2009 American Control Conference
Hyatt Regency Riverfront, St. Louis, MO, USA
June 10-12, 2009

Detecting Intrusion Faults in Remotely Controlled Systems

Salvatore Candido and Seth Hutchinson

Abstract— **In this paper, we propose a method to detect an unauthorized control signal being sent to a remote-controlled system (deemed an "intrusion fault" or "intrusion") despite attempts to conceal the intrusion. We propose adding a random perturbation to the control signal and using signal detection techniques to determine the presence of that signal in observations of the system. Detection of these perturbations indicates that an authorized or "trusted" operator is in control of the system. We analyze a worst case scenario (in terms of detection of the intrusion), discuss construction of signal detectors, and demonstrate our method through a simple example of a point robot with dynamics.**

I. INTRODUCTION

Recently, there has been increased interest in networked control systems. While the possibility of controlling systems over established wireless, shared, or public networks has many benefits, it also requires that security join established performance characteristics such as performance, reliability, and efficiency during the design process. With this in mind, researchers have begun to address security issues specific to control systems, e.g. [1], [2], and have continued to develop control schemes to identify malfunctioning or malicious systems components, e.g. [3], [4], [5].

Typically protection of communication between an operator and a remote control system involves cryptography. Whether using encryption, message signing, or authentication, protection from a unauthorized user, an intruder, relies on a secret key unknown to intruders [6], [7]. If an intruder finds away to disrupt the communication channel and route an alternate control signal to replace the operator's signal, it is possible that the loss of control may go undetected for some length of time. If the intruder's goal is to conceal the intrusion and the operator's nominal control input is known or can be deduced in advance, the intruder can mimic that control input and the operator will be oblivious to intrusion. This could be desirable because the intruder may want to take control of the system as soon as possible, to ensure that the intrusion will be successful, but not utilize that control until a time when the operator has no recourse to stop the intruder's malicious actions.

If alerted quickly enough, the operator could possibly take action to preempt the malicious action of the intruder. However, typically no mechanism is in place to detect loss of control if the intruder is content to wait patiently for the opportune moment to act. Our goal is to automate the detection of such an intrusion.

To this end, we propose an augmentation to typical security systems that uses the constant observations of the system which are already present in many control applications. We propose to send a randomized secret signal through the control system, which will allow the operator to authenticate that his signal is reaching and controlling the system. Chosen properly, small perturbations in the system's input should produce subtle, yet detectable fluctuations in observations of the system. Since this signal is randomized, the intruder cannot replicate it from a priori knowledge. Thus, if it is detected during observation of the system, it provides good evidence that the operator's control signal is, in fact, driving the remote system. If these fluctuations cannot be found in the output, the operator is made aware that the system may be compromised and can begin taking measures to assure or regain control.

It is important to emphasize that our proposed method is intended to add an additional failsafe to an already secured system, not stand alone as a security scheme. It is meant for cases where the intruder is able to prevent the operator's control signal from reaching the plant and replace it with another signal. It is not meant for situations where the intruder is able to gain control of the the computer on board the remote system. The type of attacks we consider, while not comprehensive, are increasingly important because of the escalating amount of control communication being relayed over networks with intermediate routing points that may vulnerable to or owned by an intruder.

Our approach to this problem is, to the knowledge of the authors, novel but the mathematical tools used to implement the solution are similar to those used in a filtering based approach to active fault detection [8]. In active fault detection [9], [10], [11], [12], [13] and active parameter detection approaches [14], [15], a control signal is designed for the purpose of driving the system in such a way as to allow the operator to decide, based on output, which of a set of possible models best describes the system. In our case, the control input is chosen to be random so the intruder cannot reproduce it. We, therefore, design an optimal decision rule that decides if the sequence of observations is best explained by the random signal being present or not.

There is a significant conceptual resemblance between our approach and digital watermarking (e.g. [16]). We wish to embed a hidden signal in a sequence of observations of our system and detect its presence to verify the control signal being sent to the system is legitimate. However, there are a number of significant differences between this and the

S. Candido is in the Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801, USA candido@illinois.edu

S. Hutchinson is a Professor in the Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801, USA seth@illinois.edu

978-1-4244-4524-0/09/$25.00 ©2009 AACC

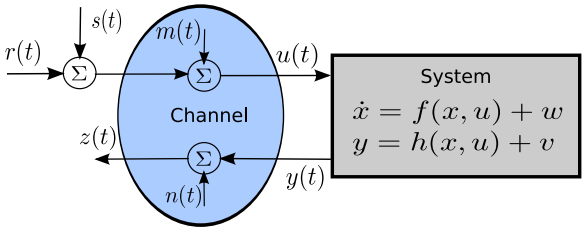

Fig. 1. Remote control system block diagram.

canonical watermarking problem. For example, the output of the system, the signal to be watermarked, is not known when the auxiliary signal is applied and there are no concerns about an intruder intentionally removing the watermark.

