19th Mediterranean Conference on Control and Automation
Aquis Corfu Holiday Palace, Corfu, Greece
June 20-23, 2011

Fault Modelling Using a Mixture of Conditional Gaussian Transitions

Dejan P. Jovanović
Department of Mathematics
The University of Queensland
Qld 4072 AUSTRALIA
Email: dejan.jovanovic@uqconnect.edu.au

Ross S. McVinish
Department of Mathematics
The University of Queensland
Qld 4072 AUSTRALIA
Email: r.mcvinish@uq.edu.au

Philip K. Pollett
Department of Mathematics
The University of Queensland
Qld 4072 AUSTRALIA
Email: pkp@maths.uq.edu.au

***Abstract*—To model a fault that can be caused by more than one source, a mixture of conditional Gaussian transitions is proposed. The conditional means are modelled by recurrent neural networks. An expectation-maximization (EM) algorithm is used to estimate model parameters. By grouping known types of faults it is possible to form a bank of different fault models.**

***Index Terms*—Fault modelling, mixture of Gaussian transitions, neural networks, EM algorithm.**

I. Introduction

The model-based fault detection methods are a very important class of techniques which can reveal unwanted deviations in system characteristics [19]. There are two kinds of model: process models and signal models. The first is used when both the input and output variables of the system are available, and the second when only output measurements are available. Once an appropriate model is adopted it can be used to generate residuals, being the difference between observed signal values and model predicted values. Once evaluated, the residuals can provide information about system behaviour.

In the existing literature on fault diagnosis, a variety of different process and signal models have been proposed [3], [10], [20], [23]. However, a model involving a mixture of Gaussian transitions has not yet been applied in the context of fault diagnosis. Using this approach, we propose a methodology for designing a bank of models. The proposed algorithm can be applied as either a process model or as a signal model.

Our approach is similar to one introduced by Newbold and Ho [18], and studied further by Hanlon and Maybeck [6] and Semoushin *et al.* [22]. The Kalman filter bank is applied to diagnose multiple faults in a system and a general likelihood ratio test (GLRT) is used to detect changes in system behaviour. We apply these ideas to a model with a mixture of Gaussian transitions in order to deal with various operational modes. However, instead of dealing only with first-order and second-order moments, we propose to represent the model in terms of conditional distributions. One advantage that a mixture of conditional Gaussian distributions has over other models [6], [18], [22] is that nonlinear effects due to system faults can be captured [14]. We use recurrent neural networks [8] as our model for the nonlinear effects in the mean of the signal. This approach allows different causes of a single fault to be modelled.

II. Mixture of Gaussian Transitions

The signals produced by most faults are nonstationary, nonlinear time series. This makes modelling difficult if a particular fault has multiple sources. There are two difficulties in resolving problems associated with these signals: selection of an appropriate model capable of capturing multi modality and estimating the parameters of the selected model. Our approach is to use a mixture of conditional Gaussian transitions, as follows:

$$F(y_t|\mathcal{F}_{t-1}) = \sum_{k=1}^{K} \alpha_k \Phi\Big(\frac{y_t - h_k(\mathcal{F}_{t-1})}{\sigma_k}\Big), \qquad (1)$$

where $F(y_t|\mathcal{F}_{t-1})$ is the conditional cumulative distribution function (cdf) of the observation y_t at time t, $\mathcal{F}_{t-1}$ is a dependence vector (a vector of previous samples) taking values in $\mathbb{R}^m$ with $m = |\mathcal{F}|$, $\Phi(\cdot)$ is the cdf of the standard normal distribution, α_k is a mixture weighting coefficient, h_k is a non-linear mapping $h_k : \mathbb{R}^m \to \mathbb{R}$, σ_k^2 is the variance of component k and K is a number of mixture components. Equation (1) is a generalization of the finite mixture model discussed in [4], [16], [24]. Our mixture representation was chosen because (i) estimating parameters for this type of model is possible using the EM framework, (ii) the most likely fault type can be detected by monitoring changes in the likelihood ratio for given set of observations, and (iii) using Monte Carlo techniques, the future behaviour of the signal can be predicted. Whilst our model has several attractive features, there are some drawbacks. First, the choice of h_k is not necessarily obvious, and will depend on the problem at hand. Second, h_k can be nonlinear in its parameters. And, third, there are significant problems associated with estimating K, the order of dependence in $\mathcal{F}_{t-1}$ and monitoring the standard error.

