Multiple Fault Detection and Isolation Using the Haar Transform, Part 1: Theory

C. K. H. Koh
School of Mechanical and
Production Engineering,
Nanyang Technological University,
Singapore

J. Shi
Industrial and Operations
Engineering Department

W. J. Williams
Electrical Engineering and Computer
Science Department

J. Ni
Mechanical Engineering and Applied
Mechanics Department

University of Michigan,
Ann Arbor, MI 48109

Most manufacturing processes involve several process variables which interact with one another to produce a resultant action on the part. A fault is said to occur when any of these process variables deviate beyond their specified limits. An alarm is triggered when this happens. Low cost and less sophisticated detection schemes based on threshold bounds on the original measurements (without feature extraction) often suffer from high false alarm and missed detection rates when the process measurements are not properly conditioned. They are unable to detect frequency or phase shifted fault signals whose amplitudes remain within specifications. They also provide little or no information about the multiplicity (number of faults in the same process cycle) or location (the portion of the cycle where the fault was detected) of the fault condition. A method of overcoming these limitations is proposed in this paper. The Haar transform is used to generate sets of detection signals from the original measurements of process monitoring signals. By partitioning these signals into disjoint segments, mutually exclusive sets of Haar coefficients can be used to locate faults at different phases of the process. The lack of a priori information on fault condition is overcomed by using the Neyman-Pearson criteria for the uniformly most powerful form (UMP) of the likelihood ratio test (LRT).

1 Introduction

Many detection schemes based on thresholding of the measured signal have several limitations which result in high rates of missed detection and false alarms. A method is proposed in this paper to overcome some of these shortcomings which are common in manufacturing processes where: (i) fault detection is highly dependent on the experience and skill of the person and thus prone to human error, (ii) there is a high level of process noise from lack of process repeatability, (iii) fault signals are frequency or phase shifted with no change in amplitudes, (iv) *a priori* information on fault conditions are normally unavailable for setting thresholds, and (v) multiple faults are found within the same cycle.

The approach in this paper is based on the orthogonal Haar transform and has four major advantages: (i) it is not dependent on *a priori* knowledge of the statistical properties of the fault signals, (ii) it can localize the position of the fault, aiding in its identification, (iii) it can detect faults which do not affect the amplitude of the process signal, and (iv) it can detect multiple faults occurring within the same cycle. Unlike the more familiar orthogonal Fourier transform which uses complex exponentials as its basis functions and is ideal for sinusoidal narrow-band signals, the Haar transform is especially well-suited to represent spectrally wide-band signals [5]. Fault signatures which are frequency or phase shifted, and invisible to time domain control limits, can now be easily detected. The Haar transform is defined over the entire signal length, and its wavelet structure allows the detection of multiple faults occuring within the same press cycle. The Haar transform is also one of the most efficient transformation algorithms in terms of computational speed and memory usage [2], [7], and thus ideal for implementation as an on-line monitoring system.

This paper is the first of a two part series, and is organized as follows. Section 2 introduces the Haar transform and derives the mapping between the time and Haar domains. The development of the global detector is described in section 3 and the summary and conclusion are found in section 4.

2 The Haar Transform

2.1 The Continuous Haar Transform. The Haar transform is a member of a class of nonsinusoidal orthogonal functions [1]. It consists of rectangular waves distinguished by time scalings and time shifts. The set of continuous Haar functions $\{h(r, m, t)\}$ is periodic, orthonormal and complete, and was proposed by Alfred Haar in 1910. The Haar orthonormal sequence is defined on the open interval $[0, 1)$ and can be generated by the recurring relation (1):

$$h(0, 0, t) = 1, \quad t \in [0, 1)$$

$$h(r, m, t) = \begin{cases} 2^{r/2}, & \dfrac{m-1}{2^r} \le t < \dfrac{m-1/2}{2^r} \\ -2^{r/2}, & \dfrac{m-1/2}{2^r} \le t < \dfrac{m}{2^r} \\ 0, & \text{elsewhere } \forall t \in [0, 1) \end{cases} \qquad (1)$$

where $r \in 0 \le r < \log_2 n$ is the scale, n is the number of subintervals in $[0, 1)$ and $m \in 1 \le m \le 2^r$ is the number of functions with scale r.

