Modeling and detecting localized nonlinearity in continuum systems with a multistage transform

Paul H. Bryant[*]

BioCircuits Institute (formerly Institute for Nonlinear Science), University of California, San Diego, La Jolla, California 92093, USA

J. M. Nichols[†]

Naval Research Laboratory, Code 5673, Washington, DC 20375, USA

(Received 19 August 2009; revised manuscript received 24 November 2009; published 19 February 2010)

A general method is presented for modeling spatially extended systems that may contain a localized source of nonlinearity. It has direct applications to structural health monitoring (SHM) where physical damage may cause such nonlinearity and also communications channels which may exhibit localized nonlinearity due to bad electrical contacts or component nonlinearity. The method uses a multistage nonlinear transform in order to model the system dynamics. We discuss the application to SHM and provide a preliminary test of the method with experimental data from a randomly shaken beam with loose bolts. We discuss the application to telecommunications, provide an experimental observation of symmetric nonlinearity in a "bad" electrical contact, and provide a preliminary test of using this method to remove nonlinear echo (and thereby improve data rate) on a telephone line used for data transmission.

DOI: 10.1103/PhysRevE.81.026209 PACS number(s): 05.45.Tp, 02.50.Fz, 84.40.Ua

I. INTRODUCTION

A. Purpose and potential applications

The modeling and analysis of a nonlinear system possessing memory is, in general, an extremely challenging task. In this work, we present a technique for modeling spatially extended dynamical systems that are primarily linear but may contain a localized nonlinear region. Modeling the response of these types of systems is of importance to a number of fields of research. In telecommunications, for example, a localized nonlinearity in a transmission line could be caused by bad electrical contacts or by the (slight) nonlinearity of electronic components. In this case the generated nonlinear model itself is of critical importance as it might be used not just to detect nonlinearity but also to correct signals for the resulting nonlinear distortion or nonlinear echo, leading to an improvement in data rate. In structural health monitoring (SHM), localized physical damage to a structure, such as a ship's hull, is often modeled as a localized nonlinearity [1,2]. Forecasting the evolution of the damage is predicated on one's ability to reliably detect and characterize the nonlinearity. In aeroelasticity, localized free-play nonlinearities can cause instabilities that result in degraded performance or even system failure [3]. There are likely to be other applications as well. In addition to presenting the theory, we give experimental results for a simple structural test system, a shaken beam with loose connecting bolts. The method is demonstrated to be a reliable detector of bolt loosening. Following the SHM results, more details are provided on how the method can be applied to improve the data rate of telecommunication systems, and some experimental results are given demonstrating the removal of nonlinear echo in a transmission line. We also provide an experimental observation of symmetric nonlinearity in a bad electrical contact and briefly discuss the physics involved. Based on these example systems, it should be straightforward to apply the method to other cases that may be of interest to the reader.

B. Background for the application to damage detection

Structural systems are often subjected to an input (driving) sequence described by a random process, i.e., are subject to random vibrations. SHM is a field of study that seeks to measure these response vibrations and use them to infer something about the "health" of the structure (e.g., damaged vs undamaged). Because structural damage will often manifest as a nonlinearity, any approach that can detect nonlinearity using response vibration data is an excellent candidate for damage detection. Researchers in the SHM field have increasingly relied on the linear or nonlinear distinction in designing damage detection strategy [2,4–6].

Most nonlinearity detection strategies focus on the fundamentally different ways in which linear and nonlinear systems act on an input. For nonlinear systems, a stationary jointly Gaussian input will typically result in a highly non-Gaussian response, thus any proposed modeling technique must be capable of capturing this effect. Linear systems, on the other hand, are unable to generate statistical moments in the response that are not already present in the input. They will, however, frequently introduce "memory" or autocorrelation to an input signal. These basic properties of linear and nonlinear systems are the basis for much of the literature in "nonlinearity detection" when analyzing random processes. The surrogate data method, for example, tests whether the data are consistent with the response of a linear system to a stationary input characterized by an autocorrelation function and a possibly non-Gaussian marginal distribution [7–10]. All other detected statistical properties are assumed to result from nonlinearity acting on the input and their presence is therefore an indicator of nonlinearity.

C. Outline of the method (for all applications)

Rather than focus on detecting deviations from linear system behavior, this work is aimed at creating a generic non-

[*] pbryant@ucsd.edu

[†] jonathan.nichols@nrl.navy.mil

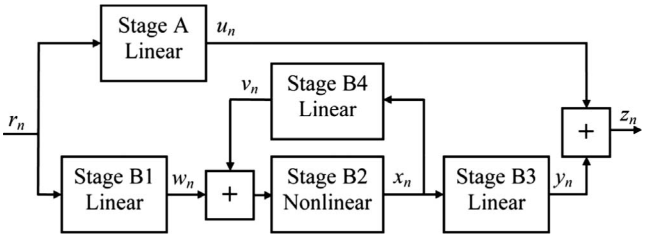

FIG. 1. Multistage map consisting of a linear map A in parallel with a nonlinear map B, which is itself made up of four stages. Inputs and outputs are assumed to be scalar time series except for r_n, which could optionally have more than one dimension, possibly sampled at more than one location. The linear stages have memory and therefore depend on previous time steps, while the nonlinear stage, B2, is either memoryless or has very limited memory. The combination of A and B is a model for a localized nonlinearity in a spatially extended linear system. The multistage map is an approximation to a multistage TD transform [11].

