Hybrid SVM-HMM Diagnosis Method for Rotor-Gear-Bearing Transmission System

Qiang Shao
Department of Mechanical Engineering, University of Dalian Nationalities
Dalian Liaoning Province, China
Email: sq@dlnu.edu.cn

Changjian Feng
Department of Mechanical Engineering, University of Dalian Nationalities
Dalian, Liaoning Province, China
Email: fcj@dlnu.edu.cn

***Abstract*—No stationary time series are occurring when the plant proceeds to an abnormal state or a transient situation from a normal state. So it is necessary to identify the type of fault during its early stages for the selection of appropriate operator actions to prevent a more severe situation. This paper proposes a new architecture for identification of the time series. It converts the output of support vector machine (SVM) into the form of posterior probability which is computed by the combined use of sigmoid function and Gauss model, it acts as a probability evaluator in the hidden states of hidden Markov models (HMM). Experiments show that the architecture is very effective.**

***Index Terms*—HMM; SVM; identification; pattern recognition**

I. INTRODUCTION

In modern and unmanned machining systems, including dedicated transfer lines, flexible manufacturing systems, and Reconfigurable Manufacturing Systems (RMS), one crucial component is a reliable and effective monitoring system to monitor process conditions, and to take remedial action when failure occurs, or is imminent. Vibration monitoring method is adopted because it's cheapness and convenience. Hoverer the monitoring vibration signals are usually some nonstationary time series. Detection and identification these time series are belong to the problem of dynamic pattern.

Time often plays a secondary role: it should be incorporated in the feature extraction procedure. For practical recognition tasks, the assumption of stationarity of the class distributions may not be hold. Alternatively, information in sequences of feature vectors may be used for recognition. We will call both groups of problems dynamic pattern recognition problems. A dynamic pattern is a multidimensional pattern that evolves as a function of time.

A set of feature vectors can be looked upon as the result of independent draws from a multi-dimensional distribution. All temporal information should now be present in each feature vector. Identification problem may then be based on the dissimilarity of a set of newly measured feature vectors with respect to a set of known templates.

HMMs have been proved to be one of the most widely used tools for learning probabilistic models of dynamical time series[1-3]. HMM can model dynamical behaviors variation existing in the system through a latent variable (hidden states). HMM is good at dealing with sequential inputs, while SVM shows superior performance in classification. Furthermore, the former approach usually provides an intra-class measure while the latter proposes inter-class difference. Since these two classifiers use different criteria, they can be combined to yield an ideal one. The output of SVM is converted into the form of posterior probability which is computed by the combined use of sigmoid function and Gauss model, it acts as a probability evaluator in the hidden states of HMM. This paper introduces the SVM-HMM method to identification of nonstationary time series of rolling element bearing. The results show the proposed method is effective.

II. SVM AND ITS OUTPUT PROBABILITY

A. SVM for Classification

Support vector machine (SVM) has been widely used in the pattern recognition and regression due to its computational efficiency and good generalization performance. It was originated from the idea of the structural risk minimization that was developed by Vapik in 1970's [4]. In order to introduce to HMM, we should convert real-valued output of SVM to probability form.

The power of SVMs lies in their ability to transform data to a high dimensional space where the data can be separated using a linear hyperplane. The optimization process for SVM learning therefore begins with the definition of a functional that needs to be optimized in terms of the parameters of a hyperplane. The functional is defined such that it guarantees good classification[5-8]. On the training data and also maximizes the margin. The points that lie on the hyperplane satisfy,

$$\mathbf{w} \bullet \boldsymbol{x} + b = 0 \tag{1}$$

doi:10.4304/jcp.7.9.2324-2329

where W is the normal to the hyperplane and b is the bias of the hyperplane from the origin.

