Detection of Hidden Markov Model Transient Signals

BIAO CHEN
Syracuse University

PETER WILLETT, Member, IEEE
University of Connecticut

Addressed here is the quickest detection of transient signals which can be represented as hidden Markov models (HMMs), with the application of detection of transient signals. Relying on the fact that Page's test is equivalent to a repeated sequential probability ratio test (SPRT), we are able to devise a procedure analogous to Page's test for dependent observations. By using the so-called forward variable of an HMM, such a procedure is applied to the detection of a change in hidden Markov modeled observations, i.e., a switch from one HMM to another. Performance indices of Page's test, the average run length (ARL) under both hypotheses, are approximated and confirmed via simulation. Several important examples are investigated in depth to illustrate the advantages of the proposed scheme.

Manuscript received July 12, 1999; revised February 26, 2000; released for publication March 16, 2000.

IEEE Log No. T-AES/36/4/11371.

Refereeing of this contribution was handled by J. A. Tague.

This research was supported by ONR through NUWC Newport under Contract N66604-97-M-3139.

Authors' addresses: B. Chen, Department of EECS, 121 Link Hall, Syracuse University, Syracuse, NY 13244, E-mail: (bichen@syr.edu); P. Willett, Department of Electrical and Systems Engineering, U-157, University of Connecticut, Storrs, CT 06269, E-mail: (willett@engr.uconn.edu).

I. INTRODUCTION

A major signal processing challenge facing sonar designers is the automatic detection of passive acoustic signatures. In many cases these signatures take the form of a transient statistical burst; although excellent techniques exist for the detection of such changes under an assumption of statistical independence, there is evidence that this assumption is in many cases unreasonable. In fact, many underwater acoustic transients, either man-made or of biological origin, undergo a cycle of behavior (attack, tonality, decay, etc.), and therefore do not admit an independence assumption.

The dependencies of transient signals are sometimes disguised—not necessarily sample-to-sample correlation, but rather a structural form. As an illustration, consider bursts of random transients buried in Gaussian noise. The appearance is of increased variance of the observations, as illustrated in Figs. 1(a) and 1(b). The cyclic nature of the increased-variance observations, i.e., the on-off nature of the random signal, clearly suggests dependence between clumps of transients, yet all observations are uncorrelated and hence recognition of this dependence using standard techniques would be problematic. To detect such transients, first consider blockwise processing. If the increased variance transient appears in the whole data window, then it is easy to see the likelihood ratio test amounts to an energy detector. However, if the transient appears in only part of the data window, the usual likelihood ratio technique does not apply if the location is unknown. Heuristic approaches have been proposed under various assumptions [7, 15, 16, 25], but their performances rely on the validity of assumptions about the transient signal. For example, in [16], it was assumed that partial knowledge, such as the burst length, is available, and that there is only one burst within a data window; while in [7, 15, 25], assumptions are made regarding the frequency occupancy of the transients.

On the other hand, a change detection scheme such as the CUSUM (cumulative sum) procedure does not require the knowledge of the starting or ending points of the transient signal. Specifically, for a CUSUM scheme, data are processed sequentially as they arrive and a detection is declared whenever the CUSUM exceeds a threshold. The standard Page's test designed for the detection of Gaussian increased variance transients, however, does not fully utilize the on-off property of the transient signals. For a certain transient strength (the variance increase in this case), the detection probability depends on the transient length [10]. Thus the short-duration nature of each component burst could evade a system detection due to the fact that the CUSUM statistic would tend to decrease during the quiescent period. There is no

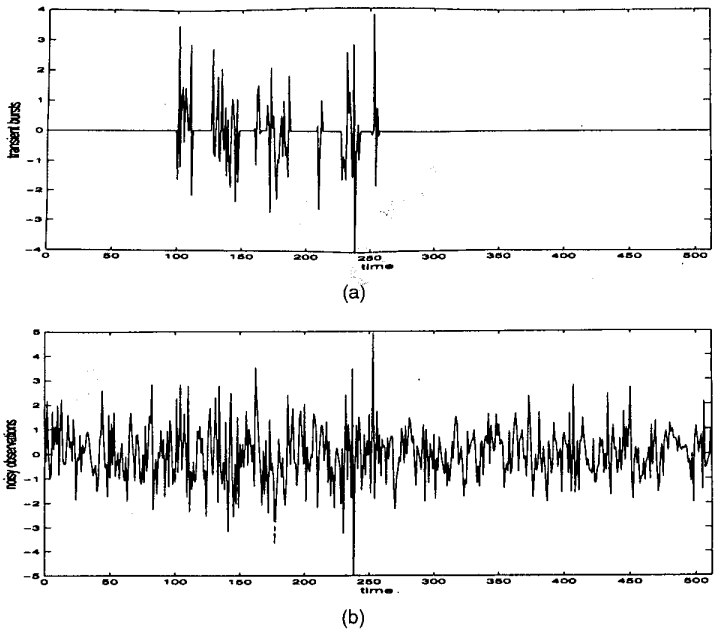


Fig. 1. Examples of Gaussian-bursts transient signal. Transient shown is transient C in Table I. Transient starting time is 100. (a) Generated transient burst. (b) Transient buried in Gaussian background.

aggregation of CUSUM statistic between different clumps if the quiescent period is long enough such that the CUSUM statistic is at or near zero when the next burst starts. The point of this work is that this can be overcome if the cyclic behavior is integrated to the transient model—that is, the quiescent period in between two bursts is now part of the transient model. A natural choice that captures the (stochastic) transition between "on" and "off" states is a two state hidden Markov model (HMM).

Although important, the above example is somewhat simplistic. However, the idea of capturing the possible dependence structure in a transient signal is of great value for improving detection probability. Such dependence not only exists in the time domain, where the occurrence of transient follows cycle of distinct Markov states, it could also occur in the frequency domain where the frequency band of transient signal exhibits contiguity from time to time [24]. In fact, in a paper by the same authors [2], such an HMM structure was used to form the transient (frequency line) detector which exhibits performance close to optimal. Improvement over these detectors, via use of amplitude and phase information, was investigated later [3].

In [2] the detection of a frequency line modeled as an HMM is achieved using blockwise processing. That is, for a fixed data window, the likelihood functions under HMM-present and -absent hypotheses are computed for the block of data, and the detector obtained thereafter. Such blockwise processing loses appeal in practice since the transient signal could start and end at arbitrary unknown times. On the other hand, a sequential change detection scheme, if available, is more attractive in both its easy implementation, performance gain (in terms of both detection probability and delay to detection), and robustness to the unknown starting time of the transient signal.

In each of the above examples, while the transient-present observations could be represented as an HMM, those for transient-absent are independent random noises. A more general model is that both signal-present and signal-absent observations are all HMMs with different parameters. A realistic example is an underwater acoustic system where there are always some background transients (biological sound, for example) going on and our interest is in the detection of those man-made transients. Hence, a generic problem is to detect a "new" HMM in the presence of an "old" one. However, as pointed out in [8], the superposition of two HMMs with common observation space is equivalent to a single HMM with an expanded state whose state transition matrix is simply a Kronecker product of that of the two superimposing HMMs. Therefore our focus here is to detect a possible switch from one HMM to another. A treatment of the detection of superimposed HMM transients is in [9].

The goal of this work is therefore to pursue a CUSUM-like procedure for the detection of change in HMM, i.e., we want to build a Page detector that can quickly detect a switch from one HMM to another. The organization is as follows. Section II gives the necessary background in both Page's test and HMMs.

In Section III, a Page detector based on the idea of a repeated SPRT (sequential probability ratio test, to be discussed in the following section) is proposed for the detection of an HMM transient, followed by performance prediction in terms of average run length (ARL) in Section IV. We study in detail a few important examples in Section V to illustrate the practical implementation of the proposed scheme, as well as the performance gain as opposed to alternative detectors. Section VI gives concluding remarks.

Throughout this work, we deal only with discrete state HMMs, although these may have real- and/or vector-valued observations. We call those HMMs with discrete (finite or countable alphabet) observations discrete, while otherwise we call them continuous. Further, for simplicity, we assume that all HMMs involved are ergodic and distributed according to their stationary distributions at commencement. Such a stationarity assumption amounts to requiring that the initial state of the underlying Markov chain follows its stationary (equilibrium) distribution. The stationary assumption is purely a technical requirement for theoretical development. Our simulation results indicate that the procedures we develop work equally well for a nonstationary HMM.

II. BACKGROUND

A. Sequential Probability Ratio Test

We first revisit the general formulation of the SPRT. The reader is encouraged to refer to [4] for an excellent and thorough treatment of these. Suppose we have the following hypothesis test:

$$\begin{aligned} H &: \mathbf{x}^n = \{x_1, x_2, \ldots, x_n\} \sim P_H, \qquad n = 1,2,\ldots \\ K &: \mathbf{x}^n = \{x_1, x_2, \ldots, x_n\} \sim P_K, \qquad n = 1,2,\ldots . \end{aligned} \tag{1}$$

Note here n is not a fixed number; data arrives continually, and a decision is made when the evidence warrants. For any specified n, the likelihood ratio is

$$\Lambda(n) = \frac{P_K(\mathbf{x}^n)}{P_H(\mathbf{x}^n)} = \frac{P_K(x_1)}{P_H(x_1)} \prod_{i=2}^{n} \frac{P_K(x_i \mid x_{i-1}, \ldots, x_1)}{P_H(x_i \mid x_{i-1}, \ldots, x_1)}. \tag{2}$$

The SPRT, also called Wald's test, consists of choosing real numbers $0 < B < 1 < A < \infty$ and defining

$$N^* = \min\{n \geq 1 : \Lambda(n) \geq A \text{ or } \Lambda(n) \leq B\} \tag{3}$$

to be the stopping time. We say H (K) is accepted if $\Lambda(N^*) \leq B$ ($\Lambda(N^*) \geq A$).

