Analyst

PAPER

View Article Online
View Journal | View Issue

Cite this: *Analyst*, 2014, **139**, 5927

Karhunen–Loève treatment to remove noise and facilitate data analysis in sensing, spectroscopy and other applications

V. V. Zaharov,[ad] R. H. Farahi,[ac] P. J. Snyder,[a] B. H. Davison[a] and A. Passian*[abc]

Received 17th July 2014
Accepted 8th September 2014

DOI: 10.1039/c4an01300j

www.rsc.org/analyst

Resolving weak spectral variations in the dynamic response of materials that are either dominated or excited by stochastic processes remains a challenge. Responses that are thermal in origin are particularly relevant examples due to the delocalized nature of heat. Despite its inherent properties in dealing with stochastic processes, the Karhunen–Loève expansion has not been fully exploited in measurement of systems that are driven solely by random forces or can exhibit large thermally driven random fluctuations. Here, we present experimental results and analysis of the archetypes (a) the resonant excitation and transient response of an atomic force microscope probe by the ambient random fluctuations and nanoscale photothermal sample response, and (b) the photothermally scattered photons in pump–probe spectroscopy. In each case, the dynamic process is represented as an infinite series with random coefficients to obtain pertinent frequency shifts and spectral peaks and demonstrate spectral enhancement for a set of compounds including the spectrally complex biomass. The considered cases find important applications in nanoscale material characterization, biosensing, and spectral identification of biological and chemical agents.

1 Introduction

Geometric and molecular identification of materials by probing at contact or at a distance, remains a major challenge in physical, chemical, and biological detection, sensing, and material characterization.[1–3] While this challenge thrives from lack of adequate mechanical, electromagnetic, and chemical probes, in cases where such measurements are possible, the detected signal is typically associated with random fluctuations that can render the fine spectral features inaccessible.[4,5] Point and standoff detection of complex molecules that make up chemical and biological agents are increasingly explored for applications ranging from medicine[6,7] to security. In nanometrology, depending upon the local morphology and concentration of the sought compound or the desired detection threshold, and random characteristics of both the sample/environment and the measuring systems, the acquired data can contain a variety of stochastic components that in some cases can dominate the useful signal despite employment of phase sensitive detection.

Many material characterization and biosensing experiments, employ the micro-cantilever force sensor platform[8] or the atomic force microscope probes, as well as laser reflectometry and interferometry that require operation in atmospheric condition, dense gases and fluids, leading to further sensitivity difficulties due to gas kinetic and hydrodynamic dissipation and coupling as well as background absorption and scattering. For example, by using quantum cascade lasers and invoking a pump–probe scheme, we recently demonstrated acquisition of spectral fingerprints of substances from a variable standoff distance.[1,9] However, the pump–probe measured data are often accompanied by a high level of random fluctuations that obstruct the systematic pattern. As for any laser remote sensing measurement, the nature of those fluctuations are uncorrelated or weakly correlated random noise, and weakly or strongly correlated instrumental errors.[3] Here, aiming toward better molecular recognition, we treat the measurement results of the Brownian motion of a micro-oscillator; and the infrared absorption of molecularly complex materials, by employing spectral analysis and Karhunen–Loève Transform (KLT). In the case of the Brownian motion of the cantilever, the KLT possesses advantages in the data processing, by virtue of its inherent suitability for handling non-stationary random functions. Furthermore, KLT can detect signals embedded in noise when values of the signal-to-noise ratio is extremely small (∼40 to −50 dB).[10,11] Among the variety of different noise suppression techniques, KLT[12,13] plays a unique role despite its implementation complexities and higher computational cost