linear model of the system response. The model we propose for this type of system takes the form of a multistage time-domain (TD) transform [11] approximated as a multistage map (MSM), configured as shown in Fig. 1. The linear stages are "memoried," i.e., they have dependence on previous time steps, while the nonlinear stage has little or no memory. Unless stated otherwise, we will assume in the following analysis that the nonlinear stage is memoryless and one dimensional. In some cases, the "self-interaction" or "feedback" stage B4 may be small and therefore omitted from the model, at least in first approximation. Note that these maps are just discrete time approximations to linear and nonlinear TD transforms and the theory could just as easily be presented in that form. The discrete form, however, is what is needed in practice since data from a real physical system will be in the form of time series sampled at some finite time step. As will be discussed in more detail below, we typically choose to represent the linear stages in the form of a finite impulse response (FIR) filter, and the nonlinear stage as a simple power series.

II. THEORY (FOR ALL APPLICATIONS)

A. Derivation of the model

We will now show how the model of Fig. 1 can be derived if we know in precise detail all characteristics of the system. The purpose of this exercise is to show that the system can be represented by a model with the indicated structure. In practice we will not need to know these characteristics but rather we will start with the general model and use an optimization procedure to fit it to data from a particular system.

We will assume that the nonlinear element has some characteristic, a nonlinear function, that typically relates an applied "force" $F(t)$ to a "response" $R(t)$. The response, for example, could be the (reversible) bending of metal in the vicinity of a crack, or the amount of current flowing through a bad electrical contact. We will also assume that this element is embedded in an extended system that is otherwise linear and that the system is driven by one or more, known or measured, input signals. We will also assume that we are monitoring the system dynamics at some location to produce what we call the target signal.

To obtain the model, the first step is to freeze the nonlinear element, i.e., set $R(t)=0$ for all forcing values. In the case of the bad contact, this would mean breaking the connection at that point so that no current would flow. The system would then be entirely linear, and if we have complete knowledge of the system we should be able to determine the exact expressions for two particular linear TD transforms [11]. The first transform generates the signal arriving at the target site as a function of the input signal(s). This transform is stage A in the model. The second generates the applied force that appears at the nonlinear site as a function of the input signal(s). This transform is stage B1 in the model. It is sufficient for our purposes to know that these linear transforms exist.

The second step is to set all inputs to zero amplitude and also to force the response of the nonlinear element to follow any function of time $R(t)$ that we wish [regardless of the force $F(t)$ actually applied to it]. As before, we have a driven linear system but with the drive signal now coming from the site of the nonlinear element. Again we can calculate two linear transforms of the drive. The first again generates the signal arriving at the target site. This transform is stage B3 in the model. The second is a self-interaction, which generates the applied forcing $F(t)$ at the nonlinear site. This can be broken into two components. The first is an "immediate" self-interaction that is directly proportional to the drive signal. The second component is a "delayed" self-interaction, which in some cases might be thought of as the response signal traveling (as a wave) away from the site of the nonlinear element and then being reflected back to it by some interaction with the linear environment. This delayed self-interaction is the feedback stage B4.

What remains is to understand how the nonlinear element is modeled by stage B2. We assume that we explicitly know how the net forcing F_N is related to the response R as some function $g(\,)$, i.e., $F_N=g(R)$. In responding to an applied force F there may be an additional force from the immediate self-interaction which, since it is linear, can be expressed as λR, where λ is some constant. (We will show this explicitly when discussing the telecommunications application.) Thus the applied force is related to the response by the function $h(\,)$ defined as

$$F = h(R) = g(R) - \lambda R. \tag{1}$$

Letting $h^{-1}(\,)$ represent the inverse of the function $h(\,)$, we can express the response as a function of the applied force:

$$R = h^{-1}(F). \tag{2}$$

The function $h^{-1}(\,)$ defines the nonlinear stage B2.

In our analysis, we have treated the nonlinear element as though it was an active component, i.e., as though it monitors the input forcing $F(t)$ and then uses a nonlinear function to calculate an output $R(t)$ with which to drive the system. But, there is no difference in effect between the actual nonlinear element and this imagined active version. The analysis is also valid for a true active element, one which might be capable of generating interesting dynamics including chaos.

B. Applying the method

In order to utilize this method, it is necessary to have available at least two time series data sets sampled simultaneously at different locations on the system of interest. These could be the measurements of some physical property or properties, such as the strain, acceleration, pressure, voltage, etc. Optionally some of these time series could represent the undistorted ambient or applied vibrations. The method involves the generation of both linear and nonlinear TD transforms (approximated as memoried maps), using one (or possibly more) time series, r_n, as input(s) and optimizing these mappings to minimize the difference between the map output and a measured time series, s_n, which we will call the target. An error measure M is needed to quantify the fit of the mapping output z_n to the target series. The mapping will be a function of a large number of parameters (to be described shortly) and can therefore be written $z_n(\vec{\theta})$. Defining $q_n(\vec{\theta}) = z_n(\vec{\theta}) - s_n$, our typical choice of error measure (or cost function) is the error variance [12]:

$$M(\vec{\theta}) = \frac{1}{N-1}\left[\sum_{k=0}^{N-1} q_{k+k_\mathrm{S}}(\vec{\theta})^2 - \frac{1}{N}\left(\sum_{k=0}^{N-1} q_{k+k_\mathrm{S}}(\vec{\theta})\right)^2\right], \quad (3)$$

where k_S is the starting index and N is the number of points to be included. Under the assumption that the corrupting noise on the measurements is independent and identically distributed (iid) Gaussian, minimization of Eq. (3) gives a maximum likelihood estimate of the mapping parameters $\vec{\theta}$. The optimization problem to be solved can be stated

$$\min_{\vec{\theta}} M(\vec{\theta}). \quad (4)$$

Note that $M(\vec{\theta})$ could be expressed as a mean square error rather than an error variance, but use of the variance will, in many cases, eliminate the need for constant coefficients in the mapping stages. Also in many cases, the data can be "demeaned" provided that the mean value is not subject to change since such change could lead to a rebiasing of the nonlinear element. If desired, an appropriate constant parameter can be added to the combined map output after optimization is completed so that the mean of the output will match the mean value of the target. The available data may be broken up into training data and testing data, with different starting indices.