The performance measures are, as in the fixed sample size test, in terms of false alarm rate and probability of missed detection:

$$\begin{aligned} \alpha &= P_H(\Lambda(N^*) \geq A) \\ \beta &= P_K(\Lambda(N^*) \leq B). \end{aligned} \tag{4}$$

The objective, as with conventional hypothesis testing, is to minimize the missed detection probability β while keeping the false alarm rate α as small as possible. Due to its sequential nature, two other indices are also important, namely the ARL under H and K, denoted respectively as T^* and D^*.[1]

$$\begin{aligned} T^* &= E_H N^* \\ D^* &= E_K N^*. \end{aligned} \tag{5}$$

Important in the design of the SPRT are Wald's approximations, which establish the relationship between error probabilities (α and β) and the design parameters (A and B):

$$\begin{aligned} A &\approx (1-\beta)/\alpha \\ B &\approx \beta/(1-\alpha). \end{aligned} \tag{6}$$

The validity of these relies on the assumption that at stopping time, the exiting value of $\Lambda(N^*)$ is identical to one of the two thresholds.

An SPRT is optimal for testing between two independent distributions, in the sense that it minimizes the ARL under both hypotheses given fixed values of error probabilities [27]. This optimality, however, does not in general carry over to dependent observations. This lack of optimality is mostly due to the lack of rigorous mathematical proof; in practice, an SPRT works well even for dependent sequences. Further, Wald's approximation still applies to the dependent case provided the excess over the boundary (to be defined later) is negligible. The derivation of Wald's approximations is solely based on the property of the *likelihood ratio*, with no specification on its structure. This turns out to be important in predicting the performance of the proposed scheme.

B. Page's Test

Page's test [17], also known as the CUSUM procedure, is an efficient change detection scheme. A change detection problem, commonly referred to as change-point problem in the statistic literature, is such that the distribution of observations is different before and after an unknown time n_0; and we want to detect the change as soon as possible. Casting it into a standard inference framework, we have the following hypothesis testing problem

$$\begin{aligned} H &: x(k) = v(k) \qquad 1 \leq k \leq n \\ K &: x(k) = v(k) \qquad 1 \leq k < n_0 \\ & \quad x(k) = z(k) \qquad n_0 \leq k \leq n \end{aligned} \tag{7}$$

where $x(k)$ are observations and $v(k)$ and $z(k)$ are all independent identically distributed (IID), with probability density functions (pdf) denoted as f_H and

[1] To distinguish, we use T^* and D^* to denote the ARL for an SPRT, and reserve T and D for the ARL of a CUSUM.

f_K, respectively. Note that under K the observations are no longer a stationary random sequence; their distribution has a switch at n_0 from f_H to f_K.

The Page decision rule, which can be derived from the generalized likelihood ratio (GLR) test [4], amounts to finding the stopping time

$$N = \arg\min_n \left\{ \left(\max_{1 \le k \le n} L_k^n \right) \ge h \right\} \tag{8}$$

where L_k^n is the log likelihood ratio (LLR) of observation $\{x_k, \ldots, x_n\}$, and $\arg\min_n f(n)$ denotes the value of n that achieves the minimum for $f(n)$. Given that the observations are IID, (8) can be easily reformulated as

$$N = \arg\min_n \left\{ \left(L(n) - \min_{1 \le k \le n} L(k-1) \right) \ge h \right\} \tag{9}$$

where

$$L(k) \triangleq L_1^k = \sum_{i=1}^{k} \left(\ln \frac{f_K(x_i)}{f_H(x_i)} \right) \tag{10}$$

with $L(0) = 0$. This is based on the fact that, given independence,

$$L_1^n = L_1^{k-1} + L_k^n. \tag{11}$$

Equation (9) allows us to write down the standard recursion for the Page's test

$$N = \arg\min_n \{ S_n \ge h \} \tag{12}$$

in which

$$S_n = \max\{0, S_{n-1} + g(x_n)\} \tag{13}$$

and

$$g(x_n) = \ln \left(\frac{f_K(x_n)}{f_H(x_n)} \right) \tag{14}$$

is the update nonlinearity.

Page's recursion assures that the test statistic is "clamped" at zero; i.e., whenever the LLR of current observation would make the test statistic S_n negative (which happens more often when H is true), Page's test restarts at zero. The procedure continues until it crosses the upper threshold h and a detection is claimed. Thus, operationally, Page's test is equivalent to a sequence of SPRTs with upper and lower thresholds h and 0. Whenever the lower threshold 0 is crossed, a new SPRT is initiated from the next sample until the upper threshold h is crossed. A typical test is shown in Fig. 2. The schemes of Figs. 2(a) (corresponding to (8)) and 2(b) (corresponding to (12)) are equivalent, but usually the latter is easier to implement.

In practice, the update nonlinearity $g(x_i)$ need not be an LLR as in (14), since this might not be available as in the case when dealing with composite hypotheses or with hypotheses involving nuisance parameters. For a nonlinearity other than the LLR, a

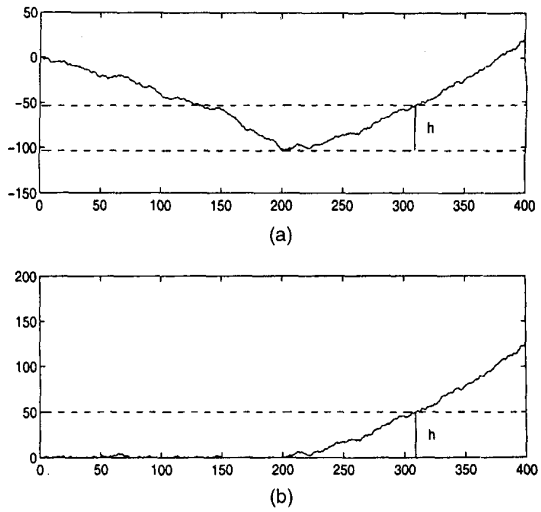


Fig. 2. Example of Page's test statistic. (a) and (b) represent (8) and (12), respectively. They are equivalent interpretations of Page's procedure, yet it is easily seen that the latter is more convenient to implement.

critical requirement for the corresponding CUSUM procedure to work is the "antipodality" condition:

$$\begin{aligned} E(g(x_n) \mid H) &< 0 \\ E(g(x_n) \mid K) &> 0. \end{aligned} \tag{15}$$

There is no false alarm rate or probability of detection involved, since we see from the implementation that, sooner or later, a detection is always claimed as long as the test is "closed" (i.e., $\Pr(N < \infty) = 1$ under both hypotheses). The performance of Page's test is therefore measured in terms of ARL under K and H. It is always desired to have a small delay to detection, usually denoted as D, while keeping the average number of samples between false alarms, denoted as T, as large as possible. Analogous to the conventional hypothesis testing problem where we wish to maximize the probability of detection while keeping the false alarm rate under a fixed level, the trade-off amounts to the choice of the upper threshold h. The relationship between h and the ARL is often calculated in an asymptotic sense using first or second order approximations, usually credited to Wald and Siegmund [23, 26].

As a final note, Page's test using the LLR nonlinearity has minimax optimality in terms of ARL, i.e., given a constraint on the average delay between false alarms, the Page's test minimizes the worst case delay to detection [12, 14].

C. Hidden Markov Models

An excellent tutorial on HMMs can be found in [21]. A discrete time finite-state HMM is a sequence of observations whose distribution at each particular time depends on the state of an underlying

Markov chain. One of its major applications is in speech modeling to capture the temporal variation of short-term spectra of a speech signal. The existence of the Baum–Welch reestimation algorithm [5], which is in fact an application of the EM algorithm [13], makes it a convenient tool for modeling dependent observations.

A discrete HMM is specified by the following parameter triple

$$\lambda = (A, B, \pi) \tag{16}$$

where

$$A = [a_{ij}] = [p(s_{t+1} = j / s_t = i)], \qquad i, j = 1, \ldots, N$$

is the state transition matrix of the underlying Markov chain, where

$$B = [b_{ij}] = [p(x_t = j \mid s_t = i)], \qquad i = 1, \ldots, N; \quad j = 1, \ldots, M$$

is the observation matrix, and where

$$\pi = [\pi_i = p(s_1 = i)] \qquad i = 1, \ldots, N$$

is the initial probability distribution of the underlying Markov states. Implicit to the above notation is the finite number of states (N) and finite alphabet of observations (M). A convenient choice of the initial probability is the stationary distribution of the underlying Markov states, so that the resulting sequence can be regarded as stationary. The joint probability for an HMM sequence is

$$p(s_1, \ldots, s_n, x_1, \ldots, x_n) = \pi_{s_1} \left[\prod_{t=1}^{n-1} a_{s_t s_{t+1}} \right] \left[\prod_{i=1}^{n} b_{s_t x_t} \right]$$

and this can be considered its defining property.