The first eight continuous Haar functions are shown in Fig. 1. Points of discontinuity are defined as the average of the limits approached from both sides of the discontinuity.

2.2 The Discrete Haar Transform. The corresponding discrete Haar functions are obtained by sampling the continuous Haar functions in Fig. 1 at the middle of each subinterval to produce an array:

Contributed by the Manufacturing Engineering Division for publication in the JOURNAL OF MANUFACTURING SCIENCE AND ENGINEERING. Manuscript received Oct. 1996; revised April 1998. Associate Technical Editor: M. Elbestawi.

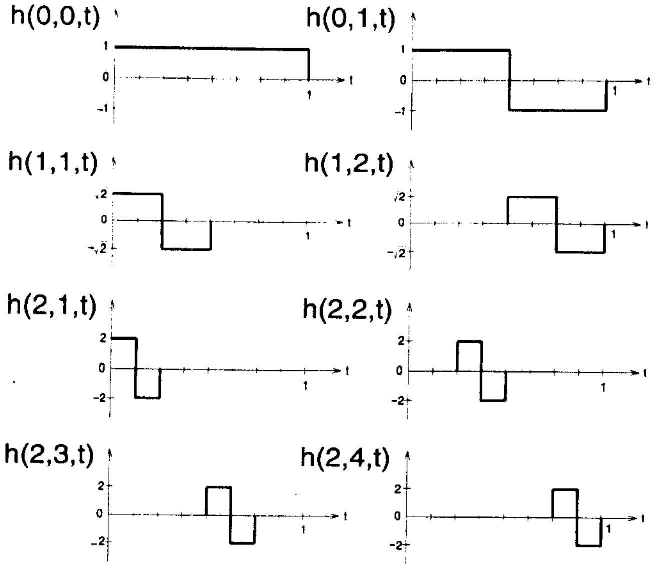

Fig. 1 Continuous Haar functions for $n = 8$ subintervals

$$H_3^T = \frac{1}{\sqrt{8}} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ \sqrt{2} & \sqrt{2} & -\sqrt{2} & -\sqrt{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{2} & \sqrt{2} & -\sqrt{2} & -\sqrt{2} \\ 2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & -2 \end{bmatrix} = \begin{bmatrix} \underline{h}_1^T \\ \underline{h}_2^T \\ \underline{h}_3^T \\ \underline{h}_4^T \\ \underline{h}_5^T \\ \underline{h}_6^T \\ \underline{h}_7^T \\ \underline{h}_8^T \end{bmatrix} \qquad (2)$$

The discrete Haar transform array is denoted by H_N where $n = 2^N$ is the number of discrete data points. Each row of H_N^T is a discrete Haar function, $\underline{h}_i$, obtained by sampling the corresponding continuous Haar function, $h(r, m, t)$. The matrix H_N^T is orthonormal by definition.

If $\underline{X} = [x_1 \;\; x_2 \;\; x_3 \;\; \ldots \;\; x_n]^T$ is a vector sequence of n discrete points, and the corresponding set of Haar coefficients is $\underline{C} = [c_1 \;\; c_2 \;\; c_3 \;\; \ldots \;\; c_n]^T$ for the Haar basis functions $\underline{h}_i$, $i = 1, 2, \ldots n$, then $\underline{X}$ and $\underline{C}$ are related by the transform pair:

$$\underline{X} = H_N \underline{C} = \sum_{i=1}^{n} c_i \underline{h}_i \qquad (3)$$

and

$$\underline{C} = H_N^T \underline{X} \qquad (4)$$

2.3 Properties of the Haar Transform. The Haar transform has several nice properties which make it particularly attractive over other orthogonal transforms for this kind of detection problem.