When used to detect nonlinearity, there must be a way to compare accuracy of the fit of the linear mapping with that of the nonlinear mapping. Our preferred form for the linear mapping (map A) is that of a FIR filter:

$$u_n = \sum_{i=1}^{N_\mathrm{I}} \sum_{k=0}^{D_{\mathrm{A}i}-1} a_{i,k} r_{i,n-k+k_{\mathrm{A}i}}, \quad (5)$$

where N_I is the number of input data sets being used (the sum over i may be omitted if $N_\mathrm{I}=1$), $a_{i,k}$ is set of adjustable coefficients, and $k_{\mathrm{A}i}$ is an offset for the index of the input time series $r_{i,n}$. The number of terms in the sum over k is often called the number of taps of the filter, which in this case is $D_{\mathrm{A}i}$. We want to optimize the values of the coefficients to generate a map which minimizes our error measure $M(\vec{\theta})$ (with $z_n = u_n$ when optimizing the linear mapping by itself, i.e., with map B turned off). The resulting map possesses memory in the sense that each output value depends on more than one input time step. If any of the offsets $k_{\mathrm{A}i}$ are positive, the mapping will be acausal, which may be acceptable if $r_{i,n}$ is not a true input to the system (as discussed below). When optimizing just over the coefficients $a_{i,k}$, the problem has an exact solution and can be reduced to a set of n linear equations in n unknowns where n is the total number of filter taps (i.e., $n = \Sigma_i^{N_\mathrm{I}} D_{\mathrm{A}i}$). Thus we can be sure that there is no linear mapping (for equivalent inputs) that can do better.

For the application to damage detection, the essence of the method is to run a nonlinear memoried mapping in parallel with the linear mapping so determined, which acts on the same set or a subset of these index values, and see if we can achieve any improvement in the error measure. The combined output of the linear and nonlinear maps is now used to calculate the error measure. If we can achieve any significant improvement through the addition of the nonlinear mapping, we then have *strong evidence* of a nonlinear process in the structure which, for this application, is indicative of structural damage.

Note that the nonlinear mapping (map B) could be implemented as a discrete Volterra series expansion, but the number of adjustable coefficients in such an expansion can in some cases be *extremely* large (see Sec. III). Instead we may choose to use the multistage map of Fig. 1, which results in a much smaller number of coefficients, and is much faster to implement.

Consider the case where the (one or more) input data sets are just measurements at certain locations and *not true* input signals. This may be of importance in the SHM application, where it may be difficult to measure the input signals directly. If these locations are relatively far from the site of the nonlinear element, then one might argue that the *true* inputs are approximately some linear transform of these *proxy* inputs. These transforms could therefore be incorporated into the transforms of stages A and B1 and the resulting MSM model would use these proxy input signals.

As will be demonstrated in a structural test system, the above-described model is able to adequately capture and quantify the influence of the nonlinearity acting on the input signal and thereby deduce the presence and severity of damage. Note that in cases where an input and/or target location is very close to the site of the nonlinearity, the corresponding linear stage may not be needed, but typically we would not know in advance where the nonlinearity will appear. A likely method for damage detection would be to continually process data using various combinations of neighboring sensors looking for a significant rise in nonlinearity above the background level.

The input stage (B1) can be approximated as an FIR map:

$$w_n = \sum_{i=1}^{N_\mathrm{I}} \sum_{k=0}^{D_{\mathrm{B1}i}-1} b_{i,k} r_{i,n-k+k_{\mathrm{B}i}}, \quad (6)$$

where $b_{i,k}$ is set of adjustable coefficients, $D_{\mathrm{B1}i}$ is the number of taps, and $k_{\mathrm{B}i}$ is an index offset which may be different from $k_{\mathrm{A}i}$. We may choose nonlinear stage (B2) to be a memoryless power series:

$$x_n = \sum_{k=1}^{P} c_k (w_n + v_n)^k, \tag{7}$$

where c_k is set of adjustable coefficients and v_n is the output of the optional feedback stage B4 (discussed below). Note that one could choose to include only certain powers, such as odd or even (related to the symmetry of the nonlinearity). In some cases one might wish to use a weakly memoried nonlinear stage that depends on a small number of previous time steps. Certain types of nonlinear stages, e.g., a bilinear function, might require an input biasing coefficient. The power series (with no missing powers) will not need a bias, as such a bias could be absorbed by suitable coefficient changes. The output stage (B3) may be approximated as another FIR map:

$$y_n = \sum_{k=0}^{D_{\mathrm{B3}}-1} d_k x_{n-k}. \tag{8}$$

In cases where delayed self-interaction of the nonlinear component plays a significant role we may include the optional linear feedback stage (B4):

$$v_n = \sum_{k=1}^{D_{\mathrm{B4}}} e_k x_{n-k}. \tag{9}$$

Note that we start the sum with $k=1$ to ensure that this stage is strictly causal.