As said before, our goal is to extend Page's test to the detection of the switch between HMMs, thus it is desired that there be an efficient way to compute the likelihood of a sequence of HMM observation given the parameter triple. Fortunately, this is readily solved using the so called forward variable. The forward variable of an HMM is defined as

$$\alpha_t(i) = p(x_1, x_2, \ldots, x_t, s_t = i \mid \lambda). \tag{17}$$

It is easily checked that the following recursion holds for the forward variable

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_{j x_{t+1}} \tag{18}$$

with initial condition

$$\alpha_1(j) = \pi(j) b_{j x_1}. \tag{19}$$

This gives us an efficient way to calculate the likelihood function of an HMM given observations up to the current time. Such efficiency is both relevant and crucial to our approach, the on-line detection of a change in the HMMs.

Although the formulation above is directed toward discrete HMMs, it can be easily modified to apply to HMMs where the observations can take on continuous values. The only modification is to replace the B matrix by a vector of density functions indexed by each state. Hence the forward recursion is written in a more general form:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(x_{t+1}) \tag{20}$$

where $b_j(x_{t+1})$ is the conditional density function (perhaps Gaussian) given the underlying Markov state at time t is j. This extension is useful when dealing with continuous observation HMMs, and we use one as an example in Section VB.

III. CUSUM PROCEDURE FOR DETECTING HMMS

A. CUSUM-like Procedure for Dependent Observations

Consider the same hypothesis test as in (7) except that f_H and f_K are general non-IID probability measures. Assume under K the observations before and after the change are independent of each other.[2] The likelihood ratio (parameterized by n_0) is then

$$\begin{aligned} \Lambda(n; n_0) &= \frac{f(X_1^n \mid K)}{f(X_1^n \mid H)} \\ &= \frac{f_H(X_1^{n_0-1}) f_K(X_{n_0}^n)}{f_H(X_1^{n_0-1}) f_H(X_{n_0}^n \mid X_1^{n_0-1})} \\ &= \frac{f_K(X_{n_0}^n)}{f_H(X_{n_0}^n \mid X_1^{n_0-1})}. \end{aligned} \tag{21}$$

The LLR is then

$$L_k^n = \ln(\Lambda(n; k)) = \sum_{i=k}^{n} \ln \left(\frac{f_K(x_i \mid x_{i-1}, \ldots, x_k)}{f_H(x_i \mid x_{i-1}, \ldots, x_1)} \right). \tag{22}$$

Previously, for the IID case, we claimed that the stopping time could be reduced to (9), and hence the Page test can be derived from a GLR perspective. Here, however, this procedure cannot be repeated exactly due to the fact that $L_1^n \neq L_1^{k-1} + L_k^n$, as seen from (22).

We recall that Page's test is equivalent to a sequence of repeated SPRTs with thresholds h and

[2]This independence assumption is a realistic model in dealing with signals buried in ambient noise. However, if there is an ambient transient, such as one of biological origin as alluded to in Section I, then the independence assumption is no longer valid. The treatment of the detection of a new HMM overlapped with an extant one is dealt with elsewhere [9] and is not discussed here.

0. This idea *can* be carried over to dependent cases, in which each individual SPRT bears the general form depicted in Section IIA. Hence we propose the following CUSUM-like procedure.

1) Start an SPRT with threshold 0 and h.
2) If the SPRT ends at time k with test statistic below zero, reinitiate another SPRT from $k+1$ as if no previous data existed. That is, recalculate the likelihood ratio based on the stationary marginal distribution.
3) Repeat the above procedure until h is crossed.

In compact form, we can write, in a manner similar to the standard Page recursion (12):

$$S_n = \max\{0, S_{n-1} + g(n;k)\} \tag{23}$$

where

$$g(n;k) = \ln\left(\frac{f_K(x_n \mid x_{n-1},\ldots,x_k)}{f_H(x_n \mid x_{n-1},\ldots,x_k)}\right) \tag{24}$$

and x_k is the first sample after the last reset, i.e., $S_{k-1} = 0$. The difference with (22) is that the conditional densities of both numerator and denominator in the logarithm of $g(n;k)$ depend on the same set of random variables, which make a Page-like recursion possible by utilizing the stationary assumption of the HMMs. Note also that such a scheme reduces to the standard Page test with a LLR nonlinearity when the observations both before and after the change are IID. Further, if the observations before change are independent, we need only replace $f_H(x_n \mid x_{n-1},\ldots,x_k)$ with $f_H(x_n)$ for the scheme to work.

The scheme presented here is in the same form as the sequential detector proposed in [1]. It was shown that this procedure is in fact asymptotically optimal in Lorden's sense, i.e., as $h \to \infty$, it minimizes the worst delay to detection given a constraint on the average time between false alarms, among all possible sequential schemes.

B. Justification of Proposed Scheme

The key property that allows a CUSUM-like procedure is the antipodality of the update nonlinearity as specified in (15). Applied to the proposed detector, it amounts to requiring $E(g(n;k) \mid H) < 0$ and $E(g(n;k) \mid K) > 0$, where $g(n;k)$ is defined in (24). We give the proof of $E(g(n;k) \mid H) \le 0$ in the following

$$\begin{aligned}
&E(g(n;k) \mid H) \\
&= \int f_H(x_n, x_{n-1},\ldots,x_k) \\
&\quad \times \ln\left(\frac{f_K(x_n \mid x_{n-1},\ldots,x_k)}{f_H(x_n \mid x_{n-1},\ldots,x_k)}\right) dx_n dx_{n-1}\cdots dx_k \\
&\le \int f_H(x_n, x_{n-1},\ldots,x_k) \\
&\quad \times \left(1 - \frac{f_K(x_n \mid x_{n-1},\ldots,x_k)}{f_H(x_n \mid x_{n-1},\ldots,x_k)}\right) dx_n dx_{n-1}\cdots dx_k \\
&= \int f_H(x_n, x_{n-1},\ldots,x_k) dx_n dx_{n-1}\cdots dx_k \\
&\quad - \int f_H(x_{n-1},\ldots,x_k) f_K(x_n \mid x_{n-1},\ldots,x_k) dx_n dx_{n-1}\cdots dx_k \\
&= 1 - \int f_H(x_{n-1},\ldots,x_k) dx_{n-1}\cdots dx_k \\
&\quad \times \int f_K(x_n \mid x_{n-1},\ldots,x_k) dx_n \\
&= 1 - 1 \\
&= 0.
\end{aligned}$$

The claim that $E(g(n) \mid K) \ge 0$ can be similarly proved. This property is in fact a direct result of the nonnegativity of the (conditional) Kullback–Leibler (KL) distance.

C. Detection of HMMs

In the previous section we have proposed a CUSUM procedure that is applicable to the case of dependent observations *provided* we have an efficient means to calculate the likelihood function. This is not always a reasonable assumption. Fortunately, for the HMM, the existence of the forward variable, together with its recursion formula as discussed in Section IIB, enables efficient computation of the likelihood function of an HMM. Specifically, the likelihood function of an HMM with parameter triple λ could be written as

$$f(x_1, x_2,\ldots,x_t \mid \lambda) = \sum_{i=1}^{N} \alpha_t(i) \tag{25}$$

where N is the total number of states and the α_ts are the forward variables defined in (17).

Now the conditional probability in (22) is readily solved as

$$\begin{aligned}
f_j(x_t \mid x_{t-1},\ldots,x_1) &= f(x_t \mid x_{t-1}, x_{t-2},\ldots,x_1,\lambda_j) \\
&= \frac{\sum_{i=1}^{N} \alpha_t(i)}{\sum_{i=1}^{N} \alpha_{t-1}(i)}
\end{aligned} \tag{26}$$

where $j = H;K$.

Although we have followed the proposed procedure to find the conditional pdf as in (26), this step can in fact be avoided since the likelihood function, defined as the sum of $\alpha_t(i)$, can be used directly by each individual sequential likelihood ratio test. But in practice, it is found, the direct use of the likelihood function as defined in (25) will cause numerical underflow as the number of

observations increases. For discrete HMMs it is easily seen from the definition of the forward variable that the likelihood decreases monotonically (and generally geometrically) with the number of observations. The conditional likelihood function defined in (26) does not suffer such a numerical problem. We need therefore to develop a way of recursively computing the conditional likelihood function in (26) without the direct use of the forward variable. This can be achieved by scaling. Define α_t' such that $\alpha_1'(i) = \alpha_1(i)$, but for $t > 1$

$$\alpha_{t+1}'(j) = \frac{\left[\sum_{i=1}^{N} \alpha_t'(i) a_{ij}\right] b_{jx_{t+1}}}{\sum_{i=1}^{N} \alpha_t'(i)}. \tag{27}$$

It is easily checked that $\sum_{i=1}^{N} \alpha_t'(i)$ is identical to $f_j(x_t \mid x_{t-1}, \ldots, x_1)$ with $j = H, K$ as defined in (26). Thus the updating nonlinearity $g(n;k)$ can be obtained recursively without computing explicitly the exact likelihood function at each time.

To summarize, for the quickest detection of HMMs, we propose the following procedure.

1) Set $t = 1$, $l_0 = 0$, where l_t denotes the LLR at time t.
2) Initialize the (scaled) forward variable α_t' using

$$\alpha_t'(j) = \pi(j) b_{jx_t}$$

for each possible state j and for both hypotheses H and K.
3) Update the LLR

$$l_t = l_{t-1} + \ln\left(\frac{\sum_{i=1}^{N} \alpha_t'(i \mid K)}{\sum_{i=1}^{N} \alpha_t'(i \mid H)}\right). \tag{28}$$

4) If $l_t > h$, declare detection of a change, stop;
If $l_t < 0$, set $l_t = 0$; $t = t + 1$; then goto 2;
If $0 < l_t < h$, continue.
5) Set $t = t + 1$;
Update the scaled forward variable α_t' using (27);
then goto 3.