While enhancing the security of the system, this method has the downside of adding additional random noise to the system dynamics. However, in some cases, small perturbations in some system components can be negligible in the performance of the overall system. For example, a convoy of trucks moving hazardous materials could slightly vary their nominal speed or position within a lane of a road. Both of these quantities can easily be monitored externally from an observer watching the convoy from above. Candidate systems to employ this method should have stable basins of attraction around their nominal trajectory, be observed in a manner that cannot easily be compromised, and not attenuate high frequency signal components beyond detection.

The contribution of this paper is to propose a new method for verifying control signals sent over a link between an operator and control system by detecting when that link is compromised. We consider different types of systems and give conditions under which we can use an efficient, recursive decision rule. In cases where we can not, we discuss the theoretical difficulties and where numerical algorithms may be sufficient. We then present an example of our method on a simple hypothetical system. Finally, we discuss some shortcomings of our method and a number of questions that will need to be answered beyond our exploratory analysis.

II. GENERAL SYSTEMS

Consider the block diagram in Figure 1. We model remote systems using stochastic, nonlinear differential equations of the form

$$\dot{x}(t) = f\left(x(t), u(t)\right) + w(t) \quad (1)$$

$$y(t) = h\left(x(t), u(t)\right) + v(t) \quad (2)$$

where $w(t)$ and $v(t)$ are random vectors corresponding to noise in the state and output models, respectively. The vector $x(t)$ is the state of the system, $f(\cdot)$ is the system equation which encodes system dynamics, and $h(\cdot)$ is the output equation. The control $u(t)$ is the input to the remote system and $y(t)$ is the observation.

This remote system is controlled over a communication channel and the desired control signal is $r(t)$. A signal $\overline{s}(t)$ is randomly generated and added to $r(t)$ by the trusted operator and then sent to the remote system. This signal is the main addition to the standard remote-controlled system model and the core idea is to detect intrusions by noting the absence of its effects in the observation of the system's dynamics. Since $\overline{s}(t)$ is generated as needed, it is unknown to all parties except the trusted operator who records it as it is sent. This means that detection of this signal in observations of the system provides good evidence that the control signal sent over the channel has not been compromised by an intruder.

We model the signal received by the remote system as $u(t) = r(t) + s(t) + m(t)$ where $m(t)$ is a random vector corresponding to channel noise. We use the notation $s(t)$ to indicate the part of the signal received by the remote system. This signal will be different depending on whether the control signal originates from the trusted operator or an intruder. The signal $s(t)$ can either equal $\overline{s}(t)$, if the trusted operator's signal reaches the control system, or some other value, if an intruder has connected to the control system. In our analysis, we consider the case where the intruder sends $s(t) = 0$ to minimize the difference between the $u(t)$ with and without $\overline{s}(t)$. The operator observes $z(t) = y(t) + n(t)$ where $n(t)$ is also a random vector and again corresponds to noise inherent in retrieving an observation over a communication channel.

Several conditions, consistent with the goals of this method, must be met for this scheme to be practical. First, the operator's observations of the system should be taken externally to the system. This is important because if the observations are sent over the same communication channel that is controlled by the intruder, the operator's information state could be manipulated. Secondly, we assume the intruder cannot simply switch control between the operator and himself instantaneously. This assumption is reasonable as this method should be used in conjunction with other security measures that will be nontrivial for the intruder to break. If an intruder can break the security instantly and at will, then the motivation to conceal an intrusion is removed. Finally, the intruder must not be able to read the control signal being sent in real time, and consequently learn $\overline{s}(t)$ in real time. Although the intruder may have prior knowledge of $r(t)$, if he can simply detect $\overline{s}(t)$ as it is being transmitted he may be able to use that information to deceive the trusted operator. This condition can often be guaranteed by use of a provably secure encryption protocol [6] for securing the operator's control signal before transmission over the channel. In this case, even if the intruder was able to obtain the encrypted version of the control signal, he could not extract the contents in real time. Some protocols commonly used to encrypt data for transmission over an insecure network are RSA [17] and AES [18].

In this problem, we propose a strategy to detect an intruder attempting to conceal an intrusion by mimicking the control strategy of the operator. Thus, the worst-case scenario will be an intruder who knows the operator's nominal reference control signal (without $s(t)$) exactly and also the probability distribution from which $s(t)$ is drawn. We consider the case where the intruder sends $s(t) = 0$ because it is the MMSE estimator for an intruder trying to predict $\overline{s}(t)$. By comparing

predicted observations with and without random perturbation to actual observations of the system, the trusted operator can determine if the random perturbations added to the signal are influencing the control system. Unfortunately, due to the stochastic nature of the system, this determination cannot be made with complete certainty in the general case. This is then a signal detection problem [19] with the signal being sent over an unorthodox transmission channel.

