The Gray-Box Approach to Sensor Data Analysis

M. Zak[1] and H. Park[1]

Model-based fault diagnosis has become an important approach for diagnosis of dynamical systems. By comparing the observed sensor values with those of the values predicted by the model, e.g., the residual, the health of the system can be assessed. However, because of modeling errors, sensor noise, disturbances, etc., direct comparison of observed and predicted values can be difficult.

In an effort to address this problem, we present a new method called the gray-box method. It is called a "gray box" because a deterministic model of the system, i.e., a "white box," is used to filter the data and generate a residual, while a stochastic model, i.e., a "black-box," is used to describe the residual. Instead of setting a threshold on the residual, the residual is modeled by a three-tier stochastic model. The linear and nonlinear components of the residual are described by an autoregressive process and a time-delay feed-forward neural network, respectively. The last component, the noise, is characterized by its moments.

The stochastic model provides a complete description of the residual, and the faults can be detected by monitoring the parameters of the autoregressive model, the weights of the neural network, and the moments of noise. The method is robust to system modeling errors and is applicable to both linear and nonlinear systems.

I. Introduction

Fault diagnosis is an important element in realizing truly autonomous vehicles. Reliable information about the operational health of the vehicle is crucial for proper mission planning and on-board intelligent decision making. To fully assess the vehicle health, the diagnostic system must have comprehensive ability to sense impending failures, rather than failures, and operational difficulties. While fixed thresholds, i.e., traditional redlines, may be sufficient for simple steady-state systems, more sophisticated diagnostic techniques are needed for unsteady operations and detection of incipient faults.

The natural starting point for a more sophisticated diagnosis is the model of the system. Fortunately, many systems, such as aircraft, spacecraft, gas turbine engines, hydraulic systems, etc., usually have well-developed dynamic models. The most straightforward application of the model for diagnosis is to compare the observed sensor readings with those predicted by a model. If the difference between the observed and the predicted values, i.e., the residual, is greater than some set threshold value, an anomaly has occurred.

---

[1] Explorations Systems Autonomy Section.

The research described in this publication was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

In practice, however, it is not straightforward to compare the observed and predicted values because the quality of the model may vary and noise may be present. If the model is inaccurate or has insufficient detail, the predicted values may differ greatly from those of the actual system. Some deviations are unavoidable since there is no theoretical description for the phenomenon. For example, secondary effects such as friction, thermal effects, sensor noise, etc., may not have simple model descriptions. In other cases, the model can be purposely coarse, i.e., contain insufficient detail, to facilitate real-time computations.

In an effort to mitigate the problem of comparing observed and predicted values, many different approaches have been developed to generate robust residuals and/or thresholds for anomalies. A comprehensive overview of model-based fault diagnosis is found in Chen and Patton [1]. These methods include adaptive threshold methods, observer-based approaches, parity relation methods, parameter estimation methods, and statistical testing methods.

In adaptive threshold methods, the threshold on the difference between the observed and predicted values is varied continuously as a function of time [2]. The method is passive in the sense that no effort is made to design a robust residual [1]. The unknown input observer (UIO) and parity relation methods are active since the residual is made to be robust to known disturbances and modeling errors. The residual is sensitive to only unknown disturbances that are likely to be anomalies or faults in the system. The drawback of these methods is that the structure of the input disturbances and modeling error must be known. In addition, the methods are applicable to mostly linear systems. The parameter estimation methods use system identification techniques to identify changes in the model parameters of the dynamical system. The advantage of this method is that the implementation is straightforward, and it can deal with nonlinear systems. The disadvantage is that a large amount of computational power may be required to estimate all of the parameters in the model. Finally, statistical testing methods use statistical techniques such as the weighted sum-squared residual (WSSR), x^2 testing, sequential probability ratio testing (SPRT), the generalized likelihood ratio (GLR), etc., to differentiate between normal noise and anomalous sensor values. The disadvantage of this method is that the residual is assumed to be a zero-mean white-noise process with known covariance matrix. The residual in many cases may not be describable in this manner.