Our work builds on that of Le *et al.* [14] and Wong and Li [26]. Both papers deal with mixtures of conditional Gaussian distributions. The first is a continuation of work of Raftery [21]. It explains how this type of mixture can

978-1-4577-0123-8/11/$26.00 ©2011 IEEE

model nonlinear time series, particularly flat stretches, bursts of activity and change points, using a simple autoregressive model (AR) of the first order ($AR(1)$) for the conditional mean. The second generalizes this to an autoregressive model of arbitrary order ($AR(p)$).

However, the linear AR model has some disadvantages [25]. Of particular interest from perspective of fault modelling, are the following restrictions. The AR model is not suitable for time series exhibiting sudden bursts of large amplitude. Moreover it deals only with symmetric data. In addition, since a linear difference equation does not have stable periodic solutions, independent of initial conditions, the AR model cannot account for *limit cycles*.

Notwithstanding these limitations, (1) is a natural generalization of the linear AR model to the case of nonlinear $h_k(\mathcal{F}_{t-1})$. Let it be assumed that process of signals of the monitored system follow a stochastic difference equation of the form

$$y_t{=}h_k(\mathcal{F}_{t-1}){+}\epsilon_{t,k}, \tag{2}$$

where h_k is an unknown smooth function governing the dynamic behaviour of component k, $\epsilon_{t,k}$ are iid with zero mean and finite variance σ_k^2, having probability density function $f_k(\cdot)$ (a standard normal distribution), and $\{y_t, t \in \mathbb{Z}_+\}$ is a time series on an arbitrary space generated by (1).

There is a variety of possible approximations of the deterministic *skeleton* $h_k(\mathcal{F}_{t-1})$ in (2), for example, the class of *linear basis expansion* models with *splines* or *wavelets* as bases [7]. However, given that our primary objective is time series modelling, a parametric model $h_k(\mathcal{F}_{t-1})$ is required that must be capable of embedding temporal sequences. It must be capable of making temporal association, that is, to generate a sequence in response to a particular input sequence, and it must have the capacity to reproduce a sequence when it observes part of it. The first is related to process models, while the second is connected with signal models within the model-based fault diagnosis framework. For the case of a process model, the dependence vector $\mathcal{F}_{t-1}$ is defined as $\mathcal{F}_{t-1}{=}\{y_{t-1}, \ldots, y_{t-p}, u_{t-1}, \ldots, u_{t-r}\}$. For a signal model it is given by $\mathcal{F}_{t-1}{=}\{y_{t-1}, \ldots, y_{t-p}\}$, where y_{t-i} is an output signal and u_{t-i} a control signal of the monitored system.

To meet the requirements imposed on $h_k(\mathcal{F}_{t-1})$, two types of neural networks are proposed. For the process models, a recurrent neural network (RNN) is suggested, described by a discrete-time nonlinear model of the form [12]

$$x_i(t){=}\lambda_i x_i(t{-}1){+}\beta_i a\Big[\sum_{j=1}^{n} w_{ij}x_j(t{-}1){+}s_i\Big], \tag{3}$$

where n is the number of neurons, x_i is the state of the i^{th} neuron, $a[\cdot]$ is a nonlinear activation function, $\mathbf{W} = [w_{ij}]_{n\times n}$ is the matrix of connection weights and s_i is a constant input (*bias*). A time constant λ_i models node dynamics and it is assumed that $-1 \le \lambda_i \le 1$. The node gain β_i can take any non-zero value. The neurons that receive an external input do not have node dynamics, but the neurons that represent an output from the network possible do.

For the signal models, a time delay neural network (TDNN) is considered. It differs from (3) in that $\lambda_i = 0$ (no node dynamics). We analyse stability conditions of RNN in order to reveal model properties. Since TDNN is a subclass of RNN, our results for RNN apply directly to TDNN.

An application of RNN in the context of *Mixture of Experts* was given in [13], but the statistical properties of RNN were not considered.