Given a labelled training data set,

$$\{\boldsymbol{x}_i, y_i\}_{i=1}^{N} \ (\boldsymbol{x}_i \in R^d \text{, and } y_i \in \{\pm 1\})$$

Where x_i is the input vector and y_i is its class label), an SVM constructs the discriminant function of classification as following:

$$f(\boldsymbol{x}) = sign(\mathbf{w} \bullet \boldsymbol{x} + b). \tag{2}$$

In order to maximize the separating margin, optimal problem can be solved. Minimize the following:

$$\phi(\mathbf{w}, \xi) = \frac{1}{2}(\mathbf{w} \bullet \mathbf{w}) + C\sum_{i=1}^{n} \xi_i \tag{3}$$

Subject to the constraints:

$$\begin{gathered} y_i((\mathbf{w} \bullet \boldsymbol{x}_i) + b) \geq 1 - \xi_i, i = 1, 2, \cdots, n \\ \xi_i \geq 0, i = 1, 2, \cdots, n \end{gathered}. \tag{4}$$

The result of the above problem is solving the maximum of the following function:

$$W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{5}$$

Subject to

$$\begin{gathered} 0 \leq \alpha_i \leq C, \quad i = 1, 2, \cdots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{gathered} \tag{6}$$

where $k(x_i,x_j)$ is kernel function.

The result discriminant function is

$$f(\boldsymbol{x}) = sign[\sum_{i=1}^{n} \alpha_i y_i k(\boldsymbol{x}, \boldsymbol{x}_i) + b]. \tag{7}$$

B. SVM's Output Probability[4,9]

General the outputs of SVM are symbols that represent the class labels. But real value outputs are considered only and converted to output probability. The real value is given by formula (2) and the value is

$$g(\boldsymbol{x}) = \mathbf{w} \bullet \boldsymbol{x} + b \tag{8}$$

Paying attention to the training samples are all normalized, then the closest points (support vectors) to the hyperplane are subject to | g(x) | =1.The points on the hyperplane are subject to | g(x) | =0, for others points, then

$$g(\boldsymbol{x}) = \pm d|\mathbf{w}| \tag{9}$$

where d is the distance between the vector x and the hyperplane. Positive and negative sign denote that the samples on the two sides of the hyperplane. Then for any sample point vector x, the formula is

$$d_x = \frac{g(\boldsymbol{x})}{|\mathbf{w}|} \tag{10}$$

To the support vectors, the formula is

$$d_{sv} = \frac{1}{|\mathbf{w}|} \tag{11}$$

Obviously g(x) is ratio of d_x and d_{sv}. Therefore we can get the output probability of SVM by Sigmoid function as followings,

$$P(C_{+1}|\boldsymbol{x}) = \frac{1}{1 + e^{-g(x)}}, \tag{12}$$

and

$$P(C_{-1}|\boldsymbol{x}) = \frac{1}{1 + e^{g(x)}}. \tag{13}$$

C. Output Probabilities of Muti-class SVM[4,9]

Binary SVM is discussed only from above. For muti-class problem we can transform the muti-calss SVM into a series of binary subtasks that can be trained by the binary SVM. In this paper one-against-one (OAO) decomposition strategy is adopted. The output probability of each binary SVM can be calculated by the method described in equation (12) and (13). Construct the feature vector as following,

$$V(\boldsymbol{x}) = [P_{ia1}(C_i|\boldsymbol{x}), \cdots, P_{iaj}(C_i|\boldsymbol{x}), P_{iaM}(C_i|\boldsymbol{x})]^T \tag{14}$$

where Piaj(Ci|x) denotes the output probability of the binary SVM determined by the i-th type and the j-th type training samples (i ≠ j).The feature vector can be transformed into output probability by Gauss model as following[9],

$$\begin{aligned} P(C_i \mid \boldsymbol{x}) &= N(V(\boldsymbol{x}), \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= 2\pi^{-\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} \\ &\quad \exp[-\frac{1}{2}(V(\boldsymbol{x}) - \boldsymbol{\mu})^T \Sigma (V(\boldsymbol{x}) - \boldsymbol{\mu})] \end{aligned} \tag{15}$$

III. Design Of Identification Of Pattern

A. HMM for Identification Problem

Identification problem of time series is defined the classification of signal type ω_j, given sequential input pattern X_t at time t. Input pattern X_t is mathematically defined as an object described by a sequence of features at time t.[10-12]

$$\boldsymbol{X}_t = (x_1, x_2, \cdots, x_d) \tag{16}$$

The space of input pattern X_t consists of the set of all possible pattern: $X_t \subset$Rd, Rd is a d-dimensional real vector space.