II. Gray-Box Method

In an effort to improve model-based fault diagnosis, we propose a new approach called the gray-box method. It is called a "gray box" because it incorporates both a "black box," i.e., a stochastic model, and a "white box," i.e., a deterministic model. It is a hybrid model incorporating elements from residual-based methods and parametric-estimation methods. It is similar to adaptive-threshold methods in that a residual is generated without any regard for robust residual generation. However, instead of examining the amplitude of the residual as in the case of the adaptive threshold methods, the structure, i.e., the model parameters, of the residual is examined. The residual generation is our white box. The residual is modeled using techniques similar to the parametric estimation methods. The method is distinct from the standard parametric estimation method in that the system identification is carried out on the residual, not on the system variables directly. The residual is parameterized, not the full system. In our terminology, the parameter estimation method is a black box.

A block diagram of the gray-box method is shown in Fig. 1. After filtering the deterministic components using the model of the system, the residual is separated into its linear, nonlinear, and noise components and is fitted to stochastic models. The parameters to the linear, nonlinear, and noise models completely describe the residual. The gray box has several advantages. First, the full model is employed rather than only the model structure, as in the case of standard parametric estimation methods. Thus the gray box takes full advantage of the information about the system. Second, the gray-box method can be made robust to modeling errors that can be taken care of during residual modeling. The model of the residual also can describe many unmodeled phenomena in the original model. Finally, the method is applicable to both linear and nonlinear systems.

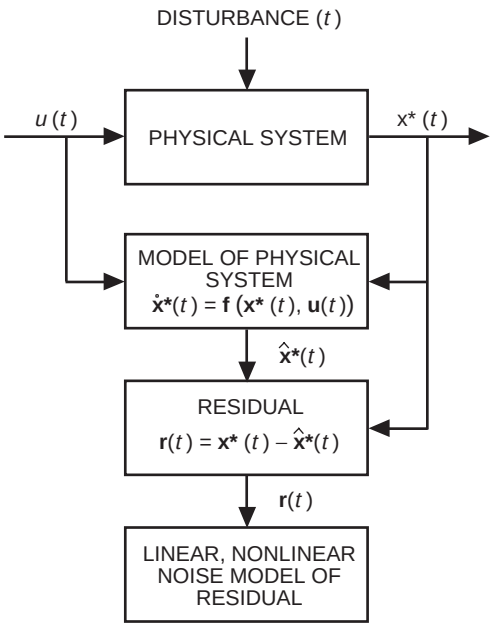

Fig. 1. The gray-box approach to modeling residual.

III. Residual Generation

Any theoretical dynamical model includes two types of components: those that describe the phenomena directly associated with the primary function of the system (such as the effect on rotor speed of torque exerted on the turbine shaft) and those that represent secondary effects (such as frictional losses, heat losses, etc.). The first type of component usually is well understood and possesses a deterministic analytical structure, and, therefore, its behavior is fully predictable. On the other hand, the second type may be understood only on a much more complex level of description (including molecular level) and cannot be simply incorporated into a theoretical model. In fact, some components may be poorly understood and lack any analytical description, e.g., viscosity of water in micro-gravity. Therefore, the first step in the gray-box approach is to filter out the contributions that are modeled, i.e., the components of the first type, and to focus on the components of the second type whose theoretical prediction is inadequate.

The residual generation is as follows. Let us assume that the theoretical model is represented by a system of differential equations:

$$\dot{\mathbf{x}}(t) = \mathbf{f}\big(\mathbf{x}(t), \mathbf{u}(t)\big) + \mathbf{y}(t) \tag{1}$$

where $\mathbf{x}(t)$ is the state variable vector, $\mathbf{u}(t)$ is the known input, and $\mathbf{f}$ is the known theoretical relationship following from conservation laws of mechanics, thermodynamics, etc. The last term, $\mathbf{y}(t)$, represents components that lack theoretical descriptions, are too complex, or are the result of modeling errors. These can include sensor noise, unknown input to the system, friction in bearings, material viscosity, and other secondary effects such as torsional oscillations of the shaft, flutter in the turbine and compressor blades, incomplete fuel combustion, etc.