1. **Computational Efficiency.** It can be performed in $(2n - 2)$ additions and subtractions and is considerably less than the $n \log_2 n$ multiplications for the fast Fourier transform. Fast algorithms for computing the Haar transform can be found in [2], [7], [8].

2. **Signal Bandwidth.** While the more familiar orthogonal Fourier transform is ideal for sinusoidal narrow-band signals, the Haar transform is especially well-suited to represent spectrally wide-band signals [5] with step-like discontinuities.

3. **Data Reduction.** The uncertainty of process repeatability makes it undesirable to monitor the process at a particular fixed instant. The lower indexed Haar functions have large regions of support and are proportional to the signal averages over these regions. Consequently, it is sufficient to monitor selected Haar coefficients rather than a long vector of consecutive time domain data. Also, most manufacturing systems are large inertia systems which act as low pass filters. The frequency content of the process signals are therefore concentrated in the lower harmonic range. Hence, higher-indexed Haar coefficients represent mostly noise and may be neglected resulting in a further reduction of data.

4. **Measure of System Entropy.** The complexity of a signal with respect to a given discrete basis can be defined as the Shannon entropy of the basis expansion coefficients [6]. The Haar coefficients can be used to estimate the entropy of the system.

5. **Isolation of Fault.** Unlike the other orthogonal transforms such as the Fourier, Rademacher, Walsh-Hadamad and the Discrete Cosine transforms, the Haar functions possess a wavelet structure which enables it to isolate localized events in the time domain.

2.4 Mapping Between the Time and Haar Domains. The mapping functions between the Haar and time domains determine the regions of support for the Haar functions and consequently the Haar coefficients to represent different segments of the press cycle [3].

If $\underline{X} = [x_1 \;\; x_2 \;\; \ldots \;\; x_n]$ is a data vector of length $n = 2^N$ points, and the discrete Haar function $\underline{h}_i$ in (1) is indexed by $i = 2^r + m$, the region of support $k_1 \leq k \leq k_2$ is given by

$$(m - 1)2^{N-r} \leq k < 2^{N-r}m \qquad (5)$$

where $k_1 = (m - 1)2^{N-r}$ and $k_2 = 2^{N-r}m - 1$.

Conversely, the Haar functions whose regions of support overlap a given data segment can be determined as follows. If the data falls completely between one of the binary boundaries, then i can be determined from

$$r = N - \left\lceil \frac{\log(k_2 - k_1 + 1)}{\log(2)} \right\rceil \qquad (6)$$

and

$$m = (k_2 + 1)/2^{N-r} \qquad (7)$$

In general, the data segment of interest, $S = \{x_j\}_{j=p}^{j=q} \in \underline{X}$, do not always lie between any convenient binary intervals. In this case, the discrete Haar functions $\underline{h}_i$ which completely or partially span S must satisfy one of the following conditions

$$\{k_1 < p\} \cap \{p \leq k_2 \leq q\} \qquad (8)$$

$$\{p \leq k_1 \leq q\} \cap \{k_2 > q\} \qquad (9)$$

$$\{k_1 \geq p\} \cap \{q \geq k_2\} \qquad (10)$$

$$\{k_1 \leq p\} \cap \{q \leq k_2\} \qquad (11)$$

where $k_1 \leq k \leq k_2$ is the region of support for $\underline{h}_i$.

The process monitoring signal can be divided into several disjoint segments. Condition (10) ensures that the sets of Haar coefficients for disjoint data segments are also mutually exclusive.

2.5 Moments of the Haar Coefficient. A Haar coefficient is a linear combination of consecutive points in the original data. Its statistical properties are therefore dependent on the behavior of these points. From the analysis in section 2.4, $\underline{h}_i$ can be expressed as

$$\underline{h}_i = [h_{i1} \quad h_{i2} \quad \ldots \quad h_{in}] \tag{12}$$

where $i = 2^r + m$ and h_{ij} is the jth element of $\underline{h}_i$. The Haar coefficient, c_i, which is the projection of the discrete signal $\underline{X}$ on $\underline{h}_i$, is given by

$$c_i = \langle \underline{h}_i, \quad \underline{X} \rangle = \sum_{j=1}^{n} h_{ij} x_j \tag{13}$$