The k observed data up to time t is defined as,

$$\boldsymbol{\Phi}_{t-k} = \{\boldsymbol{X}_{t-k+1}, \cdots, \boldsymbol{X}_{t-1}, \boldsymbol{X}_t\} \tag{17}$$

The set of possible signal classes ωj forms the space of classes Ω:

$$\boldsymbol{\Omega}(t) = \{\omega_1, \omega_2, \cdots, \omega_c\}, \tag{18}$$

where c is the number of classes.

The identification task can be considered to be the finding of function f, which maps the space of input patterns $\boldsymbol{\Phi}_{t-k}$ to the space of classes $\boldsymbol{\Omega}$.

Nonstationary time series often exhibit sequentially changing behaviours. If one short-time period is defined

to a frame, the probability of a particular frame transition is different for each type of time series. Therefore, the probability of frame's existence, and of a particular transition between frames, can be statistically modelled. The probability of specific signal is already known, and is called prior probability. When identifying a specific time series, a decision can be made only by selecting the type of signal ω with the highest a prior probability $P(\omega)$. The decision is probably unreasonable. It is more reasonable to determine the type of time series after observing the trend of time series major variables, namely, to get the conditional probability $P(\omega|\mathbf{\Phi}_{t-k})$. This conditional probability is called a posterior probability. Decision-making based on the posterior probability is more reliable, because it employs a priori knowledge together with the observed time-series data. Classification of an unknown pattern X_t corresponds to finding the optimal model $\hat{\omega}$ that maximizes the conditional probability $P(\omega|\mathbf{\Phi}_{t-k})$ over the whole time series of the type ω. One can apply Bayes rule to calculate the a posterior probability,

$$P(\hat{\omega}|\mathbf{\Phi}_{t-k}) = \max_{\omega} \frac{P(\mathbf{\Phi}_{t-k}|\omega)P(\omega)}{P(\mathbf{\Phi}_{t-k})} \tag{19}$$

The conditional probability $P(\omega|\mathbf{\Phi}_{t-k})$ comes from comparing the shapes of the time series models with the input observations, while the a priori probability P(ω) comes from the accident probability. Since $P(\mathbf{\Phi}_{t-k})$ is independent of $\hat{\omega}$,

$$P(\hat{\omega}/\mathbf{\Phi}_{t-k}) \propto \max_{\omega}\{P(\mathbf{\Phi}_{t-k}|\omega)P(\omega)\} \tag{20}$$

In fact it is difficult to calculate an a priori probability *P(ω)*, which satisfy the following equation.

$$\sum_{j=1}^{c} \omega_j = 1 \tag{21}$$

The HMM can successfully treat an identification of nonstationary time series under a probabilistic or statistically framework.

In this identification problem, the HMM is used to estimate the conditional probability $P(\omega|\mathbf{\Phi}_{t-k})$.

B. Vibration feature extraction

Linear predictors are used to predict the value of the next sample of a signal as a linear combination of the previous samples. The next sample of the signal $\overline{s}_n$ is predicted as the weighted sum of the p previous samples, $s_{n-1}, s_{n-2}, \cdots, s_{n-p}$, $\overline{s}_n$ can be expressed as

$$\begin{aligned}\overline{s}_n &= a_1 s_{n-1} + a_2 s_{n-2} + \cdots + a_p s_{n-p} \\ &= \sum_{i=1}^{p} a_i s_{n-i}\end{aligned} \tag{22}$$

The residual error e_n is defined as the different between the actual and predicted values of the next sample and can be expressed as

$$e_n = s_n - \overline{s}_n = s_n - \sum_{i=1}^{n} a_i s_{n-i} \tag{23}$$

The weighting coefficients, also referred to as the linear prediction coefficients (LPC) a_1, a_2, $\cdots$, a_p, can be calculated by minimizing some functional of the residual signal e_n over each analysis window. Different methods can be used to find the linear prediction coefficients. The coefficients of linear predictors are equal the coefficients of AR models [3,9].