The combined structure with output $z_n = u_n + y_n$ is shown in Fig. 1 with r_n representing the (one or more) input data sets. The entire parameter vector for the nonlinear mapping therefore includes $b_{i,k}$, c_k, d_k, e_k, and possibly the index offsets $k_{\mathrm{B}i}$. These parameters are adjusted in order to minimize the cost function (3). The sampling rate, the maximum power of the nonlinear stage and the number of taps used in the linear stages must also be chosen, but here one must consider the trade-off between the complexity of the model and the accuracy of the result. The coefficients of the parallel linear stage (and hence its output set u_n) may be held fixed during this process or they could be reoptimized simultaneously with those of the nonlinear mapping. Simultaneous optimization may be more critical for applications where the model itself is used, such as for telecommunications, than for applications where the main purpose is the detection of nonlinearity, such as for SHM. Care should be taken to ensure that the input to the nonlinear map (B) is restricted to the taps utilized by the parallel linear map (A). as this will make the detection process immune to *false positives* for nonlinearity.

The full nonlinear map B depends nonlinearly on its coefficients and therefore requires the use of an optimization algorithm. There are several good choices [13,14] including the Powell “direction set method,” the Brent method (a variant of Powell) and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) “quasi-Newton method” with numerically determined derivatives. We have a slight preference for the Brent method. Nonlinear optimization problems like this can be susceptible to getting stuck in a local minimum although this has not been a major problem. Note that the initial values for the coefficients should be chosen so that the initial output y_n will be exactly zero. This will ensure that $M(\vec{\theta})$ will never be larger that it was for map A alone.

When the purpose is nonlinearity detection, we note that the error measure, being essentially a mean square amplitude, may in some cases be considered to be a kind of measure of the vibrational power of the signal it is based upon. (This would seem reasonable for the case of the strain measurements which we give in our experimental results below.) After first optimizing the linear mapping (A) and then the parallel nonlinear mapping (B) we are immediately presented with three quantities: the power, P_O, of the original target time series, the residual power, P_L, after a linear correction and the residual power, P_N after both linear and nonlinear corrections were applied. The fraction of the power removed by the linear map A is

$$\Gamma_{\mathrm{L}} = (P_{\mathrm{O}} - P_{\mathrm{L}})/P_{\mathrm{O}}, \tag{10}$$

and the fraction of power removed by the addition of the nonlinear map B is

$$\Gamma_{\mathrm{N}} = (P_{\mathrm{L}} - P_{\mathrm{N}})/P_{\mathrm{O}}. \tag{11}$$

Nonlinearity and possible damage is indicated when Γ_N goes up and/or Γ_L goes down.

C. Optimization in the frequency domain

In some cases there may be an interest in optimizing the performance of the model for some particular range of frequencies. One way to accomplish this is to divide z_n, the output of the MSM, into groups of consecutive time steps and perform a Fourier transform on each group and compare this to the corresponding transform for the target time series s_n. The error measure $M(\vec{\theta})$, previously defined by Eq. (3), can be redefined as the mean square difference between the corresponding values obtained by discrete fast Fourier transforms. The contributions to $M(\vec{\theta})$ can be limited to a frequency band of interest or they can be weighted by frequency according to the level of importance. Frequency domain optimization is definitely an option for the telecommunications application (discussed below) because the transmitted data for this case is typically coded into the signal in the frequency domain. Thus it is critical for the model to work well at those frequencies which are being used to transmit data, at not so important at other frequencies.

III. EXPERIMENTAL RESULTS FOR DAMAGE DETECTION

The experimental structure to which we applied the theory is a composite beam measuring 1.219 m in length by 17.15 cm in width and 1.905 cm in thickness. The beam was bolted at both ends to two steel plates using 4 × 1.9-cm-thick bolts measuring 8.9 cm length. Each of the bolts are Strainsert instrumented bolts capable of measuring axial force. The composite material utilizes a quasi-isotropic layup consisting of (0/90) and (+/−45) 24 oz knit EGlass fabric. Excitation was provided by means of a MB Dynamics (PM50a) electrodynamic shaker, coupled to the mid-span of the beam

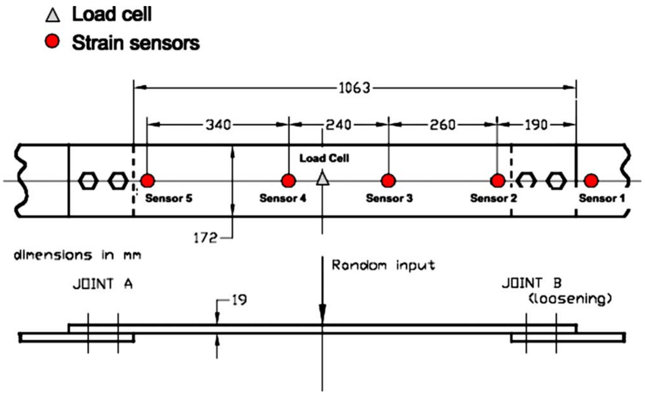

FIG. 2. (Color online) Composite beam, showing sensor locations.

through a thin aluminum rod. Between the rod and the beam is an Sensotec Model 31 load cell for recording the input signal. Note that the load cell interacts with the beam dynamics and thus is not a pure input signal.