III. Model Properties

The conditional Gaussian mixture (1) has two important properties. Since the conditional mean of each component depends on previous samples, a conditional distribution can change its shape and from unimodal it can become multimodal. The conditional expectation of y_t given previous samples $\mathcal{F}_{t-1}$ is [26]

$$E(y_t|\mathcal{F}_{t-1}){=}\sum_{k=1}^{K}\alpha_k h_k(\mathcal{F}_{t-1}), \tag{4}$$

where $h_k(\mathcal{F}_{t-1})$ is a nonlinear parametric model. It is worth noting that the accuracy of prediction depends on the model used for the conditional mean. Additionally, (1) has the conditional variance [26]

$$var(y_t|\mathcal{F}_{t-1}){=}\sum_{k=1}^{K}\alpha_k\sigma_k^2{+}\sum_{k=1}^{K}\alpha_k h_k^2(\mathcal{F}_{t-1}) - \Big(\sum_{k=1}^{K}\alpha_k h_k(\mathcal{F}_{t-1})\Big)^2. \tag{5}$$

By exploiting these properties in the context of fault modelling, a multiple source fault can be modelled, one-step prediction can be performed, and its corresponding variance calculated. However, asymptotic stationarity of the time series $\{y_t, t \in \mathbb{Z}_+\}$ is required. We establish this condition using results of Meyn and Tweedie [17], Chan and Tong [1] and Wassim and Bernstein [5].

For a model (2), let $(\mathbb{R}^m, \mathfrak{B}^m, \mu_m)$ be a probability space, where $\mathfrak{B}^m$ are the Borel sets of $\mathbb{R}^m$ and μ_m is Lebesgue measure on $\mathbb{R}^m$. Furthermore, it is postulated that $\epsilon_{t,k}$ are iid with positive probability density f_k. For some $\mathcal{F}{\in}\mathbb{R}^m$ and $A{\in}\mathfrak{B}^m$, the transition probability function $P(\mathcal{F}, A)$ of $\{y_t\}$ is given by

$$P(\mathcal{F}, A){=}\sum_k \alpha_k \int_{A-h_k(\mathcal{F})} f_k(t)\mu_m(dt). \tag{6}$$

Similarly, $P^{(n)}(\mathcal{F}, A){=}P(\mathcal{F}_{t+n}{\in}A|\mathcal{F}_t{=}\mathcal{F})$ are the n-step transition probabilities with $P^{(1)}{=}P$. In this way, a time series $\{y_t, t{\in}\mathbb{Z}_+\}$ with a transition probability $P(\mathcal{F}, A)$ can be viewed as a Markov chain on $(\mathbb{R}^m, \mathfrak{B}^m, \mu_m)$.

There are three important properties of Markov chains formulated in [17] that are relevant to the problem discussed here. The first is φ-irreducibility. The Markov chain is said to be φ-irreducible if, for some finite measure $\varphi(A)>0$, there is an $n>0$ such that $P^{(n)}(\mathcal{F},A)>0$ for all $\mathcal{F}\in\mathbb{R}^m$. In other words, all parts of the state space can be reached whatever the initial point. The next is aperiodicity: there is no regular pattern in return times to states. Finally, a property that gives the rate of convergence and makes a link with the deterministic part of the model (2) is *geometric ergodicity*. For any measure π on $(\mathbb{R}^m,\mathfrak{B}^m,\mu_m)$, let $\|\cdot\|$ denote the total variation of π. A Markov chain is *geometrically ergodic* if there is a constant $\rho>1$ such that

$$\lim_{n\to\infty}\rho^n\|P^{(n)}(\mathcal{F},A)-\pi(A)\|=0,\quad \forall\mathcal{F}\in\mathbb{R}^m, \tag{7}$$

where π is invariant probability measure, that is,

$$\pi(A)=\int P(\mathcal{F},A)\pi(d\mathcal{F}),\quad \forall A\in\mathfrak{B}^m. \tag{8}$$

If the associated Markov chain is geometrically ergodic, then the distribution of the time series will converge to π geometrically quickly, in which case it is said to be *asymptotically stationary*.

For asymptotic stationarity, geometric ergodicity must be proved for each component in (6). We use a result of Chan and Tong [1] to establish a connection between geometric ergodicity of $\{y_t\}$ and the existence of a Lyapunov function of the parametric model $h_k(\mathcal{F}_{t-1})$ in (2). Once a Lyapunov function is identified for $h_k(\mathcal{F}_{t-1})$, geometric ergodicity follows from Theorem 4.2 and Section 5 of [1].