The expected value and variance of c_i are thus

$$E[c_i] = E[\sum_{j=1}^{n} h_{ij} x_j] = \sum_{j=1}^{n} h_{ij} E[x_j] \tag{14}$$

and

$$\text{var}\ [c_i = \text{var}\ [\sum_{j=1}^{n} h_{ij} x_j]$$

$$= \sum_{j=1}^{n} (h_{ij})^2 \sigma_j^2 + \sum_{j=1}^{n} \sum_{\substack{k=1 \\ j \neq k}}^{n} h_{ij} h_{ik} \text{ cov } (x_j, x_k) \tag{15}$$

Equations (14) and (15) determine the control (or threshold) limits on the respective Haar coefficients. Since c_i is a linear combination of several random variables, $x_j, j = 1, \ldots n$, with finite mean and finite variances, the Central Limit Theorem implies a normal Gaussian distribution for c_i, $i = 1, \ldots n$, if n is sufficiently large. However, if we assume that x_j is Gaussian for all j, then c_i is also Gaussian for all i without the requirement of n being sufficiently large.

2.6 Sensitivity Analysis. The number of Haar coefficients to be monitored can be further reduced by selecting only the Haar coefficients with high sensitivity to process changes but low sensitivity to process noise. Two sensitivity indices are therefore required. The first index, ξ_{ij} is a measure of the sensitivity of coefficient c_i to fault F_j. Assuming homogeneity of variance [3], ξ_{ij} can be defined as

$$\xi_{ij} = \frac{E[c_i | H_j] - E[c_i | H_o]}{\sigma_{c_i}} \tag{16}$$

where H_o and H_j are the respective hypotheses for no-fault and fault F_j condition, and σ_{c_i} is the standard deviation of the coefficient c_i (15) under H_o.

If $f_o(t)$ and $f_j(t)$ are the continuous process signals under no-fault and fault conditions, from the definition of c_i

$$E[c_i | H_o] = \int E[f_o(t)] h(r, m, t) dt$$

and

$$E[c_i | H_j] - E[c_i | H_o] = \int E[f_j(t) - f_o(t)] h(r, m, t) dt$$

$$= \int E[e_j(t)] h(r, m, t) dt$$

For a sufficiently large value of scale $\hat{r}$, the support region for $h(\hat{r}, m, t)$ is very small compared to the interval [0, 1). If $f(t) = s(t) + n(t)$, the signal of interest $s(t)$ will appear to be slowly changing over $h(\hat{r}, m, t)$ and the corresponding coefficient will represent mostly the process noise $n(t)$. A second index, ζ_i, can be used to measure the sensitivity of the coefficient c_i to the process noise. This index is simply the standard deviation of the coefficient under normal condition, and is a function of the process noise over the region of support of c_i. Thus

$$\zeta_i = \sigma_{c_i} \tag{17}$$

For a given region of support, $0 \leq p < q < 1$, a subset of $\{h(r, m, t)\}$ exists which spans the region $[p : q]$. Of these, an optimal subset of Haar coefficients [3] can be found which maximizes (16) and minimizes (17).

3 Developing the Haar Detector

One of the main challenges in the detection of process faults is the lack of *a priori* information on most of the faults. Some faults never occur at all if the equipment has been properly used and maintained. And, when faults do occur, the machine is often shut down almost immediately by an overload device. As a result, samples of fault signatures are rare to come by for statistical analysis. In this approach, Haar coefficients, which change abruptly in the presence of process changes, are used as detection signals. The statistical properties of fault signals are therefore unnecessary.