[a] Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6123, USA. E-mail: passianan@ornl.gov
[b] Department of Physics, University of Tennessee, Knoxville, TN 37996, USA
[c] Department of Chemical and Biomolecular Engineering, University of Tennessee, Knoxville, TN 37996, USA
[d] Polytechnic University of Puerto Rico, 377 Ave Ponce De Leon, San Juan, PR 00917-3006, USA

compared to FFT. In 1988 Freire and Ulrych introduced the application of the KLT to seismic profiling, where they referred to KLT as the singular value decomposition (SVD), and demonstrated its use as a new method of separating upgoing and downgoing waves in vertical seismic profiling (VSP) sections.[14] Following in 1991 Al-Yahya discussed practical techniques for reducing the computational cost of SVD in filtering seismic sections.[15] Recently, C. Maccone has elucidated the importance of KLT in applications related to broad-band and extremely feeble signals encountered in SETI (search for extraterrestrial intelligence) searches and has further emphasized the advantages of the KLT compared to Fast Fourier Transform (FFT).[10,11] KLT/SVD has also been applied to remove random noise in real time cardiac cine MRI (magnetic resonance imaging) without blurring stationary or moving edges[16] and to denoise MRS (magnetic resonance spectroscopy) data.[17] The use of KLT is also valuable in the noise reduction and estimation of recorded speech.[18,19] Furthermore, SVD can play an important role in denoising the multi-faceted sensor array data generated in electronic nose systems.[20] While KLT, SVD, and PCA (principal component analysis) have frequently been mentioned in common context, there are subtle differences among them, as for example described by J. J. Gerbrands in the context of digital image processing.[21] Here, we demonstrate that KLT provides similar advantages when applied to sensing and spectroscopy signals. For example, other sensor technologies that could benefit from KLT denoising are nanowire temperature sensors,[22] graphene-based biosensors for immunoassay of small molecules,[23] indirect absorption spectroscopy,[24] microcantilever-based detection of protein markers,[25] microcantilever-based detection of CrO_4,[2–8] mid-infrared spectroscopy of exhaled breath analysis,[26] surface-enhanced Raman scattering (SERS) detection of organophosphate pesticides,[27] and also sensing applications where surface heating by optical beams occur.[28]

In Section 2 the analysis of the random fluctuation of a micro-oscillator is discussed; KLT is introduced and applied to both ambient and photoacoustically induced responses. The random fluctuation in spectroscopic measurement is discussed in Section 3, where the experimental setup and analysis of the obtained infrared (IR) absorption spectra for a set of test compounds including Poplar cross sections, a biomass with considerable molecular complexity and of great importance in biofuel research are given. Concluding remarks given in Section 4.

2 Random fluctuations of a solid micro-oscillator

The dynamics of a free micro-cantilever is commonly acquired *via* the optical beam deflection detection employed in atomic force microscopy (AFM).[5,29] Consider $S(t)$ as the signal representing the relevant observable in the cantilever dynamics, that is, the transversal deformation $u(x, t)$ at a given point x along the length of the cantilever (assuming other oscillatory modes being negligible under the weak random fluctuations). We will begin by first treating the Brownian response of the cantilever under no driving forces. The case where a transient driving force is also presented, while the probing tip of the cantilever is engaged in contact forces with a sample surface, will be treated following the Brownian case. Upon interaction with the surface (as in the contact mode), the spectrum of the probe undergoes a redshift. Similar to molecular binding (adsorption) induced frequency shifts when biosensing with microcantilevers, such interface induced frequency shifts are amenable to the treatment here. In AFM, a free probe, that is, when the tip of the microcantilever is outside the range of interfacial force fields (typically a few nanometers away from the surface), can be described by $S(t)$. To explore the specific utility of the KLT in the treatment of the cantilever response, we also consider the case where the AFM is operated in the contact mode for the detection of the spectral properties of the sample. In particular, morphologically highly heterogenous and chemically highly complex biological samples are extremely difficult to characterize noninvasively. The specific material considered here is the biomass and the characterization method employed is the broadband infrared absorption spectroscopy. The photothermally induced transient effects are embedded in $S(t)$, which we intend to study.