In this paper, we consider the discrete problem in which observations of the system are given by the sequence $\{z(t)\}$. We will also treat $\{u_t\}$, $\{r_t\}$, and $\{\overline{s}_t\}$ as discrete sequences.We will use the notation $z_{i:j}$ to denote a history of the signal z_t between (and including) samples i and j, i.e. $z_{i:j}$ denotes $\{z_i, z_{i+1}, \cdots, z_j\}$.

The determination of whether $\overline{s}_t$ is present in u_t is made based the observations of z_t by the trusted operator. We frame our decision as a hypothesis testing problem. After observing a sequence of N samples of the output, the problem is to decide between two possible hypotheses of the transmitted signal:

1) H_0: $s_{1:N} = \overline{s}_{1:N}$, i.e. the trusted operator's control signal is received
2) H_1: $s_{1:N} = 0$, i.e. an intruder's signal is received

The decision is based on our observed output samples over that interval $z_{1:N}$. If we can say with certainty that $\overline{s}_t$ is affecting the system dynamics over a long enough time span, this provides good evidence that the trusted operator is controlling the system as no other party has a priori knowledge of $\overline{s}_t$, and the likelihood of an intruder replicating it from only knowledge of the distribution of $\overline{s}_t$ goes to zero as the number of samples increase.

There are two main issues to be overcome to implement this framework. First, we must choose a decision rule that, given $\overline{s}_{1:N}$, decides the hypothesis in a way that minimizes the probability of error. Secondly, in order to decide between these two hypotheses, we will need to compute the distributions $P_{Z_{1:N}|H_0}(z_{1:N}|H_0)$ and $P_{Z_{1:N}|H_1}(z_{1:N}|H_1)$, the probability density functions of $Z_{1:N}$ under both hypotheses. Thus, we must build a data representation that parametrizes or computes the pdf's. In cases where one cannot easily compute the distributions, an approximation may be sufficient.

The decision rule will separate all sample histories $z_{1:N}$ into two classes, $\mathcal{Z} \subseteq \mathbb{R}^N$ and $\mathcal{Z}^C = \mathbb{R}^N \backslash \mathcal{Z}$. Based on the boundary of the decision set $\mathcal{Z}$, a decision rule can be defined, $\delta : \mathbb{R}^N \rightarrow \{0, 1\}$. We accept hypothesis H_δ based on $z_{1:N}$. The decision rule and $\mathcal{Z}$ have the following relationship

$$\delta(z_{1:N}) = \begin{cases} 1 & z_{1:N} \in \mathcal{Z} \\ 0 & z_{1:N} \in \mathcal{Z}^C \end{cases} \tag{3}$$

In Sections II-A and II-B we show decision rules assuming we have an a priori distribution on the probability of intrusion and, for the case where we do not have this information, assuming the worst case.

A. A Priori Knowledge of Probability of Intrusion

In some cases, it is feasible to estimate the frequency of attack on the system. Assuming that we can estimate the probability of intrusion, we design a decision rule that minimizes p_e, the probability of erroneous decision. Let π denote the probability that $s_t = \overline{s}_t$, i.e. the trusted operator is in control of the system. With probability $1 - \pi$, an intruder will control the system. To determine the optimal $\mathcal{Z}$ with respect to p_e, we minimize

$$\begin{aligned} \min_{\mathcal{Z}} p_e(\mathcal{Z}, \pi) = \min_{\mathcal{Z}} \Bigg[& (1-\pi) \int_{\mathcal{Z}^C} p_{Z_{1:N}|H_1}(x|H_1) dx \qquad (4) \\ & + \pi \int_{\mathcal{Z}} p_{Z_{1:N}|H_0}(x|H_0) dx \Bigg] \\ = & (1-\pi) + \min_{\mathcal{Z}} \int_{\mathcal{Z}} \big(\pi p_{Z_{1:N}|H_0}(x|H_0) \qquad (5) \\ & - (1-\pi) p_{Z_{1:N}|H_1}(x|H_1) \big) \, dx \end{aligned}$$