The estimate of the system is accomplished by substituting the observed sensor data for the evolution of the state variables, $\mathbf{x}^*(t)$, and input, $\mathbf{u}(t)$. Hence,

$$\dot{\mathbf{x}}^*(t) = \mathbf{f}\big(\mathbf{x}^*(t), \mathbf{u}(t)\big) \tag{2}$$

The residual,

$$\mathbf{r}(t) = \mathbf{x}^*(t) - \hat{\mathbf{x}}^*(t) \tag{3}$$

is generated by subtracting the solution of Eq. (2), $\hat{\mathbf{x}}^*(t)$, which is generated by using the observed state variables, $\mathbf{x}^*(t)$, from the solution of Eq. (1). Hence, the original theoretical model is the filter.

In general, the residual can be treated as another realization of some stochastic process. If the theoretical model, Eq. (1), is accurate and accounts for most physical effects, and if the observed state variables are accurate, then the residual, $|\mathbf{r}(t)|$, will be very small, i.e.,

$$|\mathbf{r}(t)| \ll |\mathbf{x}^*(t)| \tag{4}$$

and either a fixed or an adaptive threshold can be assigned as a criterion for anomalous behavior. If the system is linear and the structure of $\mathbf{y}(t)$ is known, a more sophisticated unknown input observer (UIO) filter [1] can be constructed to make the residual more robust to modeling errors and disturbances. However, in our gray-box approach, the simple form of Eq. (3) is preferred over the more robust residuals since the residual is to be modeled. If the residual is too robust, the characteristic structure of $\mathbf{y}(t)$ will become hidden.

As an example, consider the simplest gas turbine plant consisting of a turbine, 1; a compressor, 2; and a combustion chamber, 3 (Fig. 2). Ignoring the thermal inertia of the combustion chamber, one can write the following dynamic equation for the angular velocity, ω, of the shaft:

$$J\frac{d\omega}{dt} = M_1(\omega, \mu) - M_2(\omega) - M_r(t) \tag{5}$$

where J is the moment of inertia of the turbo-compressor $(1-2)$ in the axis of rotation; M_1 is the turning moment generated by the turbine; M_2 is the resistive moment applied by the compressor, bearings, etc., on the shaft; μ is the rate of fuel burned inside the combustion chamber; and $M_r(t)$ is the random moment applied by effects such as torsional vibration of the shaft, blade flutter in the compressor and turbine, propagation of pressure pulses along the pipelines, heat loss, seal leaks, etc.

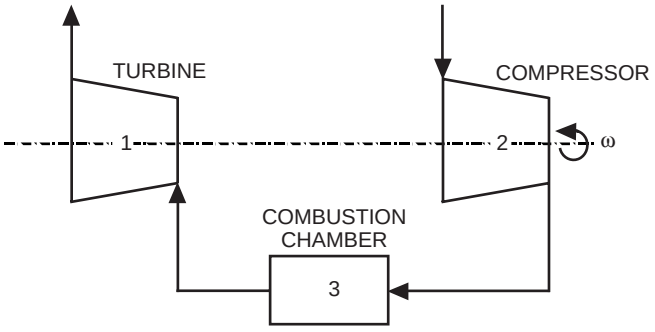

Fig. 2. Schematic of a gas turbine plant.

The conditions for stability of the solutions of Eq. (5) are

$$\frac{\partial M_1}{\partial \omega} < 0, \quad \frac{\partial M_2}{\partial \omega} > 0 \tag{6a}$$

or

$$\frac{\partial}{\partial \omega}\left(M_1 - M_2\right) < 0 \tag{6b}$$

Consequently, if one linearizes Eq. (5) with respect to a steady-state regime where the rate of fuel burn is constant, i.e,