2.1 Measurements and KLT analysis

Resonant cantilever response can be experimentally observed as a result of phonon fluctuations and interaction with particles in a medium actuating the sensor into vibrations. Within the linear response, the resonance frequencies are identified with Lorentzian peak shape functions in the frequency domain. Apart from the physical properties of the probe, the resonance spectrum strongly depends on the nature of the interacting particles. Resonance frequency shifts provide the basis for many sensing applications, where unknown analytes or adsorbates may be detected and classified. The quality-factor Q determines the resolution, and thus the precision of the resonance frequency shift measurement. In the absence of any external driving forces, $S(t)$ represents the equilibrium state of u and the accumulative random fluctuations in the entire system including the electronics noise and the Brownian oscillations of the cantilever at temperature T. Denoting the Fourier transform of the signal $S(\omega)$, the fluctuation-dissipation theorem states that $S(\omega)S^*(\omega) = (2K_BT/\omega)\Im(S(\omega))$ (which for example can take the form of eqn (5) in ref. 30), where K_B is the Boltzmann constant. Embedded in $S(t)$ is the resonant oscillations of the cantilever due to stochastic excitation as shown in Fig. 1. For a given frequency range $\Delta\omega$, the mode n dependent power spectral density of the resonant form $\langle|u_n|\rangle(\omega_n, T) = (2m^2k_BT\Delta\omega/\pi\gamma^3)/Q_n^2$, where Q_n, $n = 1, 2, \ldots$ depend on the cantilever geometry, material, elasticity and mass (m). The higher the Q, the higher the sensitivity of the sensor and the narrower the resonance peak bandwidth; hence, a shift in resonance frequency can be detected and estimated with the high precision. However, despite the possibility of high Q that can provide high sensitivity, the response of the sensor can be rather slow. For example, for a cantilever with $Q = 5 \times 10^4$ and a resonant frequency $f_r = 50$ kHz, the

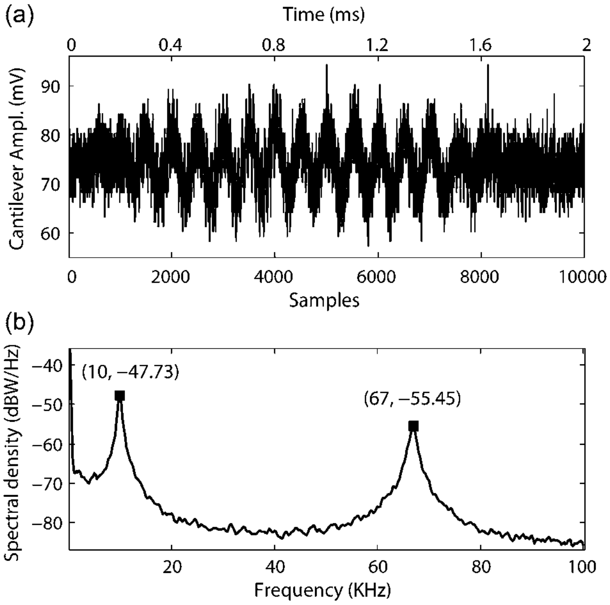

Fig. 1 An instance of the measurement of the equilibrium state of the atomic force microscope probe, the microcantilever. (a) Time domain representation of 10 000 samples at $\Delta t = 0.2$ µs, and (b) the resolved leading resonances $f_r^{(1)} = 10.025$ kHz and $f_r^{(2)} = 66.917$ kHz in frequency domain.

maximum bandwidth[29] is only 0.5 Hz. The corresponding time domain response given as $\tau = 2Q/\omega_r = 0.32$ s, which may be too long for some applications. Furthermore, the dynamic range of the high sensitivity system is also restricted due to high amplitude response at resonance. Thus, the cantilevers with high Q must be examined carefully as the higher Q oscillator would require longer time to reach a steady state. Low Q cantilevers offer faster response, *e.g.* Fig. 1 indicates that a Q-factor of ≈ 7, and therefore $\tau^{(1)} \approx 2$ ms. However, the experimentally observed low peak resolution means the shifts in resonance frequencies cannot be estimated precisely and can be undetectable. Hence, one arrives at contradictory implementation requirements: while a fast response is sought, a low sensitivity is desired. Satisfaction of both conditions is a challenge that requires a trade-off.