Since $\pi,\ p_{Z_{1:N}|H_i}(x|H_i) \geq 0$, p_e is minimized by choosing $\mathcal{Z}$ to be the set of locations where the integrand in (5) is negative. Thus,

$$\mathcal{Z} = \left\{ x \in \mathbb{R}^N : \pi p_{Z_{1:N}|H_0}(x|H_0) \leq (1-\pi) p_{Z_{1:N}|H_1}(x|H_1) \right\}$$

and the optimal decision rule is to accept H_1 if

$$\frac{p_{Z_{1:N}|H_0}(z_{1:N}|H_0)}{p_{Z_{1:N}|H_1}(z_{1:N}|H_1)} \leq \frac{1-\pi}{\pi} \tag{6}$$

is satisfied. The probability of error p_e under this decision rule can be expressed as

$$p_e = \int_{\mathbb{R}^N} \min \left\{ \pi p_{Z_{1:N}|H_0}(x|H_0), (1-\pi) p_{Z_{1:N}|H_1}(x|H_1) \right\} dx \tag{7}$$

although, in general, this quantity may be expensive or impossible to compute exactly.

If z_t is independent of z_s for all $t \neq s$, then the likelihood of any sequence $z_{1:N}$ is just the product of the likelihoods of the elements of the sequence. In this case, we can factor the likelihood ratio and the decision rule will be to accept H_1 if

$$\prod_{i=1}^{N} \left[\frac{p_{Z_i|H_0}(z_i|H_0)}{p_{Z_i|H_1}(z_i|H_1)} \right] \leq \frac{1-\pi}{\pi} \tag{8}$$

This expression is, in general, computationally feasible, while (6) may be difficult or impossible to compute. If the likelihood ratio can be factored then the likelihood ratio for $t = N+1$ can be computed by multiplying the single sample likelihood at $t = N+1$ by the likelihood ratio for $t = N$. The optimality condition in (4) can also be modified to weight the cost of the total probability of a false positive and false negative with respect to one another. However, the quantity minimized would no longer be the probability of error.

B. No Prior Knowledge of Probability of Intrusion

In cases where it is not feasible to estimate a probability of attack, the method of the previous section may not be a practical solution. If we cannot determine π a priori and prefer not to treat it as an arbitrary scaling of the decision

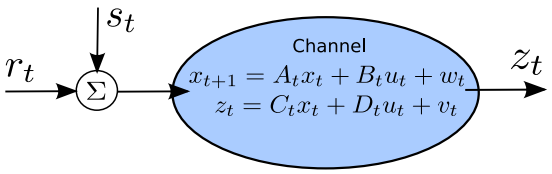

Fig. 2. Remote controlled stochastic, linear system

rule, we may choose to implement the minimax decision rule. The cost function is $\min_{\mathcal{Z}} \max_{\pi} p_e(\mathcal{Z}, \pi)$.

$$p_e(\mathcal{Z},\pi) = \left[\pi\left(\int_{\mathcal{Z}} p_{Z_{1:N}|H_0}(x)dx - \int_{\mathcal{Z}^C} p_{Z_{1:N}|H_1}(x)dx\right) + \int_{\mathcal{Z}^C} p_{Z_{1:N}|H_1}(x)dx\right] \quad (9)$$

Let $\mathcal{Z}^*(\pi)$ be the optimal $\mathcal{Z}$ (minimizes p_e) given in the previous section for a fixed π. The function $\max_{\pi} p_e(\mathcal{Z}^*(\pi), \pi)$ over $\pi \in [0,1]$ is continuous and concave [19]. From (9) we can see that $p_e(\mathcal{Z}, \pi)$ is linear in π given $\mathcal{Z}$ so, unless $\mathcal{Z}$ is chosen so that $p_e(\mathcal{Z}, \pi)$ is independent of the choice π, the π that maximizes $p_e(\mathcal{Z}, \pi)$ will be either zero or one. At the boundaries of the interval, $p_e(\mathcal{Z}^*(0), 0) = p_e(\mathcal{Z}^*(1), 1) = 0$ so the minimum average error must be strictly zero or on the interior of the interval. Using Prop. II.C.1 from [19], $\mathcal{Z}$ is a minimax rule if $\pi \in (0,1)$ and

$$\int_{\mathcal{Z}} p_{Z_{1:N}|H_0}(x|H_0)dx = \int_{\mathcal{Z}^C} p_{Z_{1:N}|H_1}(x|H_1)dx \quad (10)$$

For an arbitrary distribution, finding the set $\mathcal{Z}^*$ that satisfies (10) analytically may not be possible.

To find an approximate $\mathcal{Z}^*$ numerically, one can search $[0,1]$ for the optimal π^*. This is done by choosing π_i, the value to check at step i of the algorithm and approximating $\mathcal{Z}^*(\pi_i)$. Once the optimal decision set for π_i has been chosen by the criteria of Section II-A, compute $p_e(\mathcal{Z}^*(\pi_i), \pi_i)$ by performing a numerical integration. Based on the history of the π_i's chosen and the probabilities of error computed at those samples, one can choose π_{i+1} to reduce the interval in which π^* may lie. If we reach arbitrarily close to π^*, check to ensure it is a minimax solution by testing to see if (10) is satisfied or if π^* is zero or one. If these conditions are not met, a non-randomized minimax decision rule may not exist.