To prove the existence of a Lyapunov function for (3), re-write it as

$$\begin{aligned}\mathbf{X}_t&=\mathbf{\Lambda}\mathbf{X}_{t-1}+\mathbf{B}a(\mathbf{H}_{t-1})\\ \mathbf{H}_{t-1}&=\mathbf{W}\mathbf{X}_{t-1},\end{aligned} \tag{9}$$

where $\mathbf{X}$ is the state vector $[x_i]_{n\times 1}$, $\mathbf{\Lambda}$ and $\mathbf{B}$ are the diagonal matrices $[diag(\lambda_i)]_{n\times n}$ and $[diag(\beta_i)]_{n\times n}$, respectively, $\mathbf{H}$ is the output vector, $\mathbf{W}$ is the matrix $[w_{ij}]_{n\times n}$ of connection weights and $a(\cdot)$ is a bounded nonlinear function. Based on results in [5] (Theorem 3.1 and Theorem 4.1) it follows that for (9) there is a Lyapunov function of the form $V(\mathbf{x})=\mathbf{x}^T\mathbf{P}\mathbf{x}$, where $\mathbf{P}$ is a positive-definite matrix for which $\Delta V(\mathbf{x})<0$. Particular constraints imposed on $\mathbf{\Lambda}$, $\mathbf{B}$ and $\mathbf{W}$ are given in [12].

We deduce that the time series $\{y_t, t\in\mathbb{Z}_+\}$, generated by the Gaussian transition model with a conditional mean modelled by RNN, is asymptotically stationary, connoting that (4) and (5) take only finite values.

IV. Estimation

The EM algorithm [2] is used to estimate model parameters. Suppose that a set of observations $\{y_t, t\in\mathbb{Z}_+\}$ is generated by (1). An unobservable random choice of component k, that originates an observation y_t, is modelled by random indicator variable

$$I_{k,t}:=\begin{cases}1 & \text{if } Z_t=k,\\ 0 & \text{otherwise.}\end{cases}$$

The set of indicators $Z=\{I_{k,t}:\ k=1,\dots,K;\ t=1,\dots,T\}$ creates the missing data. Consequently this turns an incomplete data problem into a complete data problem and allows the application of the EM algorithm. Since the values of $I_{k,t}$ are unknown, it is necessary to evaluate their expectations given the observations and parameters. Once the expectation of $I_{k,t}$ has been calculated, the model parameters of $h_k(\mathcal{F}_{t-1})$ are updated. This describes one iteration of the EM algorithm. The set of parameters is $\Psi=\{\alpha_1,\dots,\alpha_K;\mathbf{W}_1,\dots,\mathbf{W}_K;\sigma_1^2,\dots,\sigma_K^2;\}$, where α_k is a mixture proportion, $\mathbf{W}_k$ is a matrix of neural network weights, σ_k^2 is the variance of mixture component k.

The likelihood function of the complete data is given by

$$\begin{aligned}L_c(\Psi|Y,Z)&=p(Y,Z|\Psi)\\ &\prod_{t=p+1}^{T}\prod_{k=1}^{K}\Big[p(z_t;\Psi)p(y_t|z_t,\mathcal{F}_{t-1};\Psi)\Big]^{I_{k,t}},\end{aligned} \tag{10}$$

where Y, Z and Ψ denote observations, the missing data and the parameters of the model, respectively.

The E-step involves computing the conditional expectation of the complete log likelihood (10), often called the Q function, and defined as follows:

$$\begin{aligned}Q(\Psi|\Psi^i)&=\mathbb{E}_{\Psi^i}\{\log L_c(\Psi|Y,Z)\}\\ &=\mathbb{E}_{\Psi^i}\{\sum_{t=p+1}^{T}\sum_{k=1}^{K}I_{k,t}\{\log\alpha_k+\log p(y_t|Y,z_t,\mathcal{F}_{t-1};\Psi)\}\}\\ &=\sum_{t=p+1}^{T}\sum_{k=1}^{K}\mathbb{E}_{\Psi^i}\{I_{k,t}|Y,\Psi\}\times\\ &\qquad\times\{\log\alpha_k+\log p(y_t|Y,z_t,\mathcal{F}_{t-1};\Psi)\},\end{aligned} \tag{11}$$

where $\mathbb{E}_{\Psi^i}\{I_{k,t}|Y,\Psi\}$ is the expectation of the hidden variable Z conditional on observation Y and a set of a model parameters Ψ. Since $\log L_c(\Psi|Y,Z)$ is a linear function of an unobservable random variable Z, the E-step requires calculating the conditional expectation of Z given the observations [16]:

$$\begin{aligned}\mathbb{E}\{I_{k,t}|Y,\Psi\}&=\mathbb{P}\{z_t|y_t,\mathcal{F}_{t-1};\Psi\}\\ &=\frac{p(y_t|z_t,\mathcal{F}_{t-1};\Psi)p(z_t|\Psi)}{p(y_t|\Psi)}\\ &=\frac{\alpha_k p(y_t|z_t,\mathcal{F}_{t-1};\Psi)}{\sum_{j=1}^{K}\alpha_j p(y_t|z_t,\mathcal{F}_{t-1};\Psi)}\\ &=\frac{\alpha_k f_k(\hat{\epsilon}_{t,k};\sigma_k^2)}{\sum_{j=1}^{K}\alpha_j f_j(\hat{\epsilon}_{j,t};\sigma_j^2)}\\ &=\tau_{t,k}.\end{aligned} \tag{12}$$

where $\hat{\epsilon}_{i,t}$ is the current estimate of the residual for a component i, defined by $\hat{\epsilon}_{i,t}=y_t-h_i(\mathcal{F}_{t-1})$.

The EM algorithm requires $Q(\Psi|\Psi^i)$ to be maximized. To facilitate this, it can be rewritten as

$$Q(\Psi|\Psi^i)=\sum_{t=p+1}^{T}\sum_{k=1}^{K}\tau_{t,k}\log\alpha_k-\sum_{t=p+1}^{T}\sum_{k=1}^{K}\tau_{t,k}\log\sigma_k - \sum_{t=p+1}^{T}\sum_{k=1}^{K}\frac{\tau_{t,k}}{2\sigma_k^2}\hat{\epsilon}_{t,k}^2. \quad (13)$$

It is maximized over α_k and σ_k subject to the necessary constraints using the Lagrange multiplier approach. This leads to the following updating equations for α_k and σ_k:

$$\alpha_k=\frac{\sum_{t=p+1}^{T}\tau_{t,k}}{T-p}, \quad (14)$$

$$\sigma_k=\sqrt{\frac{\sum_{t=p+1}^{T}\tau_{t,k}\hat{\epsilon}_{t,k}^2}{\sum_{t=p+1}^{T}\tau_{t,k}}}. \quad (15)$$

For models $h_k(\mathcal{F}_{t-1})$ that are linear in the parameters, such as the AR models and linear basis expansion models, $Q(\Psi|\Psi^i)$ can be maximized directly. However, for the neural network models considered here, maximization of $Q(\Psi|\Psi^i)$ requires an iterative technique. Taking derivatives with respect to the parameters of $h_k(\mathcal{F}_{t-1})$, the following expression is obtained:

$$\frac{\partial Q(\Psi|\Psi^i)}{\partial w_{ij}^k}=\frac{\partial}{\partial w_{ij}^k}\{-\sum_{t=p+1}^{T}\sum_{k=1}^{K}\tau_{t,k}\frac{(y_t-h_k(\mathcal{F}_{t-1}))^2}{2\sigma_k^2}\} = \sum_{t=p+1}^{T}\frac{\tau_{t,k}}{\sigma_k^2}\hat{\epsilon}_{t,k}\frac{\partial h_k(\mathcal{F}_{t-1})}{\partial w_{ij}^k}. \quad (16)$$

Equation (16) resembles the *back-propagation* (BP) learning algorithm in the case when the squared error cost function is used. However, there is an important difference. In contrast to original expression for BP, this one, given by (16), is modulated by $\tau_{t,k}$ (the expectation of a latent variable) and σ_k^2 (variance of component k). As in BP, using the gradient-descent method, the weight is updated using

$$\Delta w_{ij}^k=\sum_{t=p+1}^{T}\frac{\tau_{t,k}}{\sigma_k^2}\hat{\epsilon}_{t,k}\frac{\partial h_k(\mathcal{F}_{t-1})}{\partial w_{ij}^k}. \quad (17)$$

From this it is apparent that the M-step is batch gradient learning.

Most of our work here is devoted to modelling a single fault coming from more than one source. Consequently the EM based framework is applied. However, if there is a single source, then it is still possible to use maximum likelihood methodology for neural network training. By simple arithmetic starting from the likelihood function

$$L_c(\Psi|Y)=\prod_{t=p+1}^{T}p(y_t|\mathcal{F}_{t-1};\Psi), \quad (18)$$

it is possible to show that the maximum likelihood based training procedure gives the following equations for batch learning within one step:

$$\Delta w_{ij}=\sum_{t=p+1}^{T}\frac{\hat{\epsilon}_t}{\sigma^2}\frac{\partial h(\mathcal{F}_{t-1})}{\partial w_{ij}} \quad (19)$$

$$\sigma=\sqrt{\frac{\sum_{t=p+1}^{T}\hat{\epsilon}_t^2}{T-p}}. \quad (20)$$

where $\hat{\epsilon}_t$ is the current estimate of the residual, given by $\hat{\epsilon}_t=y_t-h(\mathcal{F}_{t-1})$.