$$\mu = \mu_o = \text{constant} \tag{7}$$

Eq. (5) can be reduced to the form

$$\dot{\omega} = -\gamma\omega + \Gamma(t) \tag{8}$$

where,

$$\gamma = \frac{1}{J}\left[\frac{\partial M_1(\omega,\mu)}{\partial \omega} - \frac{\partial M_2(\omega)}{\partial \omega}\right]_{\mu=\mu_o} > 0 \tag{9}$$

and

$$\Gamma(t) = \frac{M_r(t)}{J} \tag{10}$$

The $\Gamma(t)$ represents a stochastic force, and Eq. (8) is a Langevin equation whose formal solution is

$$\omega(t) = \omega_o e^{-\gamma t} + \int_0^t e^{-\gamma(t-t')}\Gamma(t')dt' \tag{11}$$

subject to the initial condition

$$\omega = \omega_o \text{ at } t = 0 \tag{12}$$

This solution is the only information that can be obtained from the sensor data. The first term in Eq. (11) is fully deterministic and represents all of the theoretical knowledge about the plant. The second term includes the stochastic force, Eq. (10), and is stochastic. Hence, the stochastic process described by Eq. (11) represents only a part of the sensor data. Substituting the measured sensor data, ω^*, into the theoretical model, Eq. (8), the original stochastic force is immediately exposed as the inverse solution:

$$\Gamma(t) = \dot{\omega}^* + \gamma\omega^* \tag{13}$$

Equation (11) shows that the more stable the model, i.e., the larger the value of γ, the less the stochastic force, $\Gamma(t)$, contributes to the sensor data since

$$0 < e^{-\gamma(t-t')} < 1 \text{ at } t > t' \tag{14}$$

In other words, for highly stable dynamical models, the stochastic forces become deeply hidden in the sensor data. However, using the theoretical model as a filter damps the deterministic components and amplifies the stochastic components. This effect of damping deterministic and amplifying unknown components, i.e, sensor noise, modeling errors, etc., is important if the residual is to be modeled properly.

IV. Residual Modeling

For the model of the residual, we start with a traditional description of sensor data given in the form of a time series that describes the evolution of an underlying dynamical system. It will be assumed that this time series cannot be approximated by a simple analytical expression and is not periodic. In other words, for an observer, the future values of the time series are not fully correlated with the past ones, and, therefore, they are apprehended as random. Such time series can be considered as a realization of an underlying stochastic process that can be described only in terms of probability distributions. However, any information about this distribution cannot be obtained from a simple realization of a stochastic process unless this process is stationary. Then the ensemble average can be replaced by the time average. An assumption about the "stationarity" of the underlying stochastic process would exclude from consideration such important components of the dynamical process as linear and polynomial trends, or harmonic oscillations. Thus, a method is needed to deal with nonstationary processes.

Our approach to building a dynamical model of the residual is based upon progress in three independent fields: nonlinear dynamics, theory of stochastic processes, and artificial neural networks. From the field of nonlinear dynamics, based upon the Takens theorem [3], any dynamical system that converges to an attractor of a lower (than original) dimensionality can be simulated with a prescribed accuracy by the time-delay equation:

$$x(t) = F\big(x(t-\tau), x(t-2\tau), \cdots, x(t-m\tau)\big) \tag{15}$$

where $x(t)$ is a given time series, such as a variable in the residual vector, $\mathbf{r}(t)$, and τ, a constant, is the time delay.

It was proven that the solution to Eq. (15) subject to appropriate initial conditions converges to the original time series:

$$x(t) = x(t_1), x(t_2), \cdots \tag{16}$$

if m in Eq. (15) is sufficiently large.

However, the function F, as well as the constant τ and m, are not specified by this theorem, and the most "damaging" limitation of the model, Eq. (15), is that the original time series must be stationary since it represents an attractor. This means that for nonstationary time series the solution to Eq. (15) may not converge to Eq. (16) at all. Actually, this limitation has deeper roots and is linked to the problem of stability of the model, Eq. (15).