In general, the performance of this decision rule will not be as good as the one of the previous section but removes the requirement for an estimate of the probability of intrusion.

III. Linear Systems with Gaussian Noise

As shown in the previous section, if the likelihood ratio can be factored, we can easily compute recursive decision rules. In this section, we discuss the special case of linear systems with Gaussian noise. A wide range of physical systems can be modeled or approximated in this class of systems. We first briefly consider the special case of a memoryless system, where no state estimator is required for the decision rule, and then the more general case.

Consider the special case of a memoryless channel with Gaussian, white noise. Since estimates of Z_t do not depend on previous values of z_t, the history of the output does not need to be parametrized to compute the pdf of Z_t. Let the system equation be $z_t = D_t u_t + v_t$ and, without loss of generality, consider v_t to be both the noise from the channel and within the remote controlled system. The channel noise is $v_t \sim \mathcal{N}(0, \Sigma_{v_t})$ and $\{v_t\}$ is independent and identically distributed (iid). Since Z_t has a Gaussian distribution, we can specify its distribution with the parameters $\mathrm{E}Z_t = D_t(s_t + r_t)$ and $\mathrm{Var}(Z_t) = \Sigma_{v_t}$. The random variable Z_t will be drawn from the probability distribution $P_{Z_t|H_0}(z_t|H_0) = \mathcal{N}(D_t(\bar{s}_t + r_t), \Sigma_{v_t})$ if $s_t = \bar{s}_t$. If $s_t = 0$ then Z_t will be drawn from $P_{Z_t|H_0}(z_t|H_0) = \mathcal{N}(D_t r_t, \Sigma_{v_t})$. Since the noise between samples is independent, the likelihood of any sequence $z_{1:N}$ is just the product of the likelihoods of the elements of the sequence between so the optimal decision rule is to use the criterion of (8).

In the more general case, we consider a discrete stochastic, linear system with additive, zero-mean Gaussian noise is modeled by the standard equations

$$x_{t+1} = A_t x_t + B_t u_t + w_t \quad (11)$$
$$z_t = C_t x_t + D_t u_t + v_t \quad (12)$$

where $w_t \sim \mathcal{N}(0, \Sigma_{w_t})$ and $v_t \sim \mathcal{N}(0, \Sigma_{v_t})$ as shown in Figure 2. Again, the random variables w_t and v_t are due to both noise within the system and on the communication channel.

Once a state is added to the system, two difficulties arise. First, $\mathrm{Prob}\{Z_t = c\} \neq \mathrm{Prob}\{Z_t = c \mid z_{1:t-1}\}$ which means we lose information by computing the distribution of Z_t without taking into account previous observations. Second, since Z_{t+1} depends on the noise at t as well as $t+1$, Z_t and Z_s are no longer independent for $t \neq s$. This means the likelihood of $z_{1:N}$ cannot necessarily be factored into the products of the likelihoods of the individual observations. Without this property, the pdf of $Z_{1:N}$ could be a complicated function that is possibly infeasible to compute or represent. We deal with the first problem by using a Kalman filter [20] to recursively compute an estimate of X_t, the system's state vector, conditioned on $z_{1:t-1}$. The second problem we will bypass by testing the innovation error between the observation and prediction of Z_t. This quantity is Gaussian and uncorrelated with, thus independent of, previous innovation errors. We will compute the pdf's of sequences of innovation errors under H_0 and H_1 and use the likelihoods of the observed $z_{1:t-1}$ to decide intrusion.

Define $\tilde{z}_t := Z_t - (C_t \mathrm{E}[X_t|z_{1:t-1}] + D_t u_t)$ to be the innovation error of the observation process. It is a linear combination of the random variables Z_t and $\mathrm{E}[X_t|z_{1:t-1}]$ and the output of the system is a linear combination of the random variables X_0, $\{W_t\}$, and V_t. The system model requires these random variables to be independent so they are jointly Gaussian (jG). Since Z_t and $\mathrm{E}[X_t|z_{1:t-1}]$ are both linear combinations of these jG random variables, they themselves are jG and, since $\tilde{z}_t$ is a linear combination of Z_t and $\mathrm{E}[X_t|z_{1:t-1}]$, the innovation error is jG.

