov transition matrix corresponds to the explicit modelling of high-level prior knowledge concerning system behavior at the *state* level. In particular, it does not involve the specification of prior models for *observable data* over time since typically this is much more difficult to model. This is precisely the advantage of the HMM decomposition: the temporal behavior of the system needs only to be specified at a relatively high level.

Denote the observed data up to time t to be $\Phi_t = \{\underline{\theta}(t), \ldots, \underline{\theta}(0)\}$. The hidden aspect of the Markov model is derived from the fact that the observed data Φ_t is a stochastic function of the underlying Markov states. These states are hidden in the sense that they cannot be measured directly. It is the state identities which one wishes to estimate, hence, the purpose of the modelling is to represent the relationship between the states and the observable data such that the most likely state sequence can be inferred. Figure A-1 shows an illustration of the concept for a three-state HMM. For on-line monitoring of a dynamic system, the observed data simply consist of observed sensor data (or derived parameters) while the states reflect the underlying system states, in particular normal and fault operational states.

An estimate of the instantaneous likelihood, the probability of the observed data at time t conditioned on the state variable, $p[\underline{\theta}(t)|\Omega(t)]$, is assumed to be known. The goal is to take advantage of *all* the symptom information and to estimate $p[\Omega(t)|\Phi_t]$. It is convenient to work in terms of an intermediate variable α, where

$$\alpha_i(t) = p[\omega_i(t), \Phi_t] \qquad \text{(A-3)}$$

To find the posterior probabilities of interest, it is sufficient to be able to calculate the α's at any time t, since by Bayes' rule

$$p(\omega_i(t)|\Phi_t) = \frac{1}{p(\Phi_t)}\alpha_i(t) = \frac{\alpha_i(t)}{\sum_{j=1}^{m}\alpha_j(t)} \tag{A-4}$$

The derivation of a recursive estimate follows:

$$\alpha_i(t) = \sum_{j=1}^{m} p\Big(\omega_i(t), \Phi_t, \omega_j(t-1)\Big) = \sum_{j=1}^{m} p\Big(\omega_i(t), \underline{\theta}(t), \Phi_{t-1}, \omega_j(t-1)\Big)$$

$$= \sum_{j=1}^{m} p\Big(\omega_i(t), \underline{\theta}(t)|\Phi_{t-1}, \omega_j(t-1)\Big)\alpha_j(t-1)$$

by the definition of α_j

$$= \sum_{j=1}^{m} p\Big(\underline{\theta}(t)|\omega_i(t), \Phi_{t-1}, \omega_j(t-1)\Big)p\Big(\omega_i(t)|\Phi_{t-1}, \omega_j(t-1)\Big)\alpha_j(t-1)$$

$$= \sum_{j=1}^{m} p\Big(\underline{\theta}(t)|\omega_i(t)\Big)p\Big(\omega_i(t)|\Phi_{t-1}, \omega_j(t-1)\Big)\alpha_j(t-1)$$

assuming that $\underline{\theta}(t)$ is independent of past observations and past states, *given* the present state

$$= \sum_{j=1}^{m} p\Big(\underline{\theta}(t)|\omega_i(t)\Big)p\Big(\omega_i(t)|\omega_j(t-1)\Big)\alpha_j(t-1)$$

assuming that $\omega_i(t)$ is independent of past observations *given* the past states

$$= p\Big(\underline{\theta}(t)|\omega_i(t)\Big)\sum_{j=1}^{m} a_{ij}\alpha_j(t-1) \tag{A-5}$$

The first term is the likelihood (assumed to be known). The terms in the sum are just a linear combination of the α's from the previous time step. Hence, Eq. (A-5) provides the basic recursive relationship for estimating state probabilities at any time t.