Before [3], a different approach [4] to the same problem was developed in the statistic community. A discrete-time stochastic process can be approximated by a linear autoregressive model:

$$x(t) = a_1 x(t-1) + a_2 x(t-2) + \cdots + a_n (t-n) + z(t) \text{ as } n \to \infty \tag{17}$$

where a_i are constants and $z(t)$ represents the contribution from white noise.

As shown by [7], any zero-mean purely nondeterministic stationary process $x(t)$ possesses a linear representation as in Eq. (17) with $\sum_{j=1}^{\infty} a_j^2 < \infty$, i.e., the condition of the stationarity.

In order to apply Eq. (17), the time series Eq. (16) must be decomposed into its stationary and nonstationary components. To "stationarize" the original time series, certain transformations of Eq. (16) are required. These types of transformations follow from the fact that the conditions of stationarity of the solution to Eq. (17) coincide with the conditions of its stability, i.e., the process is nonstationary when

$$|G_i| \geq 1 \tag{18}$$

where G_i are the roots of the characteristic equation associated with Eq. (17).

The case $|G_i| \geq 1$ usually is excluded from considerations since it corresponds to an exponential instability that is unrealistic in physical systems under observation. However, the case $|G_i| = 1$ is realistic. Real and complex conjugates of G_i incorporate trend and seasonal (periodic) components, respectively, into the time series Eq. (16).

By applying as many times as required a difference operator,

$$\nabla x(t) = x(t) - x(t-1) = (1 - B)x(t) \tag{19}$$

where B is defined as the backward shift operator, one can eliminate the trend from the time series:

$$x(t), x(t-1), x(t-2), \cdots \tag{20}$$

Similarly, the seasonal components from the time series Eq. (20) can be eliminated by applying the seasonal difference operator:

$$\nabla_s x(t) = (1 - B^s)x(t) = x(t) - x(t-s) \tag{21}$$

In most cases, the seasonal differencing, Eq. (21), should be applied prior to standard differencing, Eq. (19).

Unfortunately, it is not known in advance how many times the operators, Eq. (19) or (21), should be applied to the original time series Eq. (20) for their stationarization. Moreover, in Eq. (21) the period s of the seasonal difference operator also is not prescribed. However, several methods have been developed to estimate the order of differentiation [4]. One simple estimate of the number of operations for Eq. (20) is the minimization of the area under the autocorrelation curve.

Once the time series Eq. (20) is stationarized, one can apply to it the model Eq. (15):

$$y(t) = F\big(y(t-1), y(t-2), \cdots, y(t-m)\big) \tag{22}$$

where

$$y(t), y(t-1), \cdots; \big(y(t) = x(t) - x(t-1)\big) \tag{23}$$

are transformed series, Eq. (20), and $\tau = 1$. After fitting the model Eq. (22) to the time series Eq. (20), one can return to the old variable $x(t)$ by exploiting the inverse operators $(1-B)^{-1}$ and $(1-B^s)^{-1}$. For instance, if the stationarization procedure is performed by the operator Eq. (19), then

$$x(t) = x(t-1) + F\Big(\big[x(t-1) - x(t-2)\big], \big[x(t-2) - x(t-3)\big], \cdots\Big) \tag{24}$$

Equation (24) can be utilized for modeling the residual,[2] predictions of future values of Eq. (20), and detection of structural abnormalities. However, despite the fact that Eqs. (22) and (24) may be significantly different, their structures are uniquely defined by the same function, F. Therefore, structural abnormalities that cause changes of the function F can also be detected from Eq. (22), and, consequently, for that particular purpose the transition to Eq. (24) is not necessary.

It should be noted that, strictly speaking, the application of the stationarization procedure, Eqs. (19) and (21), to the time series Eq. (20) is justified only if the underlying model is linear since the criteria of stationarity for nonlinear equations are more complex than for linear ones in the same way as the criteria of stability are. Nevertheless, there is numerical evidence that, even in nonlinear cases, the procedures of Eqs. (19) and (21) are useful in the sense that they significantly reduce the error, i.e., the difference between the simulated and the recorded data if the latter are nonstationary.