IV. APPROXIMATION TO ARL

Exact computation of the ARL involves solving Fredholm integral equations [4] which is, in general, difficult and seldom rewarding even in the IID case. If it is assumed that each test ends with the likelihood ratio exactly at a threshold (no "excess over the boundary"), then some simple and useful asymptotic results are available for the IID case. (Siegmund has addressed the excess over the boundary and come up with a more accurate result [23].) But in the case of dependence, none of the above approaches, including the exact computation via solving Fredholm integral equations, applies. In [19] Phatarfod attempted to address the performance of an SPRT between different Markov processes. Although the mathematics there is quite nice in that the author arrived at an equation similar to Wald's identity, the practical implementation of it is involved even for the simplest case of a two-state Markov process. Further, since the approach is rooted in an assumption of observable Markovianity, it does not apply to and apparently cannot be extended to an HMM.

For the general dependent case, it is proposed [4] that no analytic solution is feasible for the ARL under both hypotheses. Instead, a weak performance measure, the KL distance between the two distributions to be tested in the dependence case, is used [4]. These results are clever, but while the KL distance does reflect in a general way how well a Page test behaves given two particular distributions, it does not calibrate the effect of the choice of the threshold on the detection performance (ARL) for the Page test, and hence does not provide a stand-alone detector design. The goal in this section is therefore to try to approximate the ARL, and thus provide some guidance in the design of the Page's test for detecting an HMM, specifically in the choice of the upper threshold h. We start with T, the average number of samples between false alarms.

A. Approximation of T. Average Delay Between False Alarms

A straightforward way of obtaining T is via simulation. However, direct Monte Carlo simulation is computationally expensive due to the fact that the desired T is usually very large. In [1], an ingenious lower bound for such a procedure was obtained. However the bound is introduced for the purpose of proving the asymptotic optimality of a proposed detector. Its reliance on the "probability of finiteness" of the extended stopping time, first introduced by Lorden [12], seriously hampers its applicability to prediction of the ARL; the Page's tests used in practice are all finite with probability one. In this section, we propose an approximation method which requires only a small amount of simulation and provides a reasonably good prediction.

Recall that a CUSUM procedure can be regarded as a repeated SPRT with thresholds h and 0. Thus our approach is based on the idea of finding the relationship of ARL under H between the CUSUM procedure and the corresponding Wald test. Let N be the stopping time for the CUSUM procedure; N^* (N_1^*) be the stopping time of the corresponding SPRT that ends at 0 (h); and i be the number of zero-resettings before the final exceedance of the upper threshold h. Thus i is equivalently the number of SPRT runs in the CUSUM procedure which end at zero. Denote by α the false alarm rate of such an SPRT with h and 0 as its log thresholds. It is straightforward to see that the following relationship is true, provided the processes

involved are stationary:

$$\begin{aligned} T &\triangleq E_H N \\ &= \sum_{n=0}^{\infty}[P(i=n)(E_H N_1^* + nE_H N^*)] \\ &= \sum_{n=0}^{\infty}[\alpha(1-\alpha)^n(E_H N_1^* + nE_H N^*)] \\ &= E_H N_1^* + \frac{1-\alpha}{\alpha}E_H N^* \\ &\approx E_H N^*/\alpha \end{aligned} \tag{29}$$

where the last approximation follows from the fact that usually α is very small. Thus we have reduced the problem of estimating T to evaluating the two quantities involved in (29): $E_H N^*$, the ARL of a single SPRT under H given that the lower threshold is crossed, and α, the false alarm rate of a single SPRT.

Consider $E_H N^*$ first. Given the fact that each SPRT starts at zero and the lower threshold of the SPRT is zero, then with high probability the test will end quickly (usually in one or two steps) under H. Exact evaluation of $E_H N^*$ is not possible, but approximation of it as unity is reasonable.

The next quantity involved is the false alarm rate for a single SPRT. We have stated that Wald's approximation holds true even with dependent observations, *provided* the excess over the boundary is negligible. Starting from (6) simple algebra gives the approximate false alarm rate in terms of thresholds

$$\alpha = \frac{1-B}{A-B}. \tag{30}$$

Note here A and B are upper and lower thresholds corresponding to the original likelihood ratio. Hence for the CUSUM procedure, their corresponding values are e^h and $e^0 = 1$. Substituting them into (30), we get $\alpha = 0$, which indicates a problem.

That problem is, as usual, the excess over the boundary. The assumption behind Wald's approximation is that at the stopping time, the CUSUM statistic coincides with one of the thresholds. In practice, however, this is unlikely. For our case the corresponding lower threshold for the LLR is zero, and it is to be expected that the boundary excess is significant enough to ruin the validity of Wald's approximation. Suitable modification of Wald's approximation is attainable by addressing the excess over the boundary. To do this, define $Q_n = \{\vec{X}_\infty : N(\phi) = n, \text{ and } L_n(\vec{X}_n) \le B\}$. In words, Q_n is the set of sequences with stopping time n and decision H. Hence the decision region for H is $\Omega_H = \bigcup_{n=1}^{\infty} Q_n$, where by definition Q_n and Q_m are disjoint given $n \neq m$. Now we have

$$\begin{aligned} \beta &= \Pr(\vec{X}_\infty \in \Omega_H \mid K) \\ &= \sum_{n=1}^{\infty}\Pr(\vec{X}_\infty \in Q_n \mid K) \\ &= \sum_{n=1}^{\infty}\int_{Q_n} f(\vec{X}_\infty \mid K)\,d\vec{X}_\infty \\ &= \sum_{n=1}^{\infty}\int_{|\vec{X}_\infty : N(\phi)=n,\ L_n(\vec{X}_n)\le B|} f(\vec{X}_n \mid K)\,d\vec{X}_n \\ &= \sum_{n=1}^{\infty}\int_{|\vec{X}_\infty : N(\phi)=n,\ L_n(\vec{X}_n)\le B|} L_n(\vec{X}_n) f(\vec{X}_n \mid H)\,d\vec{X}_n \\ &= \overline{B}(1-\alpha) \end{aligned}$$

where $\overline{B}$ is defined as

$$\begin{aligned} \overline{B} &\triangleq \frac{\sum_{n=1}^{\infty}\int_{|\vec{X}_\infty : N(\phi)=n,\ L_n(\vec{X}_n)\le B|} L_n(\vec{X}_n) f(\vec{X}_n \mid H)\,d\vec{X}_n}{1-\alpha} \\ &= \frac{\sum_{n=1}^{\infty}\int_{|\vec{X}_\infty : N(\phi)=n,\ L_n(\vec{X}_n)\le B|} L_n(\vec{X}_n) f(\vec{X}_n \mid H)\,d\vec{X}_n}{\sum_{n=1}^{\infty}\int_{|\vec{X}_\infty : N(\phi)=n,\ L_n(\vec{X}_n)\le B|} f(\vec{X}_n \mid K)\,d\vec{X}_n}. \end{aligned}$$

It is easy to see that $\overline{B}$ is the posterior mean of test statistic given a correct decision for H. Therefore Wald's approximations remain reasonable after replacement of B by $\overline{B}$.

At this point, we are able to obtain α via the modified Wald's approximation provided we have a good estimate of $\overline{B}$. Further, with α available, T can be obtained as

$$T \approx 1/\alpha = \frac{e^h - \overline{B}}{1-\overline{B}} \approx \frac{e^h}{1-\overline{B}}. \tag{31}$$

The remaining issue is how to evaluate $\overline{B}$. An analytic solution is not feasible except for the simple IID case. In general it must be obtained via simulation; but to simulate $\overline{B}$ is far easier than direct simulation of T. For example, in most cases, a thousand samples will give a reasonable estimate of $\overline{B}$ while direct simulation of T in the magnitude of 10^6 would generally require over a million samples for each Monte Carlo run, and many runs would be necessary.

Assuming that $\overline{B}$ is obtained accurately, the approximation (31) actually provides a lower bound on T for the following reasons. First, the approximation in (29) introduces a bias of $-E_H N_1^* + E_H N^*$ which is always negative for $h > 0$ (it takes on average more samples for the CUSUM statistic to cross h than 0 under the H hypothesis). Second, the approximation of $E_H N^* = 1$ actually provides a lower bound on $E_H N^*$ as the stopping time $N^* \ge 1$ with probability 1. Third, we have omitted the overshoot over h; i.e., for a more accurate result, e^h in (31) should be replaced by $\overline{e}^h$ where $\overline{e}^h$ is the average value

of CUSUM statistic in the likelihood ratio scale at the stopping time (first passage over h). Nonetheless, as we see later by simulation, the above approximation does provide a reasonably close lower bound on that obtained from direct Monte Carlo simulation and therefore could serve as a guidance in threshold selection. Improvement is possible by addressing the above three points.

B. Approximation of D

The prediction of delay to detection, denoted as D, is more straightforward. For the IID case, it is intuitive to write down, for a large value of h,

$$D \triangleq E_K N = \frac{h}{E_K\{g(x)\}} \tag{32}$$

where h is the threshold and $E_K\{g(x)\}$ is the mean of the nonlinearity under K. The above equation follows directly from Wald's first equation [22] provided $S_N = h$, where S_N is the CUSUM statistic at the stopping time. This, again, is based on an assumption of negligible overshoot, which explains why it holds for only large values of h. For the dependent case with the proposed scheme, however, the nonlinearity is the logarithm of the ratio of the two conditional likelihoods whose expectations are not identical for different times.