The vibrational response of the structure was measured at five separate locations, as shown in Fig. 2, at a data rate of 1951 Hz. using a fiber optic strain sensing system with fiber Bragg gratings (FBGs) as the sensing element.

In addition to the fully tightened state, we analyzed data from three "damaged states" involving the loosening of both bolts connecting one end of the composite to the steel: finger tight, small gap (the nut holding the bolt in place is loose but the bolt is still firmly held in the bolt holes), and large gap (the nut is loose and the bolts were loosened in the holes). The damage constitutes a localized nonlinearity in the following sense: at all other locations, the beam is characterized by a stiffness parameter that is approximately constant over the range of motion to which the system is subjected, but at the site of the loose bolts, the stiffness is dependent on bending. When the beam is unbent at that site, the bolts will be in a "slack" state and so the stiffness will be very small or zero. But when the beam is significantly bent in either direction at that site the bolts will engage and the stiffness will return to near its normal value. Thus the governing equations will only be nonlinear at that location.

In order to generate the response signals, we applied 30 s of dynamic loading and recorded the structural response from all five FBGs as well as the excitation (using a load cell). The dynamic loading was chosen to conform to a random process described by the Pierson-Moskowitz frequency distribution for wave height in order to mimic the type of loading this component would be subject to on a ship structure.

TABLE I. Nonlinear power fraction, Γ_N, obtained for various bolt conditions. Other than the fully tight case, these represent varying degrees of damage. Columns labeled input and target list the sensor numbers from which the corresponding time series were taken (with L used to represent the load cell output). Note that all of the "damaged" cases are significantly more nonlinear than the fully tight case, i.e., the damage has been successfully detected by the analysis.

Input	Target	Fully tight (%)	Finger tight (%)	Small gap (%)	Big gap (%)
L	1	0.070	0.650	10.100	19.100
1	L	0.035	0.460	2.630	5.200
3	2	0.159	2.270	5.030	1.980
5	2	0.388	1.310	6.830	2.180

We used 38400 data points as training data and 6400 as testing data. We tried a variety of input and target combinations and also a variety of control parameters. In virtually all cases we could strongly detect the nonlinearity produced by the "big gap" and "small gap." For most combinations we were also able to reliably detect the "finger tight" case. We ran some tests with $k_A = k_B = 0$, $D_A = 39$, $D_{B1} = D_{B3} = 20$, and $P = 3$ (which requires a total of 82 adjustable coefficients compared to 11479 for the equivalent Volterra map [15]). Results are shown in Table I. In the first case, we used the load cell as the input and sensor 1 as the target. The dramatic increases in Γ_N compared to the fully tight result clearly identifies all of the loose bolt trials, including finger tight, as nonlinear and therefore as probably damaged. Switching the input and target we again can clearly identify the nonlinearity including the difficult finger tight case. Note that some configurations have a higher background nonlinearity (the tight result) than others—this does not appear to be noise, but rather an unknown source of nonlinearity, possibly some slight beam damage or a defect in the sensors. The last two cases, which use sensor 2 as the target, show a smaller response to the big gap than to the small gap. One possible explanation is that the nonlinear signal is weak at some locations and strong at others and that these locations change with gap size. The ability to detect the damaged state appears to be at least comparable to the surrogate method used in Ref. [5] which examined data from the same experimental system. We plan to make more detailed comparisons in future studies.

IV. TELECOMMUNICATIONS APPLICATION

A. Bad electrical contacts

We now return to the telecommunication application. In addition to the (slight) nonlinearity of electronic components, a common source of nonlinearity in such systems is bad electrical contacts, which may occur where there are splices in a long transmission line. Since we do not know of a good reference on this topic, we will discuss it briefly here. There are at least two mechanisms that can result in such "non-ohmic" behavior which are related to solid state devices which have been extensively studied. Specifically, these devices are metal oxide metal (MOM) diodes [16,17] and metal-metal point contact diodes [18,19]. MOM diodes are also referred to as metal insulator metal (MIM) and metal barrier metal (MBM) diodes.

In MOM diodes the dominant conduction mechanism is quantum tunneling of electrons through the thin oxide layer. The oxide layer does not have to be very thick and thus could form naturally in electrical contacts that are exposed to the air. The expected nonlinear behavior in this tunneling case is relatively high resistance at zero voltage, which de-

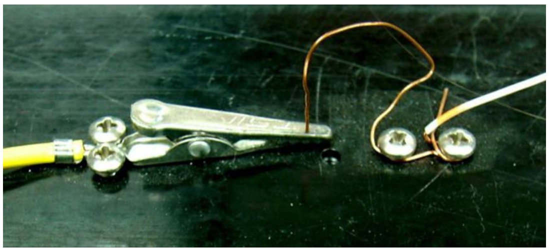

FIG. 3. (Color online) "Bad electrical contact" formed by touching a copper wire to the top of a slightly tarnished alligator clip [20].

creases somewhat as the voltage increases. In metal to metal point contacts, the nonlinearity comes about when the diameter of the contact area between the two bulk conductors is comparable with or smaller than the mean free path of the electrons in the metal. The nonlinearity in this case may be due to an inelastic electron-phonon scattering mechanism and is sometimes referred to as nonlinear spreading resistance. The energy dependence of this effect leads the resistance to exhibit the opposite type of nonlinear behavior, i.e., it increases as the voltage increases. Expansion and contraction due to thermal cycling of an electrical contact with the weather can lead to a poor connection where these types of nonlinear effects may occur.