Vibration signals are nonstationary. Therefore, the future behaviour of a vibration signal is unpredictable. However, when the signal is divided into several small windows, quasistationary behaviour can be observed in each window. Thus, future behaviour of the vibration signal can be predicted separately in small windows under the restriction that a different model is used for each window.

In this approach, as illustrated in Fig.1, the signal is divided into windows of equal length. Each window is coded into a feature vector, which consists of a set of linear prediction coefficients for that window. The feature vectors for all windows are combined together to form a feature matrix. We will interchangeably use observation matrix and feature matrix throughout the rest of the paper. In this way, the vibration signal is a feature or observation matrix, which will then be used for training HMMs.

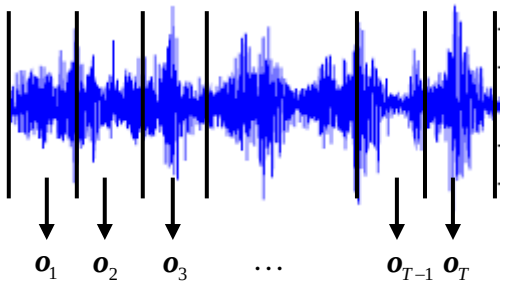

Fig.1 Vibration feature extraction

The observation matrix is, $\boldsymbol{O} = [\boldsymbol{o}_1 | \boldsymbol{o}_2 | \boldsymbol{o}_3 \cdots \boldsymbol{o}_{T-1} | \boldsymbol{o}_T]$, where the $\boldsymbol{o}_i$ is the vector of linear prediction coefficients for *i*-th window signal. It is equivalent to $\boldsymbol{X}_t$ which is described by the equation (16).

C. Application of HMM

By using the HMM, the pattern variability in the parameter space and time can be modelled effectively. HMM uses a Markov chain to model the changing statistical characteristics that exist in the actual observations of dynamic process signals. The Markov process is therefore a 'double' stochastic procedure that enables the modelling of not only spatial phenomena, but also time-scale distances. HMM's parameters are estimated from the Baum-Welch algorithm, An HMM is trained for each specific time series from both a set of training data, and an iterative maximum-likelihood estimation of model parameters from observed time-series data. Incoming observations are classified by calculating which model has the highest probability of producing such an observation.

The following parameters are needed to define a HMM:

The number of state, N

The transition probability distribution, $\boldsymbol{A} = \{a_{ij}\}$,where,

$$a_{ij} = P\{q_{t+1} = j \mid q_t = i\} 1 \le i, j \le N \tag{24}$$

The note q_t denotes the current state i.e., the probability of being in state j at time t+1 provided that the state at time t is i.

Observation probability distribution of each state, $\boldsymbol{B} = \{b_j(k)\}$,where,

$$b_j(k) = P(\boldsymbol{o}_k \mid q_t = j),\ 1 \le j \le N, 1 \le k \le M \tag{25}$$

where o_k and M denote the k-th observation and number of distinct observations, respectively. If the observation is modeled as continuous, a continuous probability density function must be specified for each state.

The initial state distribution, $\boldsymbol{\pi} = \{\pi_i\}$,where,

$$\pi_i = P(q_1 = i), 1 \le i \le N \tag{26}$$

It denotes the probability of the i-th state being the initial state.

The compact notation $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \pi)$ is used to represent a HMM.

Learning a HMM consists of two steps: (i) inference step where the posterior distribution over hidden states is calculated; (ii) learning step where parameters (such as initial state probability, state transition probability, and emission probability) are identified. The well-known forward-backward recursion allows us to infer the posterior over hidden states efficiently, More details on HMM can be found in [13].

D. Identification Method by HMM

Assume possible S types nonstationary time series are existed, each type of time series is modelled by a distinct HMM. Each HMM is trained by each training set constitutes an observation matrix $\boldsymbol{O}$. The following steps are involved in design of identification.