For HMMs with finite-state Markov chains, however, there hold some strong results which can be used to facilitate our analysis. Specifically, suppose λ_0 and λ_1 are two HMM parameters with ergodic Markov chains, or equivalently, its transition matrix is both aperiodic and irreducible. Then it was shown that the following quantity always exists [18]:

$$H_{\lambda_1}(\lambda_0) = \lim_{n\to\infty} \frac{1}{n} \log(P(\mathbf{x} \mid \lambda_0))$$

where $\mathbf{x}$ denotes samples up to time n from the HMM with parameter λ_1. Note $H_{\lambda_1}(\lambda_1)$ always exists for HMMs with ergodic Markov chains as it actually defines the entropy rate (ergodicity allows us to replace time average with ensemble average). Hence the following limit always exists:

$$\begin{aligned} \lim_{n\to\infty} g_n(\lambda_1, \lambda_0) &= \lim_{n\to\infty} \frac{1}{n} \log\left(\frac{P(\mathbf{x} \mid \lambda_1)}{P(\mathbf{x} \mid \lambda_0)}\right) \\ &= H_{\lambda_1}(\lambda_1) - H_{\lambda_1}(\lambda_0). \end{aligned} \tag{33}$$

The existence of such a quantity was also demonstrated later in [11] through extensive numerical examples. It is easy to notice that the limit of $g_n(\lambda_K, \lambda_H)$ is essentially the asymptotic per sample CUSUM statistic used in our proposed scheme, with λ_1 and λ_0 being the parameters of the HMMs under hypotheses K and H, respectively. Thus it plays a role similar to that of $E_K\{g(x)\}$ in (32) with an IID sequence, provided h is large enough. A typical plot

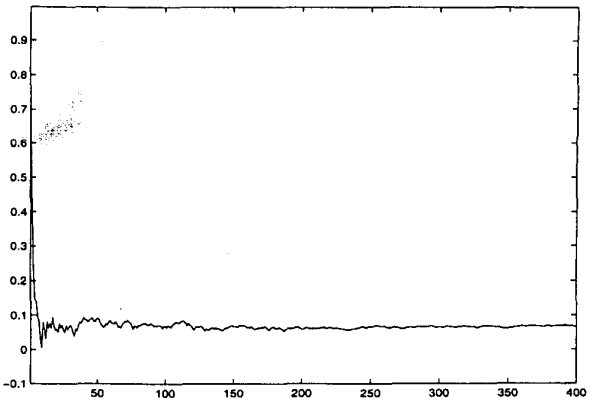


Fig. 3. Typical behavior of $g_n(\lambda_K, \lambda_H)$ from (33), of HMM under transient-present hypothesis, plotted against n. Example is discrete HMM in Section VA, whose nonlinearity stabilizes around 0.068.

of $g_n(\lambda_1, \lambda_0)$ is shown in Fig. 3 using the discrete HMM example of the next section.

Further, by invoking ergodicity, we see $H_{\lambda_0}(\lambda_0) - H_{\lambda_0}(\lambda_1)$ defines exactly the KL distance between two HMMs. We comment here that T is also a function of the KL distance between the two HMMs although the relationship is not as straightforward as with D. This justifies the use in [4] of KL distance as a (weak) performance measure of Page's test.

Knowing that T is exponential while D is linear in h, we can conclude, as in the IID case, that T is asymptotically exponential in D. Thus we can define, as in the IID case, the asymptotic efficiency of a Page's test

$$\eta = \lim_{h\to\infty} \log(T)/D. \tag{34}$$

Naturally η is a means to compare detection strategies: the larger η, the better. But perhaps more important is that fact that we *can* define it, or rather that the limit is non-trivial. The log-linear behavior is why Page-like procedures work. If an operating point (h) yields false alarms on average every million samples, and this is deemed too high, then it is possible to raise this to a false alert once in 10^{12} samples at the expense only of an approximate doubling in delay to detection.

V. EXAMPLES

In this section, several examples illustrate the application of the proposed scheme. The first example, between two discrete HMMs, serves for illustration due to its simplicity. Nonetheless, discrete HMMs have very important application in speech processing, where the underlying Markov state usually represents individual phonemes. The transition between different states (phonemes) is governed by phonological rules of each word or phrase. Detection of the change from one HMM to the other serves as a preliminary step in speech segmentation, and its effectiveness greatly affects further processing.

The second example is that involving increased-variance Gaussian-bursts, as discussed in Section I. The emphasis in this example is how to use an HMM to capture some important features of the transient signal in order to achieve better detection performance.

The last example is on the detection of a narrowband transient whose center frequency varies slowly in time. Here each state of the Markov chain represents an individual frequency band (could be the frequency bin output of a discrete Fourier transform (DFT), or the output from a bank of narrowband filters). The slow-varying nature of the center frequency results in frequency contiguity in the time-frequency domain. This allows us to represent the "wandering" of the frequency line among different frequency bands as transitions among different states, thus setting up the model for the application of the proposed procedure.

A. Discrete HMMs

The example is taken from [11] where the two HMMs that need to be distinguished have the following parameters:

$$H: A_0 = \begin{bmatrix} 0.800 & 0.150 & 0.05 & 0.00 \\ 0.070 & 0.750 & 0.12 & 0.06 \\ 0.050 & 0.140 & 0.80 & 0.01 \\ 0.001 & 0.089 & 0.11 & 0.80 \end{bmatrix}$$

$$B_0 = \begin{bmatrix} 0.30 & 0.40 & 0.20 & 0.10 \\ 0.50 & 0.30 & 0.10 & 0.10 \\ 0.10 & 0.20 & 0.40 & 0.30 \\ 0.40 & 0.30 & 0.10 & 0.20 \end{bmatrix}$$

$$K: A_1 = \begin{bmatrix} 0.400 & 0.250 & 0.15 & 0.20 \\ 0.270 & 0.450 & 0.22 & 0.06 \\ 0.350 & 0.140 & 0.40 & 0.11 \\ 0.111 & 0.119 & 0.23 & 0.54 \end{bmatrix}$$

$$B_1 = \begin{bmatrix} 0.10 & 0.15 & 0.65 & 0.10 \\ 0.20 & 0.30 & 0.40 & 0.10 \\ 0.30 & 0.30 & 0.10 & 0.30 \\ 0.15 & 0.25 & 0.40 & 0.20 \end{bmatrix}.$$

The stationary distributions could be computed as follows:

$$H: \pi_0 = [0.2042 \quad 0.3484 \quad 0.3266 \quad 0.1208]'$$

$$K: \pi_1 = [0.2931 \quad 0.2430 \quad 0.2459 \quad 0.2180]'.$$

An example test statistic is shown in Fig. 4. The transient HMM, specified by A_1 and B_1, starts at $n = 100$ and ends at $n = 500$. It can be seen that at $n = 100$, the test statistic responds promptly to the occurrence of the transient HMM, while after $n = 500$, the decrease of the CUSUM statistic reflects the end of the transient.

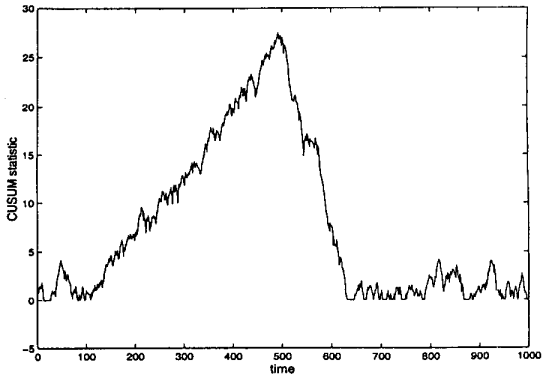


Fig. 4. Detection between four-state HMMs. Transient HMM starts at $n = 100$ and ends at $n = 500$. CUSUM statistic is seen to respond well to appearance of new HMM.

Next we study the performance in terms of ARL using both numerical simulation and the approximation method as described in Section IV. $\overline{B}$, the average value of test statistic of individual SPRT that ends below the lower threshold is found to be 0.801. Also, the asymptotic value of the nonlinearity $g_n(\lambda_1, \lambda_0)$ defined in (33) was found to be 0.068. We plot out the approximation of both T and D against h together with the simulation result. (See Fig. 5.) As expected, the prediction for T is pessimistic due to the approximation. On the other hand, the prediction for D conforms well with the simulation.

B. Increased Variance Gaussian Transients

1) *The Model*: In this section, we study in some detail the example of increased-variance Gaussian-bursts transient. In this model, white Gaussian noise with variance σ_n^2 is always present. If a transient occurs (i.e., if $n > n_0$ in (7)) then added to this is a sequence of independent random variables having variance σ_t^2 (if $u(n) =$ on) or 0 (if $u(n) =$ off). The *hidden* sequence $\{u(n)\}$ forms a Markov chain, as discussed below; but provided $u(n) = u(n-1)$ with high probability, a transient appears as "clumps" of increased-variance observations.