In Fig. 3 we show a simple test system consisting of a contact between a slightly tarnished alligator clip and a piece of copper wire. As shown in Fig. 4 this simple system is found to exhibit both kinds of nonlinear behavior depending very sensitively on the exact positioning of the contact. Initial resistance in these nonlinear cases is typically in the range of 5–80 Ω, with the higher values tending to be associated with the tunneling type of behavior. Wire to wire contacts were also found to show such nonlinear effects.

B. Equalization and echo cancellation

There are at least two ways in which our method could be used in telecommunications. In the first, it could be used as an enhancement of the usual "equalization" process that is applied to a transmitted data signal after it is received. This would normally correct for memoried linear effects of the line, using, for example, a FIR map such as that of the parallel linear stage A in Fig. 1. (Note that in this case the desired coefficients for the map are actually those that effectively invert the impulse response of the transmission line.) The appropriate coefficients could be obtained by a training process using a known test sequence transmitted on the line. By adding the nonlinear map B, we now have the possibility of also correcting for a nonlinear contact (or any other nonlinear element) in the line if one exists. Note that the correction does not have to be perfect, only good enough to result in an improvement in the maximum data rate that the line will support. A second way that the method can be used is to improve what is known as echo cancellation. Often data are transmitted simultaneously in both directions on one transmission line. This is the case with digital subscriber line (DSL), which is the method commonly used to transmit data at high speeds over ordinary telephone lines, which consist of twisted pairs of wires. Telephone lines, which may be quite old, will often contain nonlinear bad contacts. If there are any reflections of the outgoing signal at some point in the line, these will combine with the incoming signal and will act as noise, reducing the maximum data rate that is possible for that signal. The purpose of echo cancellation is to use the known outgoing signal to drive a digital filter which has been trained to generate a model of the expected echo. This is done in real time and subtracted from the incoming signal to remove the actual echo from it.

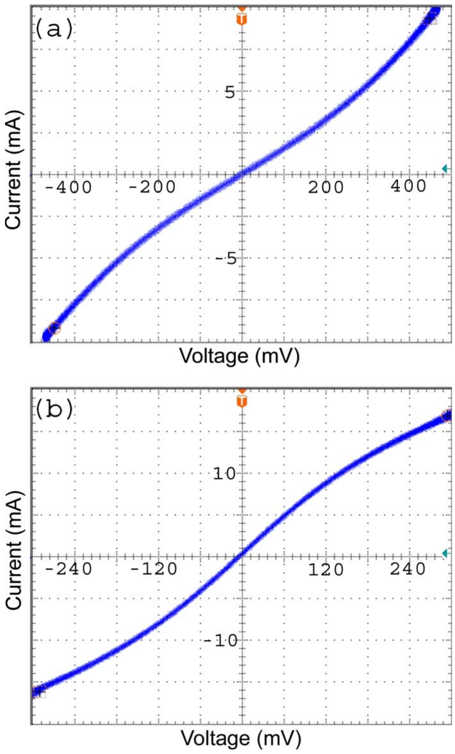

FIG. 4. (Color online) Two nonlinear I-V characteristics obtained from the contact of Fig. 3 by positioning the wire at different locations on the clip. Note that the curves are highly symmetric. The behavior is extremely sensitive to contact positioning or movement. Other cases (not shown) yield "normal" (linear low resistance) contacts as well as additional nonlinear cases. (a) The "antisaturating" type, where resistance decreases with voltage. This case may be a result of tunneling through a thin oxide layer which is the mechanism of a MOM diode. (b) The "saturating" type, where resistance increases with voltage. This case may be exhibiting nonlinear spreading resistance associated with a metal-metal point contact of very small cross section.

We now show explicitly, for the case of echo cancellation, how to derive the multistage nonlinear model of Fig. 1, following the discussion of Sec. II A. As before, the purpose of this exercise is to show that the system can be represented by a model with the indicated structure.

Assume that there is a single nonlinear contact in one of the wires of the twisted pair transmission line. The system can be described by three parts: the transmission line up to the contact, the contact itself, and the transmission line from the contact to the end of the line. The contact will be treated like a nonlinear resistance, R_X. The voltage V_X across R_X

corresponds to the "force" F discussed in Sec. II A and the current I_X through it corresponds to the "response" R. The I-V characteristic of this element is the function $g(\,)$. We will assume that the line can be characterized approximately by a line resistance R_0.

Following the first step in the discussion of Sec. II A, we freeze the nonlinear element so that the current through it is zero. If the line is otherwise ideal and nondissipative then the incident signal $V_i(t)$ arriving at R_X is simply a delayed version of that which entered the line at the source. In reality of course, the line will have attenuation and other linear effects and it is the function of stage B1 to model these effects. Since the contact is frozen we know that an incoming signal should be completely reflected at that point and the sum of the incoming and reflected signals will have double the amplitude of the incoming signal alone. Thus the output of stage B1 will be $2V_i$ as this is the effective contribution of this incoming signal to the forcing of the nonlinear element. Any linear "reflections" occurring within the first section of the transmission line and returning to the input (including the linear reflection occurring at the site of the nonlinear element) are modeled by the parallel linear stage A, whose output constitutes the linear contribution to the echo.