Step 1: The fist step is build an HMM λ_s for each type signal. In other words, we must estimate the model parameters $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \pi)$ that optimize the likelihood of the training set observation sequence or maximize $P(\boldsymbol{O} \mid \lambda_s)$, the probability of observation sequence O given model λ_s. The method is the Baulm-Welch re-estimation algorithm, also known as expectation maximization (EM) approach.

Step 2: Given unknown observation sequence, probabilities of all possible modes are calculated. The model with the highest likelihood is considered to be the best candidate for representing the specific time series. i.e.[3,9]

$$s^* = \arg\max_{1 \le s \le S} [P(\boldsymbol{O} \mid \lambda_s)] \tag{27}$$

IV. HYBRID SVM-HMM ARCHITECTURE

One significant drawback in SVMs is that, they are inherently static classifiers. they do not implicitly model temporal evolution of data. HMMs have the advantage of being able to handle dynamic data with certain assumptions about stationary and independence. Taking advantage of the relative strengths of these two classification paradigms we have developed a hybrid SVM-HMM architecture using our Baum-Welch training method. This system provided all components for the HMM portion of the hybrid system architecture. For estimating SVMs we used a publicly available toolkit, stprtoolbox[2]. The flow chart of SVM-HMM training is as Fig.2.

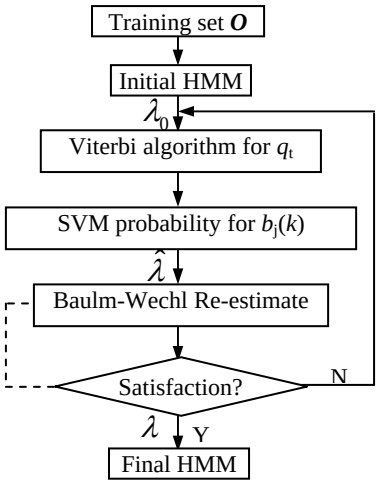

Fig.2 Hybrid SVM-HMM training

SVM is mainly used to calculate the observation probability as equation (25). Therefore it is necessary to assume the number of sub-state of the j-th hidden sate of HMM is k.

V. EXPERIMENTS AND RESULTS

A. Experiment Setup

The major objective of this paper was the experimental investigation of vibration signatures due to localized wear/damage in bearing outer race and gear tooth. Vibration results from three cases of a combination of bearing and gear.

1) The undamaged bearing and the gear set with no induced damage/wear.

2) The damage bearing and the gear set with no induced damage/wear.

3) The damage bearing and the gear set with one single tooth damage gear.

In order to perform a parametric study of the effects of bearing and gear damage on the vibration signatures of the system, vibration study for three different cases above were carried out using the test rig shown in Figure 3.

The test rig consists of two identical spur gears on two shafts with one attached to the electric motor driver while the other is attached to a water-braking system to provide loading onto the gears, each shaft is supported by two

bearings. The driver of the gear test rig consists of a 75Hp motor connected through a belt-pulley driving system that can 6 provide a maximum speed of up to 8000 rpm.

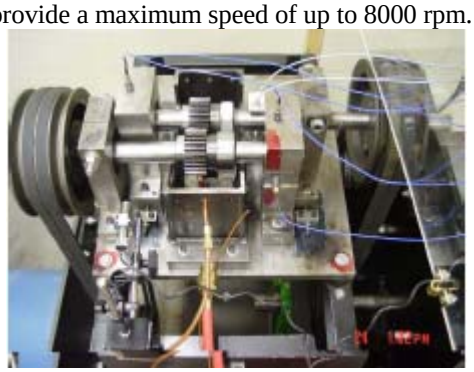

Fig. 3 Rotor-gear-bearing test rig

Using a shaft speed 20 Hz (1200 rpm), both rotor speed and bearing carrier speed were also measured using optical encoders. Vibration data were acquired through a set of accelerometers (one accelerometer in x-direction and one accelerometer in y-direction on the bearing box) to a computer-based high-speed analog-to-digital system. The sampling rate of the vibration data was set to be 6000 Hz. There were around 300 samples per revolution of the rotor (32768 in total). Vibration signals for approximately 109 revolutions were acquired to be stored in computer for fault identification vibration signature analysis.