We note that the durations of both quiescent and active times are assumed random. For any specific transient, though, it is usually reasonable to assume we have access to *a priori* information regarding the *average* duration time of both transient-active and -quiescent period. Due to the sampling in practice, such average duration is usually represented by an average number of samples of each state, which we denote by n_{on} and n_{off}, respectively. Further, define $p = 1/n_{\text{on}}$ and $q = 1/n_{\text{off}}$. We propose the following transition matrix for the switch between on and off state

$$A = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}. \tag{35}$$

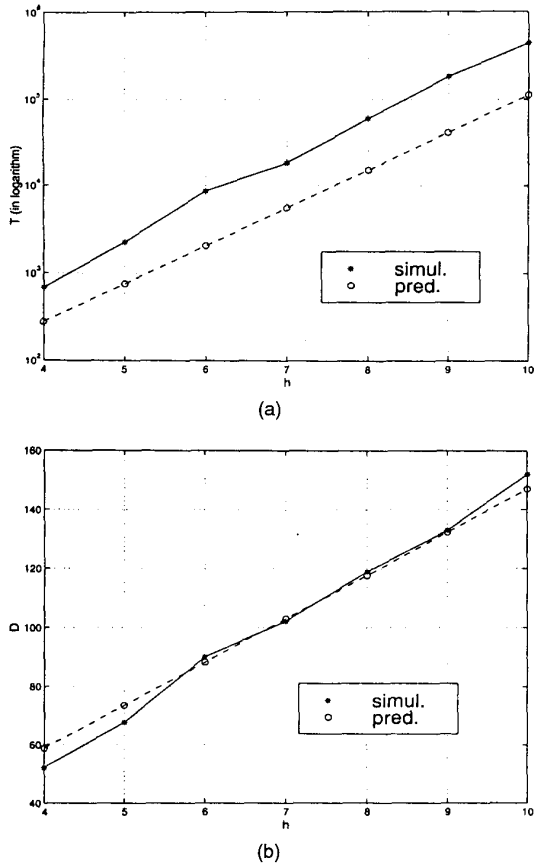


Fig. 5. ARL for four-state HMM example. (a) T as function of h. (b) D as function of h.

Note the stationary distribution is

$$\left[\frac{q}{p+q}, \frac{q}{p+q}\right] = \left[\frac{n_{\text{on}}}{n_{\text{on}}+n_{\text{off}}}, \frac{n_{\text{off}}}{n_{\text{on}}+n_{\text{off}}}\right]$$

which are exactly the average number of samples each state is occupied.

2) *Comparison with Standard Page Test*: As mentioned in Section I, for increased-variance Gaussian-bursts, a standard Page test could be used for detecting such a transient without using the relationship between neighboring bursts. In [28], it was demonstrated that the standard Page's test was among the best detectors for such transients. In this section, we investigate the performance gain over the standard Page detector achievable by incorporating the structural relationship between different bursts using HMMs.

We first use both the HMM-based Page test and standard Page test in a blockwise fashion.[3] That is,

[3]Such a realization is not consistent with the philosophy of Page's test, nor should it be implemented this way in practice. Nonetheless, it provides an easy way to compare Page's test to those detectors admitting only blockwise processing such as the energy detector. Further, the resulting receiver operating characteristic (ROC) curve provides illustration of performance gain.

TABLE I
Gaussian Burst Transients to be Studied

	Transient A	Transient B	Transient C
Signal variance σ_s^2	1	1	1.5
Per sample SNR	0 dB	0 dB	1.76 dB
Number of bursts	8	15	8

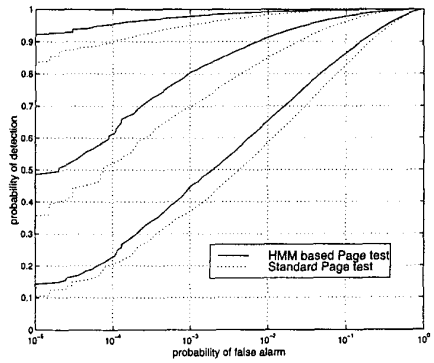


Fig. 6. ROC curves for Gaussian-bursts example. Three pairs of curves correspond to transients A, B, C of Table I, referenced from bottom to top.

we first divide input data sequence into data blocks. Within each block, we run the standard Page test and the HMM-based Page test, and declare a detection if the CUSUM statistic exceeds the threshold within the block. In our simulation, the block size is $N = 512$. The average burst length and quiet period are $n_{\text{on}} = 8$ and $n_{\text{off}} = 10$ samples, respectively, resulting in a transition matrix

$$A = \begin{bmatrix} 0.875 & 0.125 \\ 0.100 & 0.900 \end{bmatrix}.$$

The noise variance $\sigma_n^2 = 1$. We study three different transients with different signal variances and number of bursts within each block. These three transients are listed in Table I.

The standard Page test uses the LLR nonlinearity to update the CUSUM statistics. Such a nonlinearity is easily shown to be

$$g(x) = x^2\left(\frac{1}{2\sigma_n^2} - \frac{1}{2\sigma_t^2}\right) - \log\left(\frac{\sigma_t}{\sigma_n}\right) \qquad (36)$$

where $\sigma_t^2 = \sigma_n^2 + \sigma_s^2$ is the variance of noise plus transient samples. For the HMM-based Page detector, we use the algorithm described in Section IIIC.

A typical example of such a transient is shown in Fig. 1 using parameters specified by transient C in Table I. Fig. 1(a) is the transient signal and Fig. 1(b) is the noisy observations within a block. The ROC curves for the transient models using both the standard Page procedure and the HMM-based procedure are shown in Fig. 6. In all three cases the HMM-based Page test exhibits significant improvement over the standard Page test. The

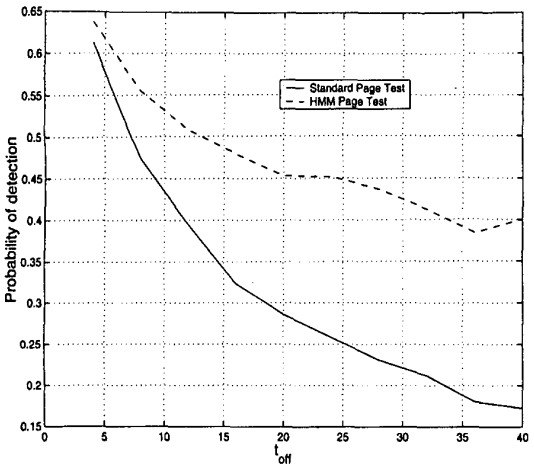


Fig. 7. Probability of detection as function of t_{off}. Here t_{on} is fixed at 8 throughout simulation and false alarm rate is kept at 0.005.

underlying reason that the HMM-based Page test outperforms the standard Page test is that the "quiet" period is built into the model itself to account for the cyclic behavior of the transient bursts.

To study further the reason for this performance divergence, we simulate the detection performance of both detectors as a function of n_{off} for a fixed $n_{\text{on}} = 8$ and a given false alarm rate. In Fig. 7, in which the false alarm is fixed at 0.005, we see first that the performances of both detectors degrade as n_{off} increases. This is not surprising; perhaps less expected is that the performance divergence between the standard Page detector and the HMM-based Page detector *widens* as n_{off} increases. The reason of this behavior can be explained as following. Denote $f_1(\cdot)$ and $f_2(\cdot)$ as pdfs of Gaussian distributed random variable with variance σ_t^2 and σ_n^2 respectively. Then the nonlinearity of the standard Page test is simply

$$g_{\text{page}}(x) = \log\left(\frac{f_1(x)}{f_2(x)}\right) \tag{37}$$

which can be further simplified to (36). Now write the forward recursion of α', defined in (27), as

$$\begin{aligned}\alpha'_{t+1}(1) &= \frac{\alpha'_t(1)a_{11} + \alpha'_t(2)a_{21}}{\alpha'_t(1) + \alpha'_t(2)} f_1(x) \\ \alpha'_{t+1}(2) &= \frac{\alpha'_t(1)a_{12} + \alpha'_t(2)a_{22}}{\alpha'_t(1) + \alpha'_t(2)} f_2(x)\end{aligned} \tag{38}$$

where a_{ij} are elements of the transition matrix A. Using the fact that $a_{11} + a_{12} = a_{21} + a_{22} = 1$, we can show that the coefficients of $f_1(x)$ and $f_2(x)$ add to unity. Through substitution of (38) to (28), we get the nonlinearity used in the HMM-based Page test

$$g_{\text{hmm page}}(x) = \log\left(\frac{\lambda f_1(x) + (1-\lambda) f_2(x)}{f_2(x)}\right) \tag{39}$$

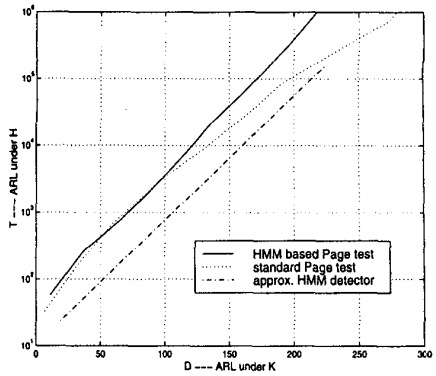


Fig. 8. T versus D for Gaussian-bursts example.

in which λ varies with time. The key, therefore, is that the HMM-based Page procedure uses an update appropriate for detection of Gaussian *mixture* statistics, rather than a single (increased-variance) Gaussian. Thus, lower variance samples from a "quiet" period have less effect; further, since $a_{21} = 1/t_{\text{off}}$ (cf., (35)), it is reasonable that the weight on the increased-variance Gaussian (λ in (39)) decreases with an increase in the average time between bursts t_{off}. Thus an HMM based Page scheme is more ready to accept a series of low-variance observations (a "quiet" period between bursts) as just that, rather than as evidence refuting the presence of a transient.