The jump in the value of a Haar coefficient c_i can be modeled as a mean shift or a variance shift or both. Assuming that c_i has a normal Gaussian distribution under all conditions, the problem can be formulated by the following four hypotheses:

Case 1. No fault condition:

$$H_o : c \sim N(m_o, \sigma_o^2)$$

Case 2. Change in mean:

$$H_1 : c \sim N(m_1, \sigma_o^2)$$

Case 3. Change in variance:

$$H_2 : c \sim N(m_o, \sigma_1^2)$$

Case 4. Change in mean and variance:

$$H_3 : c \sim N(m_1, \sigma_1^2)$$

where c is the observation of c_i arriving at the detector. The mean values of c_i under (H_o, H_2) and (H_1, H_3) are denoted by m_o and m_1 respectively. The variance of c_i under (H_o, H_1) and (H_2, H_3) are denoted by σ_o^2 and σ_1^2 respectively.

Different algorithms can be derived for each of these hypotheses. In practice, it is easy to estimate the values of m_o and σ_o^2 using (14) and (15), but there is often no way of estimating m_1 and σ_1^2 with confidence. It will be assumed that the process noise remains unchanged throughout so that $\sigma_1^2 = \sigma_o^2$. Hence, the fault condition can be modeled by hypothesis H_1, and the fault detection problem consequently reduces to distinguishing between hypotheses H_0 and H_1.

3.1 The Likelihood Ratio Test. The likelihood ratio test (LRT) for c_i can be expressed as

$$\Lambda(R) = \frac{\dfrac{1}{\sqrt{2\pi\sigma_1^2}} e^{-(R-m_1)^2/2\sigma_1^2}}{\dfrac{1}{\sqrt{2\pi\sigma_o^2}} e^{-(R-m_o)^2/2\sigma_o^2}} \underset{H_o}{\overset{H_1}{\gtrless}} \eta \tag{18}$$

where R is a particular value of c. Taking the natural logarithm of (18),

$$l(R) = \frac{1}{2\sigma_o^2}(R - m_o)^2 - \frac{1}{2\sigma_1^2}(R - m_1)^2 \underset{H_o}{\overset{H_1}{\gtrless}} \log(\eta) - \log(\sigma_o) + \log(\sigma_1) \qquad (19)$$

Assuming homogeneity of variance, $\sigma_o^2 = \sigma_1^2 = \sigma^2$ in (19), so that

$$l(R) = \frac{1}{2\sigma^2}(R - m_o)^2 - \frac{1}{2\sigma^2}(R - m_1)^2 \underset{H_o}{\overset{H_1}{\gtrless}} \log(\eta) \qquad (20)$$

or

$$R \underset{H_o}{\overset{H_1}{\gtrless}} \frac{2\sigma^2 \log(\eta) + (m_1^2 - m_o^2)}{2(m_1 - m_o)} = \gamma > 0 \quad \text{if } m_1 > m_o \qquad (21)$$

and so

$$R \underset{H_o}{\overset{H_1}{\lesseqgtr}} \frac{2\sigma^2 \log(\eta) + (m_1^2 - m_o^2)}{2(m_1 - m_o)} = \gamma < 0 \quad \text{if } m_1 < m_o \qquad (22)$$

where γ is a function of η, m_o and m_1. The jump in the value of R is the same as having a Gaussian density with a larger absolute mean.

It is clear from the structure of the test (21)–(22) that an ordinary LRT can be designed for a particular m_1. However, m_1 is not a predictable value and therefore cannot be part of the test. The uniformly most powerful (UMP) form of the LRT is used to circumvent this obstacle. A UMP test must be as good as any other test for every m_1 and exists if and only if the LRT for every m_1 can be completely defined (including threshold) without knowledge of m_1 [9]. The existence of the UMP in the sense of the Neyman-Pearson criterion is proven in the following section.