V. Model Fitting

The models, Eqs. (22) and (24), that have been selected in the previous section for detection of structural abnormalities in the time series Eq. (20) have the following parameters to be found from Eq. (20): the function, F; the time delay, τ; the order of time delays, m; the powers, m_1 and m_2, of the difference, $(1-B)^{m_1}$, and the seasonal difference, $(1-B^s)^{m_2}$; and the period, s, of the seasonal operator.

The form of the function F we have selected for the residual is shown in Fig. 3. After stationarization, the linear component is fit using the Yule–Walker equations [4], which define the autoregressive parameters a_i in Eq. (17) via the autocorrelations in Eq. (20). If sufficient, the residual left after removal of the linear component, $w(t)$, can be directly analyzed and modeled as noise.

If the linear model of the residual leads to poor model fitting, the best tool for fitting the nonlinear component of the residual may be a feed-forward neural network that approximates the true extrapolation mapping by a function parameterized by the synaptic weights and thresholds of the network. A rigorous proof [6] states that any continuous function can be approximated by a feed-forward neural net with only one hidden layer and, thus, is selected for fitting the nonlinear component after the linear component is removed using Eq. (17). Hence, $w(t)$ is sought in the following standard form of a time-delay feed-forward network:

$$z(t) = \sigma\left\{\sum_{j=1} W_{1j}\sigma\left[\sum_{k=1}^{m} w_{jk} z(t - k\tau)\right]\right\} \tag{25}$$

where W_{1j} and w_{jk} are constant synaptic weights, and $\sigma(x) = \tan h(x)$ is the sigmoid function.

[2] The residual $\mathbf{r}(t)$ is assumed to be in the form of a discrete-time series. This is a valid assumption given that the gray box will be implemented on a digital computer.

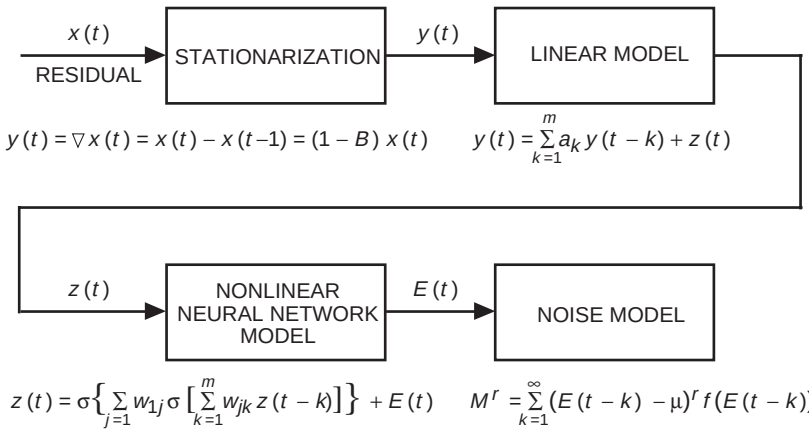

Fig. 3. Description of the residual data.

The model fitting procedure is based upon minimization of the mean standard error:

$$E\left(W_{1j}, w_{jk}\right) = \sum_{i} \left(z(t-i) - \sigma \left\{ \sum_{j=1} W_{1j} \sigma \left[\sum_{k=1}^{m} w_{jk} z(t - k\tau - i) \right] \right\} \right)^2 \tag{26}$$

The error measure Eq. (26) consists of two parts:

$$E = E_1 + E_2 \tag{27}$$

where E_1 represents the contribution of a physical noise, while E_2 results from nonoptimal choice of the parameters of the model Eq. (25).

There are two basic sources of random components in E_1. The first source is chaotic instability of the underlying dynamical system; in principle, this component of E_1 is associated with instability of the underlying model, and it can be represented based upon the stabilization principle introduced by [5]. The second source is the physical noise, imprecision of the measurements, or human factor, such as multi-choice decisions in economical or social systems, one's driving habits in the case of the catalytic converter of a car, etc.