A more natural performance measure for the Page test (both standard and HMM-based) is in terms of ARL. In Fig. 8 we present the simulation result of T versus D for both standard Page's test as well as the HMM-based Page detector. The advantage of the HMM-based Page detector is obvious. Also plotted is the approximation using the method described in Section IV which gives a lower bound of the ARL curve. Note that for the Gaussian bursts transient and a standard Page test, the log linearity of T with respect to h need not be true. This is so because the assumption intrinsic to the standard Page test, that the samples be identically distributed for the transient-present data, is no longer valid.

C. Detection of Slowly Varying Narrowband Transients

The idea of modeling a slowly varying frequency line via an HMM appears to be first due to Barrett and Streit [2, 24]. In their work, a frequency line is assumed to be a constant-amplitude sinusoid with slowly varying frequency, buried in Gaussian noise. The output envelope of a narrowband filter (e.g., DFT) therefore assumes either a Rayleigh (noise only) or a Rice (noise plus transient) distribution. Our model is slightly different in that we do not assume constant amplitude for the frequency line. Instead, we assume that the transient is a narrowband Gaussian process, an assumption more in line with [7, 15, 25]. Variation of the center frequency could arise from the

Doppler shift due the motion of the transient source. Thus the slowly varying nature, or contiguity, of the frequency is governed by the continuity of the target motion. If we pass such a narrowband Gaussian transient buried in Gaussian noise through a bank of narrowband filters, and we take the magnitude square of each filter output, we then have the following scenario. The observations are now a vector sequence with the dimension equal to the number of narrowband filters. Under H, our observations are an IID vector sequence with each element in the vector being unit-exponentially distributed (assuming appropriate normalization). Under K, there exists one element to each vector that has an exponential distribution with increased scale parameter 1 + SNR (signal-to-noise ratio), i.e., whose pdf is written as $p(z_t(k)) = (1/(1 + \text{SNR})) * \exp(-z_t(k)/(1 + \text{SNR}))$. Such a cell corresponds to the frequency band in which the transient resides. The approach is to model the transient index k as a Markov chain with a certain transition probability A. Such a transition matrix should have the property that those elements at or near the diagonal have larger values to account for the "slowly varying" nature of the signal. In [24] the authors used a Gaussian approximation; that is, each row of the transition matrix follows a Gaussian density centered at its diagonal element. This is based on the assumption that the frequency bin of the transient at the next time slot is Gaussian distributed with mean value equal the current frequency bin occupancy. It was found however, that the detection performance is insensitive to the transition matrix as long as the large diagonal (and near diagonal) element requirement is satisfied.

We remark that the authors in [2, 24] obtained an observation matrix by setting a threshold for each individual frequency bin, and therefore transformed the problem into a discrete HMM. The advantage of such an approach is its simple implementation. However, the thresholding operation results in a loss of information contained in the output amplitude. In [3], such amplitude information was kept only for those bins that pass the threshold. Treating the transient as a continuous HMM, i.e., using a vector of density functions instead of the observation matrix obtained via thresholding, enables us to fully exploit the information carried in the magnitude of the DFT outputs.

An example is shown in Fig. 9. The total number of narrowband filters is taken as 8, thus resulting in an 8-state Markov chain. The transition matrix is obtained using a standard Gaussian approximation with appropriate normalization. We choose SNR = 1. The frequency occupancy of the narrowband transient is shown in Fig. 9(a). The transient starts at $n = 30$. Fig. 9(b) is a time-frequency plot of the simulated magnitude-square output from the bank of narrowband filters which assumes either unit exponential or exponential with scale parameter 2 depending on whether the transient is present. It is hard to discern the frequency line from these data. The CUSUM statistic using the HMM-based Page test, however, can easily pick up at the start of the transient, as indicated in Fig. 9(c).

We further compare the performance of our detector with the maximum power detector (MPT) which has better performance than the HMM-based detector of [2]. The MPT statistic simply sums up the maximum output bin from each vector output. Such a detector could be derived from a GLR perspective with the assumption that at any time instant, one and only one frequency bin could contain the possible transient. The superiority of the HMM-based detector over MPT is obviously seen from the ROC curves in Fig. 10. The transient used is the same as those depicted in Fig. 9. Such performance gain, however, is at the cost of increased computational complexity.

In the above comparison we have compared blockwise processing (the MPT detector) with a sequential method (our HMM-based Page detector). For a comparison which is perhaps more appropriate, we develop the Page detector using the maximum power output by getting the optimal LLR nonlinearity, which turns out to be

$$g(t) = \log\left[\frac{n-1}{n}\frac{1-e^{-y(t)/\mu_1}}{1-e^{-y(t)/\mu_0}} - \frac{\mu_0}{\mu_1}e^{y(t)/\mu_0 - y(t)/\mu_1}\right]$$

where t is the time index and $y(t)$ is the maximum bin output (see the Appendix). The Page detector is therefore simply the threshold testing of the following statistic,

$$S_t = \max(0, S_{t-1} + g(t)). \tag{40}$$

The dotted line in Fig. 10 is in fact obtained using S_t in (40) in a blockwise fashion, which has inferior performance to both the HMM-based Page detector and MPT detector. The simulation result regarding the ARLs of the MPT and HMM-based Page detectors are shown in Fig. 11: the HMM-based Page detector clearly outperforms the Page detector based on the MPT statistic.

VI. SUMMARY

The problem studied in this paper is how to quickly detect an HMM transient. By decomposing a Page's test to a repeated SPRT, we were able to derive a CUSUM-like procedure for the detection of a distributional change with dependent observations, to justify it, and to approximate its performance as parametrized by the threshold. Utilizing the forward variable of an HMM, such a procedure was applied to HMM transients. We have investigated in detail its application to several important transient detection problems. Our emphasis is to stress the usefulness of HMM in capturing certain features of the transient that can be utilized to improve the detector performance. The advantage of the

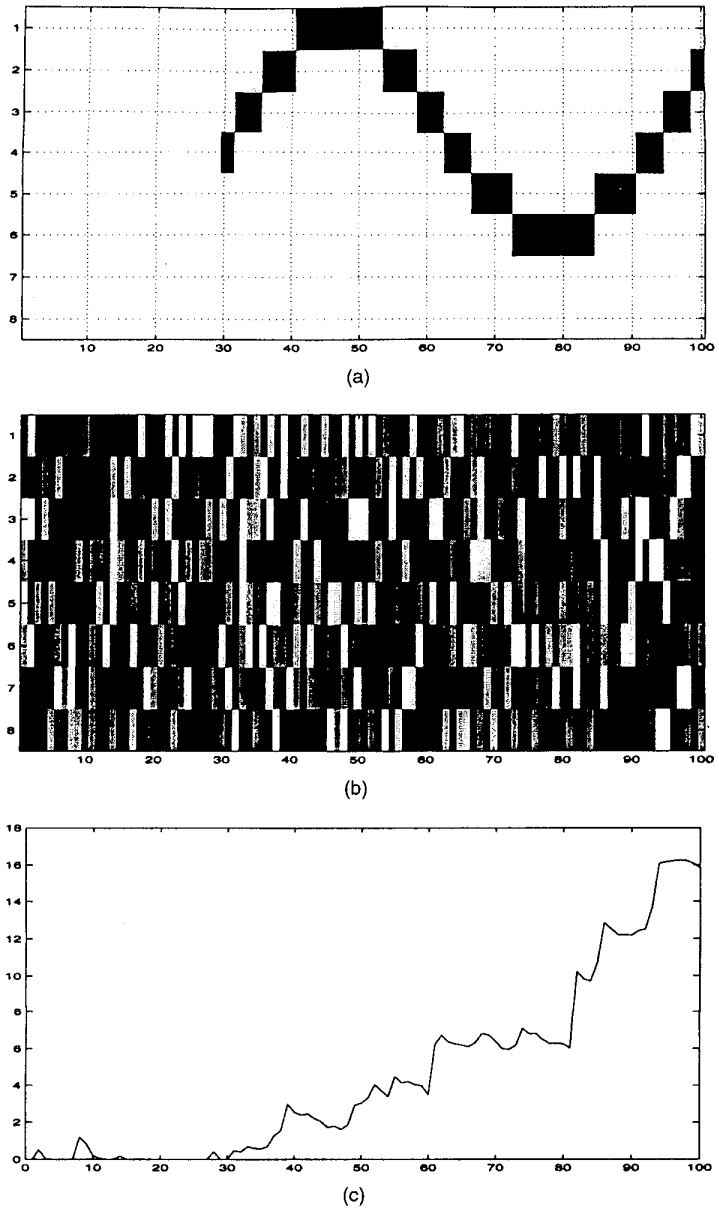


Fig. 9. Detection of narrowband transient with slowly varying center frequency. (a) Frequency occupancy of transient. Frequency line starts at 30th sample. (b) Typical plot of frequency bin output with SNR = 1. Frequency line hard to distinguish from noisy background. (c) CUSUM statistic of HMM-based Page detector. As expected, it responds promptly at occurrence of frequency line.

HMM-based CUSUM procedure was confirmed by simulation.

APPENDIX. DERIVATION OF PAGE DETECTOR BASED ON MAXIMUM DFT OUTPUT FOR NARROWBAND TRANSIENTS

Denote by $\mathbf{x}(t)$ the magnitude square DFT outputs at time t. The statistic of interest is then

$$Y(t) = \max(\mathbf{x}(\mathbf{t}))$$

where we assume $\mathbf{x}(t)$ is of size n. Under the noise-only hypothesis, $\mathbf{x}(\mathbf{t})$ follow identical exponential distributions with parameter μ_0. Under the noise-plus-narrowband-signal alternative, one of the DFT bin outputs follows an exponential probability law with increased mean value μ_1, while the rest have parameter μ_0. Our purpose is to find the LLR of $Y(t)$.