3.2 Proof of Existence of the UMP. Using the Neyman-Pearson criterion, the probability of false alarm is denoted by $P_F = \alpha$, where α is the prespecified level of significance. For the case where $m_1 > m_o$

$$\begin{aligned} P_F &= \Pr[\text{choose } H_1 | H_o \text{ true}] \\ &= \int_\gamma^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-(R-m_o)^2/2\sigma^2} dR \\ &= \text{erfc}\left[\frac{\gamma - m_o}{\sigma}\right] = \alpha \end{aligned} \qquad (23)$$

where erfc is the error function integral [4]. Therefore

$$\gamma = m_o + \sigma \, \text{erfc}^{-1}(\alpha) \qquad (24)$$

Let $\lambda = (\gamma - m_o)/\sigma$. A value of λ can be found for a given α [4]. Rewriting,

$$\gamma = m_o + \lambda\sigma \qquad (25)$$

and substituting into (21)

$$R - m_o \underset{H_o}{\overset{H_1}{\gtrless}} \lambda\sigma \qquad (26)$$

where $\lambda\sigma \geq 0$ always. The UMP test now reduces to the form of a mean change detector.

If $m_1 < m_o$, then

$$\begin{aligned} P_F &= \int_{-\infty}^{\gamma} \frac{1}{\sqrt{2\pi}\sigma} e^{-(R-m_o)^2/2\sigma^2} dR \\ &= \text{erf}\left[\frac{\gamma - m_o}{\sigma}\right] = \alpha \end{aligned} \qquad (27)$$

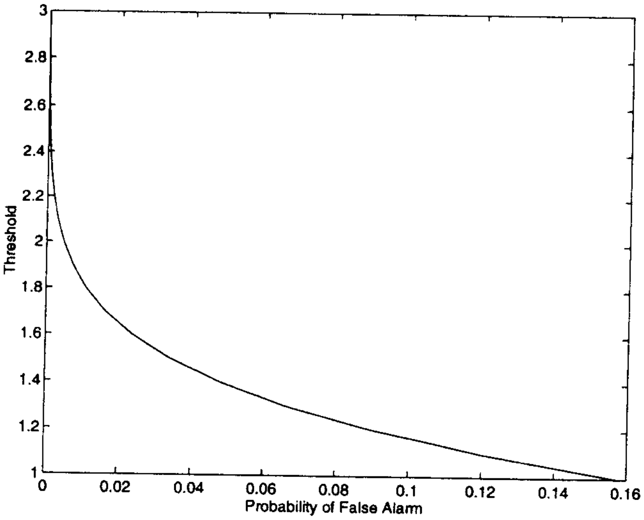

Fig. 2 Relationship between threshold (λ) and probability of false alarm ($P_F = \alpha$)

and

$$\gamma = m_o + \sigma \, \text{erf}^{-1}(\alpha) \qquad (28)$$

so that

$$R - m_o \underset{H_o}{\overset{H_1}{\lesseqgtr}} -\lambda\sigma \qquad (29)$$

From the sensitivity analysis of the Haar coefficients in section 2.6, R is known to be highly dependent on the shape of the signal. The same fault would therefore affect the Haar coefficient the same way it occurs each time. This satisfies the condition that for the UMP test to exist, R can take on only values greater than or equal to m_o (26), or R can take on values less than or equal to m_o (29), but not both at the same time for the same fault.

Taking advantage of the symmetry of the two directional tests, (26) and (29) can be combined as a test of the absolute difference between R and m_o, that is

$$|R - m_o| \underset{H_o}{\overset{H_1}{\gtrless}} \lambda\sigma \qquad (30)$$

The probability of detection, P_D is defined as

$$\begin{aligned} P_D &= \Pr[\text{choose } H_1 | H_1 \text{ true}] \\ &= \int_\gamma^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-(R-m_1)^2/2\sigma^2} dR \\ &= \text{erfc}\left[\frac{\gamma - m_1}{\sigma}\right] \end{aligned} \qquad (31)$$

Equation (31) gives an upper bound on the power of the test for any value of m_1.

The threshold γ (24) is independent of the value of m_1. The performance of the test, P_D varies with m_1 (31), but knowing m_1 will not make it better. This is the uniformly most powerful form of the LRT. The necessity of estimating the magnitude of the fault has been avoided by using the UMP test.