Figure 4 shows the test gear with single tooth damage. Figure 5 shows the test bearing with the damage of outer race.

Fig. 4 Gear with single tooth damage

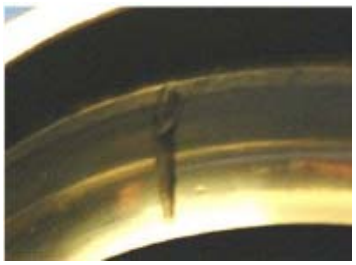

Fig. 5 Bearing with outer race damage

In the Time Signal figures, only 1600 points are chosen within 32768 total points. Time signals with undamaged the gear and undamaged bearing of y-direction are shown in Figure 6. Time signals with undamaged gear and damaged bearing of y-direction are shown in figure 7. Time signals with damage gear and damage bearing of y-direction are shown in figure 8. Only y-direction vibration signals are considered.

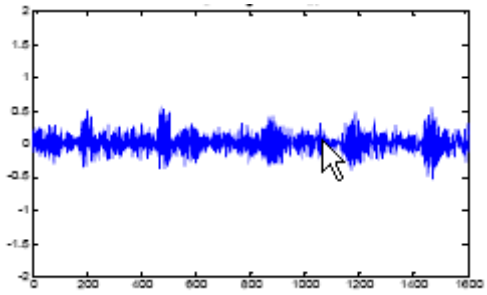

Fig. 6 Time signals for undamaged gear & undamaged bearing

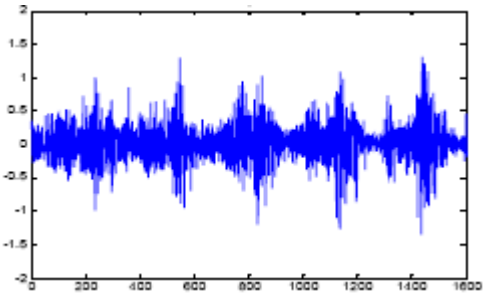

Fig. 7 Time signals for undamaged gear & damaged bearing

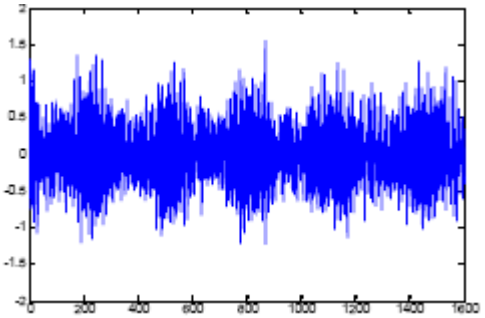

Fig. 8 Time signals for damaged gear & damaged bearing

B. Features Extraction by Wavelet Packet Decomposition

The features used for a specific fault must only be correlated with this specific fault and uncorrelated (or unnoticeable corrected) with all the other faults. To illustrate this, if the energy of a certain frequency band is used as one of the features for type-one fault, then the energy of this band must only be affected by the presence of type-one fault and be unchanged (or minimally affected) by the presence of other faults. Thus a new feature extraction method is considered. The signal is divided into several small window segments of equal length. Wavelet packet decomposition (WPD) is applied to the segments of the signal in each window. A detailed review of WPD can be found in [7]. The feature vector for each window consists of selected node energies of the WPD. This process of WPD is illustrated .

Figure 1 shows the three wavelet packet decomposition. (i,j) represents the i-th layer and j-th node. For example, (3, 0) represents the third layer and 0 th node of WPD. We obtain the coefficients of all the nodes. The time signals S_{ij} are reconstructed from the coefficients. All the nodes in third layer are considered. Then the whole signal is obtained as following,

$$S = S_{30} + S_{31} + \cdots + S_{37} \tag{28}$$

Thus the energy of node is as following,

$$E_{3j} = \int \left| S_{3j} \right|^2 dt = \sum_{k=1}^{n} \left| x_{jk} \right|^2 \tag{29}$$

Where x_{jk} represent amplitude of reconstructed discrete signal. Therefore the observation feature vector is constructed as following.

$$\boldsymbol{o}_t = [E_{30}, E_{31}, \cdots, E_{37}] \tag{30}$$

C Experiment Results

Initial HMM is as figure 2. The number of components of gauss mixture is 5. The result of 20 times test of single fault diagnosis is illustrated in table 1.