The parameters of the distribution $\tilde{z}_t$ conditioned on $Z_{1:t-1}$ are $\mathrm{E}[\tilde{z}_t|Z_{1:t-1}] = 0$ and $\mathrm{Cov}(\tilde{z}_t|Z_{1:t-1}) = C_t\mathrm{Cov}(X_t - \hat{x}_t|z_{1:t-1})C_t^T + \Sigma_{v_t}$. Since $\tilde{z}_t$ is jG, this completely specifies the distribution $\tilde{z}_t \sim \mathcal{N}(0, C_t\mathrm{Cov}(X_t - \hat{x}_t|Z_{1:t-1})C_t^T + \Sigma_{v_t})$.

Now that we have determined the distribution of $\tilde{z}_t$, in order to use it at every stage of detection we will use the Kalman filter to store and propagate the parameters of the distribution as new observations from the system become available [20]. Define $\hat{x}_t := \mathrm{E}[X_t|Z_{1:t-1}]$ and $\Sigma_{\hat{x}_t} := \mathrm{Var}(X_t - \hat{x}_t|Z_{1:t-1})$. The Kalman filter updates these two quantities recursively from initial conditions $\hat{x}_0$ and $\Sigma_{\hat{x}_0}$ based on the uncertainty the initial state. We will only mention that the Kalman filter is the optimal MMSE estimate for linear systems with Gaussian noise. For more information on the Kalman filter, see [20], [21].

Our strategy is to use two Kalman filters to estimate the state of the system. Both are given $\{z_s\}_{s=0}^{t-1}$ and one uses $u_t = r_t + \overline{s}_t$ as the input, the other uses $u_t = r_t$ as the input. As described above, we use these estimators to construct the parametrization of the pdf of $\tilde{z}_t$, the innovation error between the next observation and our prediction of that observation. We will test the likelihood of $\tilde{z}_t$ using both hypotheses and we will denote the mean and covariance of $\tilde{z}_t$ working under hypothesis H_i as $\tilde{z}_{t|H_i}$ and $\Sigma_{\tilde{z}_{t|H_i}}$.

We know from the orthogonality principle [22] that $\mathrm{E}[\tilde{z}_t\tilde{z}_s] = 0$ for all $t \neq s$ because the innovation error of the MMSE estimator is orthogonal to any function of the elements of $Z_{1:t-1}$, which includes $\tilde{z}_s$ (assuming without loss of generality that $t > s$). Since $\tilde{z}_t$ and $\tilde{z}_s$ are uncorrelated and jG, they are independent. As the distribution of each $\tilde{z}_t$ is jG, the distribution of $\tilde{z}_{1:N}$ will also be Gaussian. Since all the innovations are zero mean and independent of one another, the mean of $\tilde{z}_{1:N}$ will be the zero vector of length N and the covariance matrix will be block diagonal with the i^{th} block equaling $C_t\Sigma_{\hat{x}_t}C_t^T + \Sigma_{v_t}$.

We will use the decision rules of Section II to decide intrusion using the likelihoods from the two pdf's corresponding to H_0 and H_1. The optimal decision rule when there is an a priori estimate of intrusion is analogous to (8), only using innovation errors.

$$\prod_{i=1}^{N}\left(\frac{p_{\tilde{z}_i|H_0}(\tilde{z}_i|H_0)}{p_{\tilde{z}_i|H_1}(\tilde{z}_i|H_1)}\right) \leq \frac{1-\pi}{\pi} \tag{13}$$

In the case of no estimate of π, the decision boundary can be found analytically in the case of the pdf's being Gaussian. By varying π_i and using numerical approximations to the Gaussian cdf, one can find the decision boundary that approximately satisfies (10), the condition for a minimax solution.

IV. EXAMPLE

We now present an example of our framework. Consider a point robot with mass m and second order dynamics, moving on the real number line. The operator's control input is a scalar specifying the desired position of the robot to a PD controller. A specific trajectory for the robot r_t is planned in advance by the trusted operator. At $t = 0$, an intruder may have taken over the system. If an intrusion has occurred, the goal of the intruder will be to avoid detection for as long as possible. The intruder has full knowledge of r_t, the system model, and the distribution of $\overline{s}_t$. The trusted operator's goal is to detect an intrusion without deviating (more than an additive random Gaussian signal) away from the preplanned r_t. We will assume the noise in the process model is Gaussian having zero mean and a diagonal covariance matrix with σ_i, σ_j, and σ_k as the diagonal terms. The observation shows robot's position with additive Gaussian noise distributed according to $\mathcal{N}(0,\ \sigma_o)$. There is also channel noise with distributions $\mathcal{N}(0, \sigma_m)$ and $\mathcal{N}(0, \sigma_n)$. This means our full system model is