We propose to apply the KLT for denoising the low Q cantilever data, and exploiting the high resolution properties of KLT to significantly increase the precision of the resonance frequency estimation. We therefore consider a decomposition of the stochastically generated signal $S(t)$, $t \in [t_1, t_2]$ as an infinite linear combination of orthogonal functions $\psi \in L^2[t_1, t_2]$:

$$S(t) = \sum_{k=1}^{\infty} s_k \psi_k(t), \tag{1}$$

where

$$s_k = \int_{t_1}^{t_2} S(t)\psi_k(t)\mathrm{d}t, \tag{2}$$

are uncorrelated random variables. The KLT employs a set of orthogonal functions that provides the optimum approximation of the original stochastic process in the sense of the total mean-square error (MSE) $\langle [S(t) - \tilde{S}(t)]^2 \rangle$, where $\tilde{S}(t) = \sum_{k=1}^{L} s_k \psi_k(t)$ is a truncated decomposition of $S(t)$ up to L terms. The solution is found by solving the homogeneous Fredholm integral equation of the second kind:

$$\int_{t_1}^{t_2} K_S(\tau, t)\psi_k(\tau)\mathrm{d}\tau = \lambda_k \psi_k(t), \tag{3}$$

where $K_S(\tau, t)$ is the covariance function (kernel) satisfying the Mercer theorem,[31] which postulates that a set of eigenvalues and eigenfunctions, $(\lambda_k, \psi_k(t))$, exists and form an orthonormal basis on $L^2[t_1, t_2]$, such that $K_S(\tau, t)$ can be expressed as:

$$K_S(\tau, t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(\tau)\psi_k(t). \tag{4}$$

By doing so, (2) becomes the KLT. As a result, KLT uses the best possible basis with the fastest convergence of $\tilde{S}(t)$ toward $S(t)$. However, the drawback of the KLT is the high numerical cost of determining the eigenvalues and eigenfunctions of its covariance operator, which are given by the solutions to the integral eqn (3). The sampled presentation of the $S(t)$ in our case simplifies the implementation of the KLT since the eigenvalue decomposition of a discrete covariance matrix can be employed. As an orthogonal transform, the KLT is optimum under the MSE between truncated and the actual data providing the highest convergence of the data vector into smaller dimension subspace. In contrast to Fourier transform, the orthogonal basis of KLT are the eigenvectors of the data covariance matrix. As a result, KLT exhibits powerful properties such as complete decorrelation of the measured data, squeezing the data information into the minimum number of parameters, and splitting the whole data information into two subspaces (useful data and unwanted noisy data) yielding the highest resolution of the data spectral components among other orthogonal transforms.[31,32] However, prior to KLT operation, a data pre-processing is required: the ”training” set of representative data needs to be collected, and the data covariance matrix needs to be formed, and finally by computing the data covariance matrix eigenvalue decomposition, the proper basis functions are to be found. Therefore, KLT is data dependent, and the proper basis functions are never known *a priori* except in cases where the data model is known. Nevertheless, data denoising with KLT offers the advantage of splitting the data into two subspaces, the useful data and the noisy data.

For our data denoising, the KLT algorithm includes the following steps.[31] We will first obtain the data covariance matrix $\mathbf{R} = \frac{1}{M}\sum_{i=1}^{M} S_i S_i^T, \quad M \geq N$, where M is the number of measurements runs, and N is the length of the measurement data vector. We then commence with finding the eigenvalue decomposition $\mathbf{R} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^T$, where $\mathbf{\Lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_N]$ is a $N \times N$ diagonal matrix $\mathbf{R}$, with elements sorted in descending order, and matrix $\mathbf{\Phi} = [\phi_1, \phi_2, \ldots, \phi_N]$ is an eigenvectors matrix, where ϕ_i, $i = 1, \ldots, N$ are $N \times 1$ vectors. We will then proceed by determining the threshold that divide the data space into two subspaces and define m significant eigenvalues, $m \leq N$ and forming an $N \times m$ KLT transform matrix $\mathbf{\Phi}_m = [\phi_1, \phi_2, \ldots, \phi_{\mathrm{m}}]$ that includes only eigenvectors that belong to the largest m eigenvalues of $\mathbf{R}$. Finally, we can compute direct KLT as $\mathbf{y} = \mathbf{\Phi}_m^T\mathbf{x}$, or compute inverse KLT as $\mathbf{x} = \mathbf{\Phi}_m\mathbf{y}$ for reconstruction and data denoising. The computational cost and the memory requirements of the KLT algorithm is $\mathcal{O}\{N^2\}$ *versus* $\mathcal{O}\{N \log_2 N\}$ in the case FFT.[11,31]