The last component of E_1 cannot be presented by any model based upon classical dynamics, including Eq. (22). However, as shown by [5], there are models based upon a special type of dynamics called terminal, or non-Lipschitz, dynamics that can simulate this component. In the simplest case, one can assume that E_1 represents a variance of a mean-zero Gaussian noise.

The component E_2, in principle, can be eliminated by formal minimization of the error measure Eq. (26) with respect to the parameters W_{1j}, w_{jk}, τ, m, m_1, m_2, and s. Since there is an explicit analytical dependence between E and W_{1j}, w_{jk}, the first part of minimization can be performed by applying back propagation. However, further minimization should include more sophisticated versions of gradient descent since the dependence $E(\tau, m, m_1, m_2, s)$ is too complex to be treated analytically.

VI. Anomaly Detection

As discussed in the previous section, there are two causes for abnormal behavior in the solution to Eq. (25): (1) changes in external forces or initial conditions (these changes can be measured by Lyapunov

stability and associated with operational abnormalities) and (2) changes in the parameters W_{1j}, w_{jk}, i.e., changes in the structure of the function F in Eq. (22). (These changes are measured by structural stability and associated with structural abnormalities. They can be linked to the theory of catastrophe.)

The measure we use for anomaly detection in the nonlinear component is

$$\zeta = \sum \left[\left(W_{1j} - \overset{o}{W}_{1j} \right)^2 + \left(w_{ij} - \overset{o}{w}_{ij} \right)^2 \right] \tag{28}$$

where $\overset{o}{W}_{1j}$ and $\overset{o}{w}_{ij}$ are the nominal, or "healthy," values of the parameters and W_{1j}, w_{jk}, are their current values. If

$$\zeta = |\varepsilon| \tag{29}$$

where ε is sufficiently small, then there is no structural abnormality. The advantage of this criterion is in its simplicity. It can be periodically updated, and, therefore, the structural health of the process can be easily monitored.

Similar criteria can be generated for the parameters of the linear component, a_j, and the noise component that is modeled by the variance or higher moments. Unfortunately, there is no general method for setting the threshold, ε, other than experience and heuristic methods. This is a problem faced by all fault diagnosis.

VII. Conclusion

In this article, we present a new method called the gray-box method for model-based system diagnosis. It is a hybrid model incorporating elements from residual-based methods and parametric-estimation methods. The residual is generated by filtering the measured state variable with those predicted by the system model. The residual is modeled by a three-tier stochastic model. The linear and nonlinear components of the residual are described by an autoregressive process and a time-delay feed-forward neural network, respectively. The last component, the noise, is characterized by its moments.

The faults are detected by monitoring the parameters of the autoregressive model, the weights of the neural network, and the moments of noise. The method is applicable to both linear and nonlinear systems, and computer simulations are being conducted to validate the method in practice.

References

[1] J. Chen and R. J. Patton, *Robust Model-Based Fault Diagnosis for Dynamic Systems*, Boston, Massachusetts: Kluwer Academic Publishers, 1999.

[2] R. N. Clark, "State Estimation Schemes for Instrument Fault Detection," in *Fault Diagnosis in Dynamic Systems: Theory and Application*, edited by R. J. Patton, P. M. Frank, and R. N. Clark, New York: Prentice Hall, pp. 21–45, 1989.

[3] F. Takens, "Detection Strange Attractors on Turbulence" and "Dynamical Systems and Turbulence," *Lecture Notes in Mathematics*, vol. 898, pp. 366–381, Berlin, Germany: Springer-Verlag, 1980.

[4] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis*, Upper Saddle River, New Jersey: Prentice Hall, 1994.

[5] M. Zak, "Postinstability Models in Dynamics," *Int. J. of Theoretical Physics*, vol. 33, no. 77, pp. 2215–2218, 1994.

[6] J. Hertz, A. Krough, and R. G. Palmer, *Introduction to the Theory of Neural Computations*, Redwood City, California: Addison-Wesley, 1991.

[7] M. O. Wold, *A Study in the Analysis of Stationary Time Series*, second edition, Uppsalla, Sweden: Almqvist and Wiksell, 1954.