Under H, the statistic is simply the maximum of n exponentially distributed random variables whose distribution function is easily obtained as

$$P(T_H < y) = \prod_{i=1}^{n} P(x_i < y) \tag{41}$$

$$= (1 - e^{-y/\mu_0})^n. \tag{42}$$

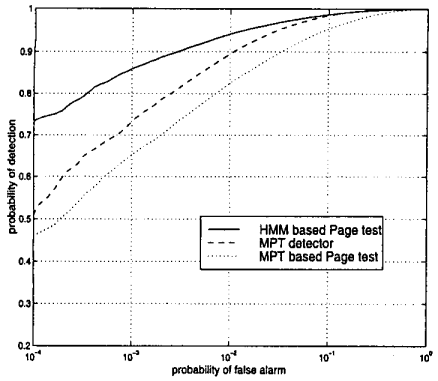


Fig. 10. ROC curves of both HMM-based Page detector and MPT for detection of narrowband transient as shown in Fig. 9.

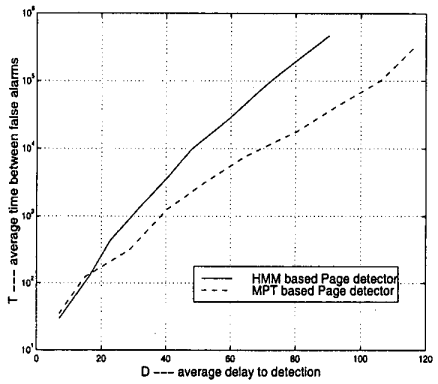


Fig. 11. ARL curves of both HMM-based Page detector and MPT-based Page detector. Superior performance of HMM-based Page detector lies in hidden Markov modeling that captures continuity of center frequency of narrowband transients.

The probability density is obtained by taking the derivative:

$$f_{Y_H}(y) = n(1 - e^{-y/\mu_0})^{n-1}\left[\frac{1}{\mu_0}e^{-y/\mu_0}\right].$$

To find the density function of the statistic under K, note that it is essentially the maximum of the following two random variables, Y_1 and Y_2, i.e., $Y_K = \max(Y_1, Y_2)$ with Y_1 following

$$f_{Y_1}(y) = n(1 - e^{-y/\mu_0})^{n-1}\left[\frac{1}{\mu_0}e^{-y/\mu_0}\right]$$

and Y_2 exponential with parameter μ_1. Hence the distribution function for Y_K is

$$P(Y_K < y) = (1 - e^{-y/\mu_0})^{n-1}(1 - e^{-y/\mu_1}).$$

The pdf is therefore

$$f_{Y_K}(y) = (1 - e^{-y/\mu_0})^{n-2}\left[(n-1)(1 - e^{-y/\mu_1})\frac{1}{\mu_0}e^{-y/\mu_0} + (1 - e^{-y/\mu_0})\frac{1}{\mu_1}e^{-y/\mu_1}\right].$$

Now the LLR is readily solved as

$$g(y(t)) = \log\left[\frac{n-1}{n}\frac{1 - e^{-y(t)/\mu_1}}{1 - e^{-y(t)/\mu_0}} - \frac{\mu_0}{\mu_1}e^{y(t)/\mu_0 - y(t)/\mu_1}\right].$$

Hence the Page detector is the just the CUSUM of the above nonlinearity clamped at 0, i.e.,

$$S_t = \max(0, S_{t-1} + g(y(t)).$$

REFERENCES

[1] Bansal, R., and Papantoni-Kazakos, P. (1986)
An algorithm for detecting a change in a stochastic process.
IEEE Transactions on Information Theory, **IT-32** (Mar. 1986), 227–235.

[2] Barrett, R., and Streit, R. (1989)
Automatic detection of frequency modulated spectral lines.
In *Proceedings of the Australian Symposium on Signal Processing, Appls.*, 1989, 283–287.

[3] Barrett, R., and Holdsworth, D. (1993)
Frequency tracking using hidden Markov models with amplitude and phase information.
IEEE Transactions on Signal Processing, **41** (Oct. 1993), 2965–2976.

[4] Basseville, M., and Nikiforov, L. (1993)
Detection of Abrupt Changes.
Englewood Cliffs, NJ: Prentice Hall, 1993.

[5] Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970)
A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains.
Annals of Mathematical Statistics, **41** (1970), 164–171.

[6] Chen, B., and Willett, P. (1997)
Quickest detection of hidden Markov models.
In *Proceedings of the IEEE CDC*, San Diego, Dec. 1997.

[7] Chen, B., Willett, P., and Streit, R. (1998)
A test of overdispersion in a data set with application to transient detection.
In *Proceedings of the CISS*, Princeton, NJ, Mar. 1998.

[8] Chen, B., and Willett, P. (1998)
Quickest detection of superimposed hidden Markov models using a multiple target tracker.
In *Proceedings of the IEEE Aerospace Conference*, Aspen, CO, 1998.

[9] Chen, B., and Willett, P.
Detection of superimposed hidden Markov model transient signals via multiple target tracking ideas.
Submitted to *IEEE Transactions on Aerospace and Electronic Systems.*

[10] Han, C., Willett, P., and Abraham, D. (1999)
Some methods to evaluate the performance of Page's test as used to detect transient signals.
IEEE Transactions on Signal Processing (Aug. 1999).

[11] Juang, B., and Rabiner, L. (1985)
A probabilistic distance measure for hidden Markov models.
AT&T Technical Journal, **64** (1985), 391–408.

[12] Lorden, G. (1971)
Procedures for reacting to a change in distribution.
Annals of Mathematical Statistics, **42** (June 1971), 1897–1908.

[13] Moon, T. (1996)
The expectation-maximization algorithm.
IEEE Signal Processing Magazine, **13** (Nov. 1996), 47–60.

[14] Moustakides, G. (1986)
Optimal stopping times for detecting changes in distributions.
Annals of Statistics, **14** (June 1986), 1379–1387.

[15] Nuttall, A. (1996)
Near-optimum detection performance of power-law processors for random signal of unknown location, structure, extent, and arbitrary strengths.
NUWC-NPT technical report 11123, Apr. 1996.

[16] Nuttall, A. (1997)
Detection capability of linear-and-power processors for random burst signals of unknown location.
NUWC-NPT technical report 10822, Aug. 1997.

[17] Page, E. (1954)
Continuous inspection schemes.
Biometrika, **41** (Jan. 1954), 100–115.

[18] Petrie, T. (1969)
Probabilistic functions of finite state Markov chains.
Annals of Mathematical Statistics, **40** (June 1969), 97–115.

[19] Phatarfod, R. (1965)
Sequential analysis of dependent observations.
Biometrika, **52** (1965), 157–165.

[20] Poor, H. (1994)
An Introduction to Signal Detection and Estimation (2nd ed.).
New York: Springer-Verlag, 1994.

[21] Rabiner, L., and Juang, B. (1986)
An introduction to hidden Markov models.
IEEE ASSP Magazine, **3** (Jan. 1986), 4–16.

[22] Ross, S. (1983)
Stochastic Processes.
New York: Wiley, 1983.

[23] Siegmund, D. (1995)
Sequential Analysis—Tests and Confidence Intervals.
New York: Springer-Verlag, 1995.

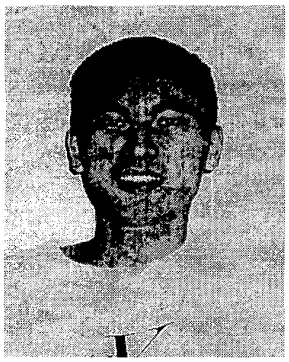
[24] Streit, R., and Barrett, R. (1990)
Frequency line tracking using hidden Markov models.
IEEE Transactions on Acoustics Speech and Signal Processing, **38** (Apr. 1990), 586–598.

[25] Streit, R., and Willett, P. (1999)
Detection of transient signals via hyperparameter estimation.
IEEE Transactions on Signal Processing (July 1999).

[26] Wald, A. (1947)
Sequential Analysis.
New York: Wiley, 1947.

[27] Wald, A., and Wolfowitz, J. (1948)
Optimum character of the sequential probability ratio test.
Annals of Mathematical Statistics, **19** (1948), 326–339.

[28] Wang, Z., and Willett, P. (2000)
A performance study of some transient detectors.
IEEE Transactions on Signal Processing, to be published.



Biao Chen received his B.E. in 1992, M.A. in 1994, both in electrical engineering, from Tsinghua University. After that, he worked at AT&T (China) Inc., Beijing, for about a year. He went to the University of Connecticut in 1995 where he obtained his Masters in statistics and Ph.D. in electrical engineering, in 1998 and 1999 respectively. After graduation, he spent a year at Cornell University as a Post-Doc Associate. He joined Syracuse University in 2000 as an Assistant Professor. His research area is in communications and signal processing, with particular interest in wireless communications.

Peter Willett (S'83—M'86) received the B.A.Sc. in 1982 from the University of Toronto, Toronto, Canada, and the Ph.D. from Princeton University, Princeton, NJ, in 1986.

He is an Associate Professor at the University of Connecticut, Storrs, where he has worked since 1986. His interests are generally in the areas of detection theory and signal processing, and, lately, particularly in the area of data fusion.