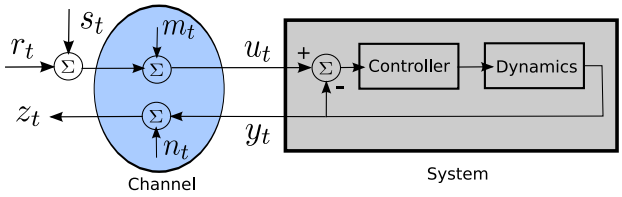

Fig. 3. Example system block diagram.

$$x_{t+1} = \begin{bmatrix} 1 & \Delta t & 0 \\ 0 & 1 & \Delta t \\ -\frac{k_p}{m} & -\frac{k_d}{m} & 0 \end{bmatrix} x_t + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u_t + w_t \tag{14}$$

$$z_t = \begin{bmatrix} \frac{k_p}{m} & \frac{k_d}{m} & 0 \end{bmatrix} x_t + v_t \tag{15}$$

where both w_t and v_t are mean zero and Gaussian.The matrix Σ_w is diagonal with diagonal entries σ_i, σ_j, and $\sigma_m + \sigma_k$ and $\sigma_v = \sigma_o + \sigma_n$. We have the option of choosing s_t, so let it be another iid, Gaussian random variable with zero mean and variance of σ_s. This is the random perturbation of which only the operator has knowledge, as it is generated online.

We will track the state trajectory of the system with two Kalman filters. We use the non-identity matrix coefficients of (14)-(15) along with the variances of v_t and w_t in the Kalman filters to estimate state. The two Kalman filters differ only in the signal u_t given to them. The first, simulating H_0, will receive $u_t = r_t + \overline{s}_t$. The other, simulating H_1, will receive $u_t = r_t$. Then the likelihood of an innovation error of $\tilde{z} = z_t - (C\hat{x}_{t|H_i} + Du_t)$ under hypothesis H_i for the sample at time t is

$$\mathcal{L}\{\tilde{z}_t|H_i\} = \left[2\pi\left(C\Sigma_{\hat{x}_{t|H_i}}C^T + \sigma_v\right)\right]^{-\frac{1}{2}} \exp\left\{-\frac{\left[z_t - (C\hat{x}_{t|H_i} + Du_t)\right]^2}{2C\Sigma_{\hat{x}_{t|H_i}}C^T + \sigma_v}\right\}. \tag{16}$$

Since this is a linear system and x_0, $\{w_t\}$, and $\{v_t\}$ are jG, we know the likelihood ratio from (8) is our optimal decision rule. Thus, we can compute the likelihood ratio recursively for every t using

$$\frac{\mathcal{L}\{\tilde{z}_{1:t}|H_0\}}{\mathcal{L}\{\tilde{z}_{1:t}|H_1\}} = \prod_{s=1}^{t}\frac{\mathcal{L}\{\tilde{z}_s|H_0\}}{\mathcal{L}\{\tilde{z}_s|H_1\}} = \frac{\mathcal{L}\{\tilde{z}_{1:t-1}|H_0\}}{\mathcal{L}\{\tilde{z}_{1:t-1}|H_1\}}\cdot\frac{\mathcal{L}\{\tilde{z}_t|H_0\}}{\mathcal{L}\{\tilde{z}_t|H_1\}}$$

starting from the initial condition where the likelihood ratio is one. If we have an a priori belief in the probability of intrusion, π, our decision rule will be

$$\frac{\mathcal{L}\left\{\tilde{z}_{1:t}|H_0\right\}}{\mathcal{L}\left\{\tilde{z}_{1:t}|H_1\right\}} \leq \frac{1-\pi}{\pi} \tag{17}$$

If this inequality is satisfied, it signals the trusted operator that an intrusion is likely to have occurred.

V. DISCUSSION

The work presented in this paper is a proposed method, and significant exploration remains before a system utilizing the principles discussed in this paper is deployed. This framework is not designed to provide foolproof, complete security for remotely controlled systems. Instead, we seek to augment current security measures by verifying the control signals sent to remote system in environments where an intrusion may go undetected due to a clever choice of control signals by the intruder.

One cause for concern with this framework is that if the remotely controlled system is, for example, a low-pass filter, a high frequency $\overline{s}_t$ may not cause any significant effects in z_t. Work will have to be done to quantify how large the magnitude of the random signal will need to be in order to be detected in sufficiently few samples.

Conversely, an $\overline{s}_t$ with more pronounced effects on the output will be easier to detect in the system output. However, it will also, by definition, perturb the remote system further away from the nominal trajectory. As the remote system moves along it's trajectory, it may be possible to determine criteria by which the distribution of $\overline{s}_t$ could be adjusted automatically based on the stability margin of the system. For example, perhaps $\overline{s}_t$ could be chosen to have a large effect in certain very stable configurations of the system but when the system is in a configuration where it could be easily pushed to instability, $\overline{s}_t$ could be attenuated or set to zero temporarily. It may also be possible to isolate the effects of $\overline{s}_t$ on the overall trajectory of the system if it is redundant.