Since the response of the nonlinear element is a current, the second step in the discussion of Sec. II A is to replace the contact with a current source $I_X(t)$. This current source is effectively driving two transmission lines: one returning to the input and the other going on to the far end of the line. Due to the line resistance, a voltage drop of $-2R_0I_X(t)$ will be generated and this will be added to the voltage generated by any incoming waves. This is the "immediate self-interaction," and therefore the coefficient λ is given by

$$\lambda = -2R_0. \tag{12}$$

Using this value in Eq. (1) and taking the inverse, we can obtain the function $h^{-1}(\,)$ that defines the nonlinear stage B2.

$I_X(t)$ also determines the transmitted wave $V_t(t)$ and the (nonlinear) reflected wave $V_r(t)$. In some cases, it may be reasonable to assume that neither of these waves is reflected back to the nonlinear element, in which case we can immediately calculate the starting amplitude of these waves, i.e., $I_X(t)R_0$ and $-I_X(t)R_0$, respectively. For echo cancellation, the reflected wave is the one of interest. It will pass backward through the first section of the transmission line to reach the input point. This is a linear process which is modeled by the linear stage B3 and depends solely on the properties of this section of the transmission line. The output of this stage constitutes the nonlinear contribution to the echo.

If we wish to include reflections of these waves which arrive back at the contact point, i.e., $V_{tr}(t)$ and $V_{rr}(t)$ which are the reflections of $V_t(t)$ and $V_r(t)$, respectively, the problem becomes considerably more complicated because some of the current $I_X(t)$ corresponds to contributions from these incoming waves. However, it is still a linear problem, i.e., a current source at the contact site driving a linear system. Thus all of these waves $[V_t(t), V_r(t), V_{tr}(t)$, and $V_{rr}(t)]$ must be linear TD transforms [11] of $I_X(t)$. They will depend on the properties of the sections of the line through which they pass. When these two extra waves are included, the model will have a feedback stage B4, otherwise it will not. Since waves incident at a current source will be reflected, their amplitudes will be doubled, and since $V_{tr}(t)$ approaches the contact from the rear, the output of stage B4 will be $2V_{rr}(t)-2V_{tr}(t)$. Note that after being re-reflected these waves are incorporated into the waves $V_t(t)$ and $V_r(t)$ that are leaving the contact site.

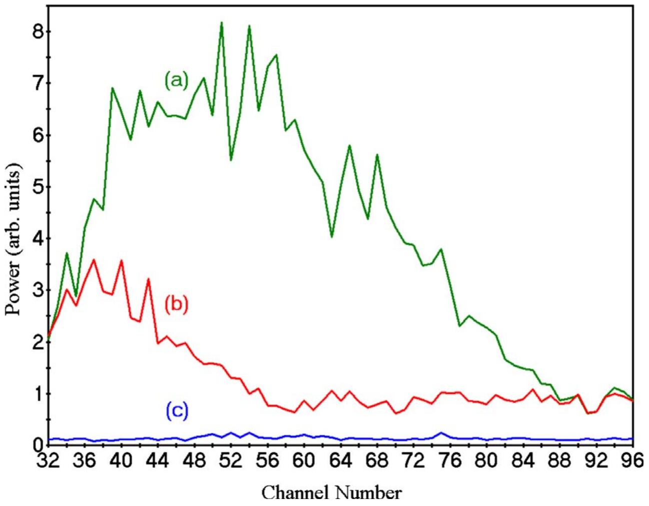

FIG. 5. (Color online) Nonlinear echo of an ADSL signal in a working telephone line with a suspected bad electrical contact. (a) Uncorrected nonlinear echo. (b) Remaining echo after nonlinear echo cancellation using the method described in this paper. (c) Background noise level with no signal on the line. The upstream signal is in channels 6 through 31 (not shown). The echo has components in the downstream band (channel 32 and above) due to mixing of the upstream frequencies at a nonlinear contact in the line. Linear echo cancellation has no effect on the nonlinear echo. The frequency of a channel is obtained by multiplying channel number by 4.3125 kHz. If implemented in a modem, this reduction in echo would lead to an increase in the maximum data rate for this line.

Now that we have shown that the MSM model is appropriate for this problem, we can proceed to optimize the model. First transmit a known outgoing signal on the line with no incoming signal present, and simultaneously capture any resulting echo. Then find a set of coefficients for the model such that when used to process the known output signal it minimizes the error measure $M(\vec{\theta})$ in modeling the echo. This model with the calculated coefficients is then the desired echo canceller to be driven by the real transmitted signal. The modeled echo would be subtracted in real time from the real incoming signal.

C. Experimental results

In Fig. 5 are shown some preliminary echo canceller results. The data were taken from a test run on a working telephone line with a suspected bad contact in the line. This was a single ended test, run from the user's end of the line. An "upstream" signal was transmitted on the line, while simultaneously capturing any incoming reflected signal. The signal generated was of the type used in asymmetric digital subscriber line (ADSL) [21], which has separated upstream and downstream bands. The full spectrum of 1.104 MHz is broken up into 256 channels each 4.3125 kHz wide. The transmitted signal contained random data in upstream channels 6 through 31 and contained no signal in channels above that. However, as is seen in Fig. 5, the echo does appear to contain significant power in the range of channels shown, 32 through 96, all of which are in the downstream band. This is easily explained as a result of the interaction of the signal with a nonlinear element in the line. For a cubic nonlinearity, new frequencies are generated that are the sums and differences of any combination of three frequencies found in the original signal. Since that signal included frequencies through channel 31, the echo might be expected to contain frequencies through channel 93. By generating a model and driving it with the transmitted upstream signal, it was possible to cancel a significant portion of the echo. The analysis used 27200 points of training data and 27200 points of testing data, with a sampling rate of 2.208×10^6 samples/s. For the model, the linear stages B1 and B3 used 17 taps ($D_{\mathrm{B1}} = D_{\mathrm{B3}} = 17$) and the nonlinear stage B2 included up to power $P=3$. This model was optimized in the frequency domain over channels 32 through 96. Offsets k_{A} and k_{B} were set to zero. The parallel linear stage A used $D_{\mathrm{A}}=33$ taps, and as expected it had no effect on the nonlinear echo. Since the echo would look like noise added to the "real" downstream signal, any reduction in that noise for a particular channel will allow a higher density of information to be transmitted on that channel. For comparison, a test run on a normal line (not shown) yielded no nonlinear echo and no improvement could be obtained through use of the model.