2.2 Free oscillation analysis

In the case of the free oscillations of the undriven cantilever, for statistical purposes, $S(t)$ is split and routed to four channels of a digitizing oscilloscope, where all channels are captured in rapid succession, with each channel measurement containing 10^4 points sampled at 200 ns. Since typically only the first few resonances can be directly observed during the random fluctuations, for practicality (for example in biosensing) we consider also the second resonance frequency of the cantilever $f_{\mathrm{r}}^{(2)} = 66.9$ kHz ($T_{\mathrm{r}}^{(2)} = 1/f_{\mathrm{r}}^{(2)} \approx 15$ μs) for adequate sampling (see Fig. 1). Denoting the bandwidth $\delta\omega$, the response time of the cantilever is $\tau_{\mathrm{r}} = 2Q/\omega_{\mathrm{r}} = (2f_{\mathrm{r}}/\delta\omega)/(2\pi f_{\mathrm{r}}) = 1/(\pi\delta\omega)$, that is, independent of f_{r} and the sampling time. For a total observation time of $t_{\mathrm{tot}} = 2$ ms, that is, $t_{\mathrm{tot}}/T_{\mathrm{r}}^{(1)} \approx 2 \times 10^{-3}/100 \times 10^{-6} = 20$ first mode oscillations, a sampling rate of 5 MHz ($\Delta t = 0.2$ μs $= T_{\mathrm{r}}^{(2)}/n$), with $n \approx 75$ chosen, the data is shown in Fig. 1. We note that $\Delta t = 0.2$ μs clearly yields an oversampling for the second resonance but is more suitable for the third resonance (since $f_{\mathrm{r}}^{(3)} \approx 15 \times f_{\mathrm{r}}^{(1)}$ for a rectangular microcantilever).

Following the procedure described in the previous section, we implemented the KLT[33] and processed the data presented in Fig. 1 to obtain a spectrum of 10^4 eigenvalues. This spectrum was acquired without truncating the matrix of eigenvalues and eigenvectors.

As can be seen from Fig. 2, the KLT spectrum has higher resolution than the reference spectrum obtained with Fast Fourier Transform (FFT) and, as a result, the resonance frequencies can be estimated with yet higher precision. However, the high level of accomplished noise can distort the result of the measurement. Thus, to start the data denoising with KLT the distribution of the eigenvalues should be found. The resulting data eigenvalues distribution is presented in Fig. 3. Considering only the largest eigenvalue, we obtain a perfect estimation of the first resonance peak as Fig. 4(a) shows. Involving other significant eigenvalues, following their diminishing amplitudes, we observe that after involving only four eigenvalues the second peak is detected as shown in Fig. 4(b). These powerful resolving features are highly desirable for frequency shift based sensing and imaging. The time domain representation with only the first four eigenvalues depicted in Fig. 4(c) clearly discerns the first and the second frequencies as the lower and the higher oscillations, respectively. The strong and narrow resonance frequency peaks allow the estimation of frequency shifts with much higher precision than those that can be provided by the reference spectrum obtained with FFT.

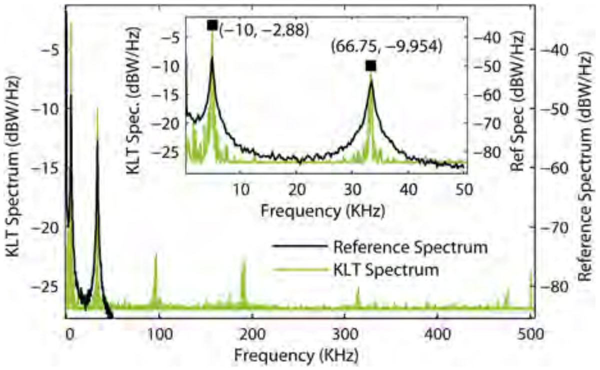

Fig. 2 Brownian response of the microcantilever in the Fourier *versus* KLT domains. Broad range (main plot) and high resolution spectra (inset) for comparison of the resonance peaks.