We mainly discussed the decision rule that used an a priori estimate of the probability of intrusion π. However, in practice, this will typically not be feasible to estimate, and using a minimax decision rule may not be desirable. However, since π simply determines the scaling factor of the likelihood ratio, it may be chosen experimentally. Another possible scenario if the operator is controlling many remote systems is to compute the likelihood ratios for each of the systems and attempt to determine if there is an outlier ratio, which would signal a possible intrusion.

For the purpose of this paper, we assumed that the intruder will sent $u_t = r_t$ and not introduce an additional perturbation. Any additional perturbation would increase the expected mean square error between the operator's and intruder's control signals and it seems would make intrusion easier for the operator to detect. However, it is possible that there is a perturbation that would be advantageous to the intruder. While more exploration must be done to determine whether sending the MMSE of s_t is an optimal intruder strategy, we feel that the preliminary analysis performed in this paper demonstrates the utility and the need for further exploration of this intrusion detection strategy.

VI. CONCLUSIONS

In summary, we have proposed a method to detect intrusions in remote-controlled systems. We have discussed the theory and implementation of this method and have shown a theoretical example of it on a simple robot system. More exploration must be done before this system is ready for deployment but we are optimistic about the possibilities of intrusion detection using this method.

REFERENCES

[1] G. Baliga, S. Graham, C. Gunter, and P. Kumar, "Reducing risk by managing software related failures in networked control systems," in *IEEE Conference on Decision and Control*, 2006, pp. 2866–2871.

[2] A. Saboori and C. Hadjicostis, "Notions of security and opacity in discrete event systems," in *CDC*, 2007, pp. 5056–5061.

[3] M. Franceschelli, M. Egerstedt, and A. Giua, "Motion probes for fault detection and recovery in networked control systems," in *American Control Conference*, 2008, pp. 4358–4363.

[4] W. Chen and M. Saif, "Output estimator based fault detection for a class of nonlinear systems with unknown inputs," in *American Control Conference*, 2008, pp. 3307–3312.

[5] R. Isermann, *Fault-Diagnosis Systems: An Introduction From Fault Detection To Fault Tolerance*. Springer, 2006.

[6] O. Goldreich, *Foundations of Cryptography*. Cambridge University Press, 2001.

[7] M. Benantar, *Access Control Systems: Security, Identity Management and Trust Models*. Springer, 2006.

[8] X. Zhang, *Auxiliary signal design in fault detection and diagnosis*. Springer Verlag, 1989.

[9] M. Simandl and I. Puncochar, "Unified solution of optimal active fault detection and optimal control," in *American Control Conference*, 2007, pp. 3222–3227.

[10] F. Kerestecioglu and M. Zarrop, "Input design for detection of abrupt changes in dynamical systems," *International Journal of Control*, vol. 59, no. 4, pp. 1063–1084, 1994.

[11] R. Nikoukhah, "Guaranteed active failure detection and isolation for linear dynamical systems," *Automatica*, vol. 34, no. 11, pp. 1345–1358, 1998.

[12] E. Courses and T. Surveys, "Stochastic change detection based on an active fault diagnosis approach," in *IEEE Conference on Decision and Control*, 2007, pp. 346–351.

[13] I. Andjelkovic, K. Sweetingham, and S. Campbell, "Active fault detection in nonlinear systems using auxiliary signals," in *American Control Conference*, 2008, pp. 2142–2147.

[14] R. Mehra, "Optimal input signals for parameter estimation in dynamic systems–survey and new results," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 753–768, 1974.

[15] M. Zarrop, *Optimal experiment design for dynamic system identification*. Springer Verlag, 1979.

[16] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, *Digital Watermarking and Steganography*. Morgan Kaufmann, 2007.

[17] R. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," *Communications of the ACM*, vol. 21, no. 2, pp. 120–126, 1978.

[18] J. Daemen and V. Rijmen, *The Design of Rijndael*. Springer-Verlag, 2002.

[19] H. V. Poor, *An Introduction to Signal Detection and Estimation*. Springer-Verlag, 1994.

[20] R. Stengel, *Optimal Control and Estimation*. Courier Dover Publications, 1994.

[21] I. Rhodes, "A tutorial introduction to estimation and filtering," *IEEE Transactions on Automatic Control*, vol. 16, no. 6, pp. 688–706, 1971.

[22] H. Stark and J. Woods, *Probability, Random Processes, and Estimation Theory for Engineers*. Prentice-Hall, 1986.