Despite the fact that estimation of parameters is restricted to the BP algorithm, there are still types of neural network that can be successfully applied in model-based fault modelling. We conclude that the methodology can be applied to recurrent networks whose training is based on BP. The types of RNN that can be used are the fully and partially RNN-like *recurrent back-propagation network*, *Jordan sequential network* and the *simple recurrent network* [8], including TDNN.

V. Simulation Results

We illustrate the algorithm by way of an example of nonlinear vibrations in engineering systems; we analyze the phenomenon of amplitude-dependent frequency [9],[11],[25]. The frequency of the observed signal depends on the amplitude of the excitation signal, which is composed of deterministic and stochastic components. Considering the amplitude-dependent frequencies as undesirable, we require a fault model capable of capturing multi-modality. We will assume that the change in signal amplitude (frequency) occurs at random.

For simplicity, assume that there are only two amplitude levels. A fault model is represented by two self-exiting autoregressive (SETAR) models, M_1 and M_2, given by (21) and (22), respectively (see [25], Section 2.14.2):

$$y_t=\begin{cases}1.6734-0.8295y_{t-1}+0.1309y_{t-2}-0.0276y_{t-3}+\epsilon_t & \text{if } y_{t-1}>0.5\\ 1.2270+1.0516y_{t-1}-0.5901y_{t-2}-0.2149y_{t-3}+\epsilon_t & \text{if } y_{t-1}\le 0.5\end{cases} \quad (21)$$

$$y_t=\begin{cases}0.30-0.80y_{t-1}+0.20y_{t-2}-0.7y_{t-3}+\epsilon_t & \text{if } y_{t-1}>3.05\\ 0.15+0.85y_{t-1}+0.22y_{t-2}-0.7y_{t-3}+\epsilon_t & \text{if } y_{t-1}\le 3.05.\end{cases} \quad (22)$$

The noise that drives these models has variance $var(\epsilon_t)=(0.005)^2$. Switching between models M_1 and M_2 occurs according to a two-state Markov chain with transition probabilities $p_{11}=p_{22}=0.95$. The training and test sets are created using simulated values. To improve neural network training, simulated values were scaled to the range $y_t\in[-0.5,0.5]$, and the training set consisted of $T=9000$ samples. Since for a given model only output measurements are available, this is obviously an example

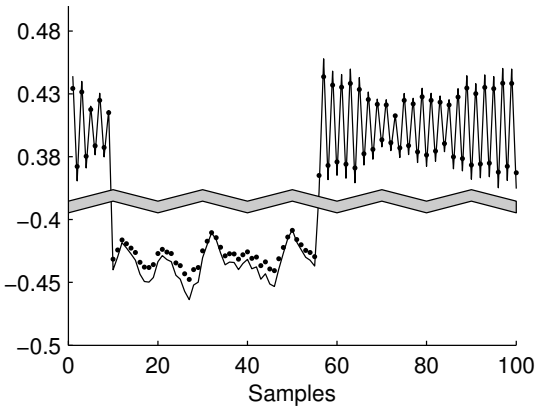

Fig. 1. *Test signal (solid) and model prediction (dotted)*

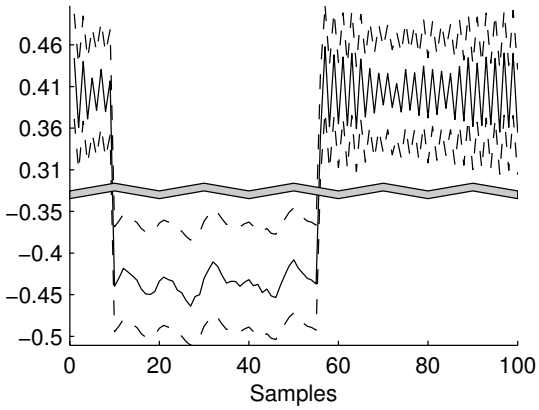

Fig. 3. *90% Prediction interval*

of the *signal model* fault diagnosis approach, discussed in Section I. Consequently, TDNN are used to approximate the nonlinear conditional mean. In order to capture model dynamics, given by (21) and (22), a two component mixture (K=2) of conditional Gaussian transitions is used.