TABLE I
THE RESULT OF DIAGNOSIS

test (20times)	Normal	Fault-1	Fault-2
Normal	20	0	0
Fault-1	1	19	0
Fault-2	1	1	18

Normal, Fault-1 and Fault-2 represent the three cases in section IV(A) separately.

The method of the multiple fault diagnosis is not test in this work.

V. CONCLUSIONS

In this paper, we attempted to construct identification classifier for nonstationary time series by integrating SVMs to HMMs. The proposed hybrid method utilized the advantages of both HMM as a detector for time varying characteristics and SVM as a powerful binary classifier. The results on the signals of the rotor-gear-bearing show that hybrid model is executable and effective..

ACKNOWLEDGMENT

This work was financially supported by Liaoning Province education department (L2010092) and University of Dalian Nationalities talent import fund(20016202).

REFERENCES

[1] Rabiner L R and Juang B H. “An Introduction to Hidden Markov Models”. *IEEE ASSP Magazine*, vol.3, pp. 4-6, January,1986.

[2] Vojtˇech Franc and V´aclav Hlav´aˇ,“Statistical Pattern Recognition Toolbox”, *http://cmp.felk.cvut.cz*, June, 2004.

[3] Hasan Ocak and Kenneth A.Loparo. “HMM-Based Fault detection and Diagnosis Scheme for Rolling Element Bearings”, *Journal of Vibration and* Acoustics, vol.127, pp. 299-306, 2005.

[4] J.Platt. “Probabilistic output for Support Vector Machine and Comparisons to regularized likelihood method in advances in large margin classifiers”, MIT Press, Cambridge, MA,USA,1999: 61-73.

[5] Ramy Saad Saman K. Halgamuge Jason Li, Polynomial kernel adaptation and extensions to the SVM classifier learning, Neural Comput & Applic (2008) 17:19–25.

[6] Xavier Capron·D´esir´e Luc Massart, Johanna Smeyers-Verbeke, Multivariate authentication of the geographical origin of wines:a kernel SVM approach, Eur Food Res Technol (2007) 225:559–568.

[7] Fethi Smach, Cedric Lemaître, Jean-Paul Gauthier, Johel Miteran, Mohamed Atri, Generalized Fourier Descriptors with Applications to Objects Recognition in SVM Context, J Math Imaging Vis(2008)30:43–71.

[8] Cecilio Angulo,Francisco J. Ruiz,Luts Gonza Lez,and Juan Antonio Ortega, Multi-Classification by Using Tri-Class SVM, Neural Processing Letters (2006) 23:89-101.

[9] Hua Jing. “Research on continuous speech recognition based on a hybrid HMM/SVM framework”. Master thesis, Harbin Institute of Technology, 2006.

[10] Kee-Choon Kwona and Jin-Hyung Kim. “Accident identification in nuclear power plants using hidden Markov models”. *Engineering Applications of Artificial Intelligence*, vol.12, pp. 491-501, 1999.

[11] Atulya Velivellia, Thomas S. Huanga and Alexander Hauptmann, Video shot retrieval using a kernel derived from a continuous HMM, SPIE-IS&T/ Vol. 6073 607311:1-10.

[12] Xiao-Bing Li, Frank K. Soong, Tor André Myrvoll, Ren-Hua Wang, optimal clustering and non-uniform allocation of Gaussian kernels in scalar dimension for HMM compression, ICASSP 2005:669-672.

[13] Feng Chang-jian, Kang-jing, Wu-bin, Hu Hong-ying, “Application in Fault Diagnosis of Rotary Machine Based on Theory Of DHMM Dynamic Pattern Recognition”, Journal of Dalian Nationalities University, No.3, pp,12-15,May,2005.