V. CONCLUSION

We have described the method of using a multistage transform as a means to model the response of general continuum systems containing a localized nonlinearity. The stages are configured as shown in Fig. 1 and may be approximated as memoried mappings. This method may be a useful tool to detect damage in physical systems such as bridges and ships where the early detection of such damage may be of great value. As a test of the method we analyzed data from a shaken beam with loose bolts. We plan to follow this with a more detailed study in the near future. We have also demonstrated with experimental results that the method may be of value for the detection and correction of nonlinear effects in communication channels such as those caused by bad electrical contacts or by the (slight) nonlinearity of electronic components. Such improvements would lead to an increase in the achievable data rate. Finally, there may be other applications of the method, yet to be discovered.

[1] K. D. Murphy and J. M. Nichols, Int. J. Non-linear Mech. **44**, 13 (2009).

[2] W. Z. Zhang and R. B. Testa, J. Eng. Mech. **125**, 1125 (1999).

[3] S. T. Trickey, L. N. Virgin, and E. H. Dowell, Proc. R. Soc. London, Ser. A **458**, 2203 (2002).

[4] J. M. Nichols, M. Seaver, S. T. Trickey, M. D. Todd, C. Olson, and L. Overbey, Phys. Rev. E **72**, 046217 (2005).

[5] J. M. Nichols, S. T. Trickey, M. Seaver, S. R. Motley, and E. D. Eisner, J. Vibr. Acoust. **129**, 710 (2007).

[6] I. Trendafilova and H. Van Brussel, Mech. Syst. Signal Process. **15**, 1141 (2001).

[7] T. Schreiber and A. Schmitz, Phys. Rev. Lett. **77**, 635 (1996).

[8] T. Nakamura, X. Luo, and M. Small, Phys. Rev. E **72**, 055201(R) (2005).

[9] R. G. Andrzejak, A. Kraskov, H. Stogbauer, F. Mormann, and T. Kreuz, Phys. Rev. E **68**, 066202 (2003).

[10] T. Schreiber and A. Schmitz, Physica D **142**, 346 (2000).

[11] A time-domain (TD) transform, as used in this paper, is a time invariant mathematical operation that acts on an input function of time to produce an output function of time. By time invariant, we mean that a shift in time of the input function produces an identical shift in time of the output function and does nothing else. If linear, it is in the form $q(t)=\int_{-\infty}^{\infty} p(t')\chi(t-t')dt'$, where $p(t)$ is the input, $q(t)$ is the output, and $\chi(\tau)$ is a linear-response function. Typically $\chi(\tau)=0$ for $\tau<0$ so that the transform is causal. If nonlinear, a TD transform can be expanded as a continuous time Volterra series. However, the nonlinear stage used in this paper is usually memoryless and one dimensional, in which case the corresponding transform is a simple nonlinear function which can be expanded as a simple power series.

[12] Equation (3) includes "Bessel's correction," an overall factor of $N/(N-1)$, that removes a bias in estimating the "true" variance when limited to a sample of N data points. It is, however, still best to avoid using small values of N. Increasing N increases the accuracy of the variance estimate and also reduces any biasing of that estimate that may result from autocorrelation of the data. Note that since Bessel's correction is an overall factor, it actually has no effect on the optimization process.

[13] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing* (Cambridge University Press, England, 2007).

[14] R. Brent, *Algorithms for Minimization without Derivatives* (Prentice-Hall, 1973), Chap. 7.

[15] For the chosen parameters, each output of the multistage map depends on 39 consecutive inputs. The equivalent Volterra map would include all cross terms up to third power amongst the 39 inputs. For D inputs and power P the total number of terms is given by $(N+D)!/(N!D!)$. Omitting the constant term we have $42!/(39!3!)-1=11479$.

[16] I. Wilke, Y. Oppliger, W. Herrmann, and F. K. Kneubühl, Appl. Phys. A: Mater. Sci. Process. **58**, 329 (1994).

[17] A. Simmons, J. Appl. Phys. **34**, 1793 (1963).

[18] R. W. van der Heijden, H. M. Swartjes, W. Herrmann, and P. Wyder, J. Appl. Phys. **55**, 1003 (1984).

[19] A. G. M. van der Heijden, A. P. van Gelder, and P. Wyder, J. Phys. C **13**, 6073 (1980).

[20] Brought to the attention of one of the authors (P.B.) by Bob Monteleone.

[21] See, e.g., T. Starr, J. M. Cioffi, and P. J. Silverman, *Understanding Digital Subscriber Line Technology* (Prentice Hall, 1999).