The relationship between the threshold (γ) and the probability of false alarm ($P_F = \alpha$) is shown in Fig. 2. The higher the threshold, the smaller the likelihood of false alarm and vice versa. Hence the two quantities are almost inversely proportional to each other. The receiver operating characteristic (ROC) curve for the UMP detector is plotted as a function of

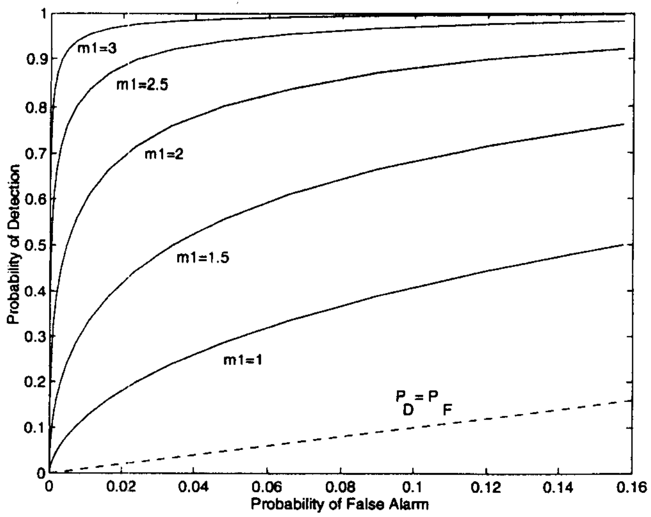

Fig. 3 ROC curve for UMP detector for different fault magnitude, m_1, with $m_o = \sigma^2 = 1$

α in Fig. 3 with values of $0 \leq \alpha \leq 1$, and $m_o = 0$, $\sigma^2 = 1$. The detector performance is clearly superior to that of a randomized test indicated by the $P_D = P_F$ line.

4 Conclusion

The complexity of a manufacturing process and the number of variables that affect the process pose a very difficult challenge for any kind of on-line fault detection system. A solution has been proposed which exploits the information in the monitoring signature using the orthogonal Haar transform. The resulting coefficients, if properly selected, can perform remarkably well with a UMP detector based on the Neyman-Pearson criteria. The Haar coefficients are ideal detection signals for a wide range of different faults. The location of these faults can be determined by partitioning the original data into disjoint segments; and associating each segment with an exclusive set of detection signals. The UMP algorithm takes care of fault signals whose statistical properties are not known in advance. The performance of the detector was investigated with respect to threshold and false alarm rate.

The authors are grateful to the referees for their helpful comments. The work is partially supported by the NSF Industrial/University Cooperative Research Center at the University of Michigan and Dr. J. Shi was partially supported by NSF CAREER award: DMI 9624402.

References

1 Ahmed, N., and Rao, K. R., *Orthogonal Transforms for Digital Signal Processing,* Springer-Verlag, Berlin 1975.

2 Fakruddin, D. B., and Parthasarathy, K., "Simplified Algorithms Based on Haar Transforms for Signal Recognition in Protective Relays," *Proceedings of the IEEE,* Vol. 73, No. 5, pp. 940–942, May 1985.

3 Koh, C. K-H, "Tonnage Signature Analysis for the Stamping Process," PhD Thesis, Univ. of Mich., Aug 1995.

4 Leon-Garcia, A., *Probability and Random Processes,* Addison Wesley, Table 3.3, pp. 126–127, 1989.

5 Mikhael, W. B., and Ramaswamy, A., "Residual Error Formulation and Adaptive Minimization for Representing Nonstationary Signals Using Mixed Transforms," *IEEE Trans. Circuits and Systems-II: Analog and Digital Signal Processing,* pp. 489–492, Vol. 39, Jul 1992.

6 Orr, R., "Dimensionality of Signal Sets," *Proc. SPIE vol. 1565, Adaptive Signal Processing,* pp. 435–446, 1991.

7 Roeser, P. R., and Jernigan, M. E., "Fast Haar Transform Algorithms," *IEEE Trans. Computers,* pp. 175–177, Feb 1982.

8 Shore, J. E., "On the Application of Haar Functions," *IEEE Trans. Communications,* pp. 209–216, Mar 1973.

9 Van Trees, H. L., *Detection, Estimation, and Modulation Theory,* New York, Wiley, 1968.