Accordingly, the TDNNs have one delayed input (p=1) each, two hidden layers with N_1=30 and N_2=25 neurons in each layer, and one output. A *hyperbolic tangent* with constant parameter λ=0.75 was used as an activation function. There were relatively few it-

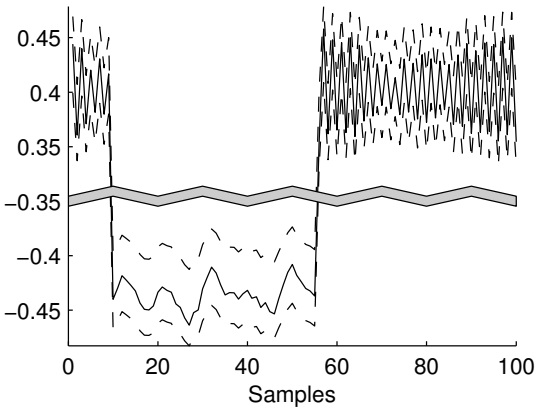

Fig. 2. *60% Prediction interval*

erations of the EM algorithm before $\Delta Q(\Psi|\Psi^i) \leq 10^{-3}$ (28 in total). In the first 25 iterations neural networks were trained with learning parameters η_1^i=$0.75\,10^{-5}, \eta_2^i$=$0.25\,10^{-5}, \eta_3^i$=$0.125\,10^{-5}$, $i=1,2$, while the remaining 3 learning parameters were reduced to η_1^i=$0.0313\,10^{-5}, \eta_2^i$=$0.0625\,10^{-5}, \eta_3^i$=$0.1875\,10^{-5}$, $i=1,2$. Finally, for the given training set, the mixture proportions were α_1=0.4237 and α_2=0.5763 for components $k=1,2$, respectively, while the variances for each of component were σ_1^2=$(0.0198)^2$ and σ_2^2=$(0.0456)^2$.

The resultant model was tested on a given test set. To have clearer graphs, we have broken ordinates in Figs. 1, 2 and 3 respectively, ignoring irrelevant parts of the signal. The top part of the figures show signals related to model M_1, while the lower part of the figures represent that relevant to model M_2.

Fig. 1 shows the first 100 samples of original test signal and the corresponding one-step ahead predictions obtained by (4). Swaps happened at the sample indices 15, 33 and 79. The prediction abilities of our model are illustrated in Figs. 2 and 3, where 60% and 90% prediction intervals are given. The dashed lines represent upper and lower boundaries, while the solid line represents the test signal.

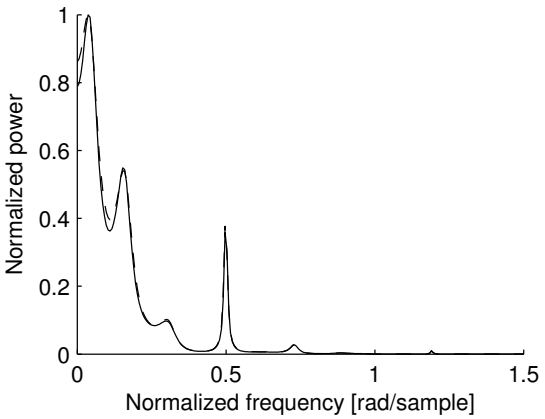

Fig. 4. *The MUSIC of test (solid) and predicted (dashed) signal*

Comparative frequency characteristics of the test and predicted values of the signal, using the multiple signal classification (*MUSIC*) [15] method, are given in Fig. 4. It is apparent that the corresponding model captured signal dynamics with negligible deviation from the original signal.

VI. Discussion

In this paper an approach to modelling a multiple source fault was proposed, based on a mixture of conditional Gaussian transitions. The conditional means were approximated by neural networks, allowing successful modelling of nonlinear dynamics. Parameters were estimated using maximum likelihood. Future research will focus on full implementation of a bank of fault detection filters using conditional Gaussian transitions.

VII. Acknowledgment

This work was supported by the Australian Research Council Centre of Excellence for Mathematics and Statistics of Complex Systems (MASCOS).