The additional assumptions made in the derivation of Eq. (A-5) (besides the first-order Markov assumption on state dependence) require some comment. The first assumption is that $\underline{\theta}(t)$ is independent of both the most recent state and the observed past data, *given* that the present state is known. This implies that the observed symptoms are assumed to be statistically independent from one time window to the next, given the state information. This will generally be true when the values of $\underline{\theta}(t)$ consist of derived parameters and τ is much greater than any significant time constants of the dynamic system. Even if it is known that the $\underline{\theta}$'s exhibit temporal correlations, this can also in principle be modelled directly in Eq. (A-5), although the model will now be much more complex. The second assumption, that the present state only depends on the previous state but not the past observations, simply reflects the causal relationship between symptoms and states.

Note that the method described above only calculates the state probabilities based on *past* information. Alternative estimation strategies are possible. For example, using the well-known forward-backward recurrence relations [9], one can update the state probability estimates using symptom information which occurred *later* in time.

References

[1] P. Smyth and J. Mellstrom, "Initial Results on Fault Diagnosis of DSN Antenna Control Assemblies Using Pattern Recognition Techniques," *The Telecommunications and Data Acquisition Progress Report 42-101*, vol. January–March 1990, Jet Propulsion Laboratory, Pasadena, California, pp. 136–151, May 15, 1990.

[2] J. Mellstrom and P. Smyth,"Pattern Recognition Techniques Applied to Performance Monitoring of the DSS 13 34-m Antenna Control Assembly," *The Telecommunications and Data Acquisition Progress Report 42-106*, vol. April–June 1991, Jet Propulsion Laboratory, Pasadena, California, pp. 30–51, August 15, 1991.

[3] J. Mellstrom, C. Pierson, and P. Smyth, "Real-time Antenna Fault Diagnosis Experiments at DSS 13," *The Telecommunications and Data Acquisition Progress Report 42-108*, vol. October–December 1991, Jet Propulsion Laboratory, Pasadena, California, pp. 96–108, February 15, 1992.

[4] P. Scholtz and P. Smyth, "Fault Detection Using a Two-Model Test for Changes in the Parameters in a Time Series," *The Telecommunications and Data Acquisition Progress Report 42-110*, Jet Propulsion Laboratory, Pasadena, California, August 15, 1992.

[5] P. Smyth and J. Mellstrom, "Fault Diagnosis of Antenna Pointing Systems Using Hybrid Neural Networks and Signal Processing Techniques," *Advances in Neural Information Processing Systems 4*, J. E. Moody, S. J. Hanson, and R. P. Lippmann (eds.), Los Altos, California: Morgan Kaufmann Publishers, pp. 667–674, 1992.

[6] P. Smyth and J. Mellstrom, "Detecting Novel Classes With Applications to Fault Diagnosis," *Proceedings of the Ninth International Conference on Machine Learning*, Los Altos, California: Morgan Kaufmann Publishers, pp. 416–425, 1992.

[7] P. Smyth, "Probability Density Estimation and Local Basis Function Neural Networks," *Computational Learning Theory and Natural Learning Systems 2*, T. Petcshe, M. Kearns, S. Hanson, and R. Rivest (eds.), Cambridge, Massachusetts: MIT Press, in press, 1993.

[8] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, San Francisco, California: Holden-Day, 1970.

[9] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

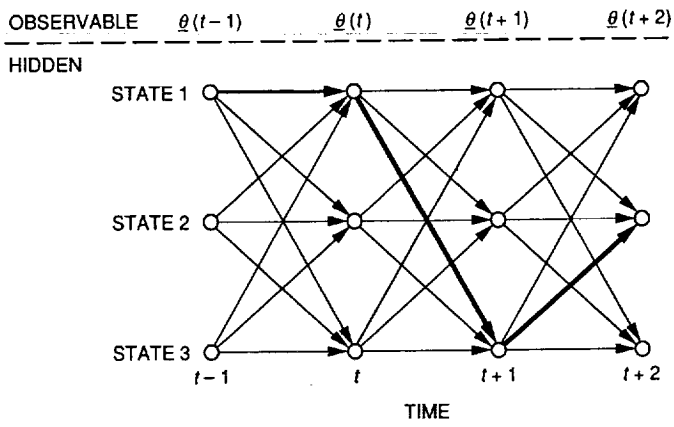

Fig. A-1. An illustrative example of a three-state hidden Markov model.