References

[1] K.S. Chan and H. Tong. On the use of the deterministic Lyapunov function for the ergodicity of stochastic difference equations. *Advances in Applied Probability*, 17(3):666–678, 1985.

[2] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[3] S.X. Ding. *Model-based fault diagnosis techniques: design schemes, algorithms, and tools*. Berlin, Heidelberg : Springer-Verlag Berlin Heidelberg, 2008.

[4] S. Fruhwirth-Schnatter. *Finite mixture and Markov switching models*. New York : Springer, 2006.

[5] W.M. Haddad and D.S. Bernstein. Explicit construction of quadratic Lyapunov functions for the small gain, positivity, circle, and Popov theorems and their application to robust stability, Part II: Discrete-time theory. *International Journal of Robust and Nonlinear Control*, 4(2):249–265, 1994.

[6] P.D. Hanlon and P.S. Maybeck. Equivalent Kalman filter bank structure for multiple model adaptive estimation (MMAE) and generalized likelihood ratio (GLR) failure detection. In *Decision and Control, 1997, Proceedings of the 36th IEEE Conference on*, volume 5, pages 4312–4317, December 1997.

[7] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning*. Springer, 2003.

[8] J. Hertz, A. Krogh, and R.G. Palmer. *Introduction to the theory of neural computation*. Redwood City, Calif. : Addison-Wesley, 1990.

[9] W. Hu and N.M. Wereley. Hybrid magnetorheological fluid-elastomeric lag dampers for helicopter stability augmentation. *Smart Materials and Structures*, 17(4):045021, 2008.

[10] R. Isermann. Model-based fault-detection and diagnosis - status and applications. *Annual Reviews in Control*, 29(1):71–85, 2005.

[11] T. Jeong and R. Singh. Inclusion of measured frequency- and amplitude-dependent mount properties in vehicle or machinery models. *Journal of Sound and Vibration*, 245(3):385–415, 2001.

[12] L. Jin, P.N. Nikiforuk, and M.M. Gupta. Absolute stability conditions for discrete-time recurrent neural networks. *Neural Networks, IEEE Transactions on*, 5(6):954–964, 1994.

[13] N. Jun and J. Tani. A model for learning to segment temporal sequences, utilizing a mixture of RNN experts together with adaptive variance. *Neural Netw.*, 21:1466–1475, December 2008.

[14] N.D. Le, R.D. Martin, and A.E. Raftery. Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models. *Journal of the American Statistical Association*, 91(436):1504–1515, 1996.

[15] D.G. Manolakis, V.K. Ingle, and S.M. Kogon. *Statistical and adaptive signal processing : spectral estimation, signal modeling, adaptive filtering, and array processing*. Boston : McGraw-Hill, 2000.

[16] G.J. McLachlan and D. Peel. *Finite mixture models*. New York : Wiley, 2000.

[17] S.P. Meyn and R.L. Tweedie. *Markov chains and stochastic stability*. New York : Springer-Verlag, 1993.

[18] P.M. Newbold and Y.C. Ho. Detection of changes in the characteristics of a Gauss-Markov process. *Aerospace and Electronic Systems, IEEE Transactions on*, AES-4(5):707–718, 1968.

[19] R. Patton, P. Frank, and R. Clark, editors. *Fault Diagnosis in Dynamic Systems*. Prentice Hall International, 1989.

[20] R. Patton, P. Frank, and R. Clark, editors. *Issues of fault diagnosis for dynamic systems*. London : New York: Springer, 2000.

[21] A.E. Raftery. A model for high-order Markov chains. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(3):528–539, 1985.

[22] I. Semoushin, J. Tsyganova, and M. Kulikova. Fault point detection with the bank of competitive Kalman filters. In *ICCS'03: Proceedings of the 2003 International Conference on Computational Science*, pages 417–426, Berlin, Heidelberg, 2003. Springer-Verlag.

[23] S. Simani, C. Fantuzzi, and R.J. Patton. *Model-based fault diagnosis in dynamic systems using identification techniques*. London ; New York : Springer, 2003.

[24] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical analysis of finite mixture distribution*. Chichester ; New York ; Brisbane : Wiley, 1985.

[25] H. Tong. *Non-linear time series : a dynamical system approach*. New York : Oxford University Press, 1990.

[26] C.S. Wong and W.K. Li. On a mixture autoregressive model. *Journal of the Royal Statistical Society Series B*, 62(1):95–115, 2000.