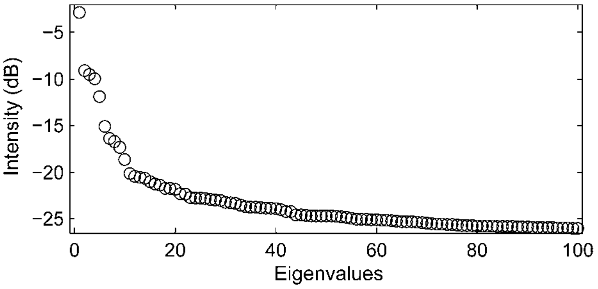

Fig. 3 Distribution of the first 100 eigenvalues obtained by decomposition of the sampled Brownian motion of the microcantilever.

2.3 Analysis of photoacoustically driven transient oscillations

In order to analyze the cantilever transient response, we engaged the cantilever tip in contact with a sample surface. We then examined the cell-wall layers of a 25 μm thick cross-section of extractive-free poplar wood that was mounted on a ZnSe substrate. An interferometric infrared source in the range of 4000–400 cm^{-1} was focused on the sample surface *via* transmission through the substrate. During the interferometer operation at 4 cm^{-1} resolution the AFM probe measures the transient surface response of the poplar due to the photothermal process. This nanomechanical interferogram is inverse transformed into the localized absorption spectra of the material. We positioned the probe in the cell corner middle lamella region within the IR illumination area in order to observe different regions of cellulose and lignin composition, and obtained high spatial and spectral resolution data as shown in Fig. 5. Following the discussions in Section 2 we applied the KLT algorithm for the driven transient oscillations including the first 20 significant eigenvalues, discarding the insignificant eigenvalues which present the random oscillations. As Fig. 5 shows the denoising with KLT allows significantly diminish the stochastic fluctuation even those presented by strong amplitudes bursts. The specific number of eigenvalues to be retained or suppressed depends on the threshold setting between the useful pattern subspace and noisy component subspace. For example, in the case of the Brownian motion, analyzing the

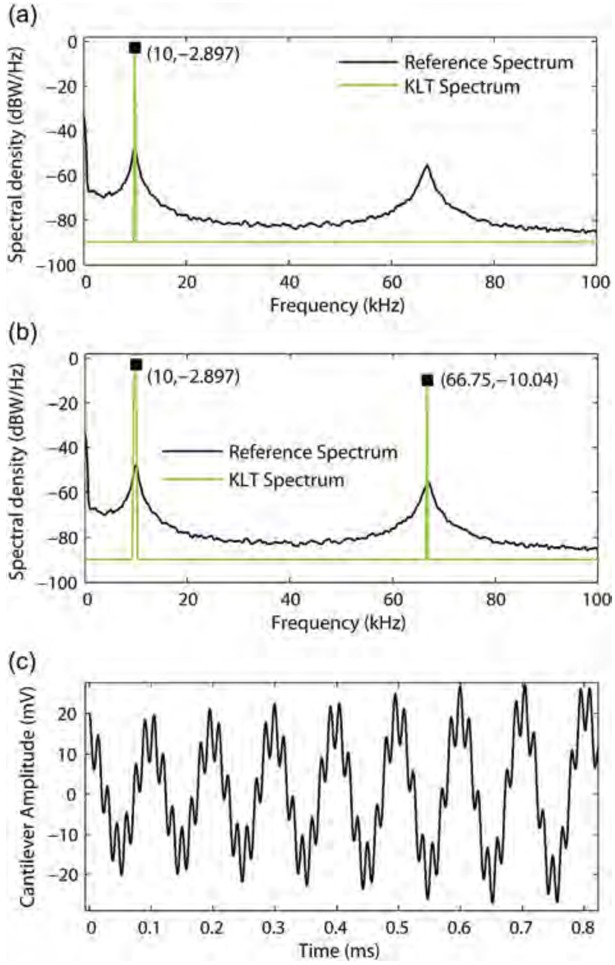

Fig. 4 Peak position estimation with KLT. Estimation of the (a) fundamental resonance employing the single largest eigenvalue, (b) the second resonance peak employing the four largest eigenvalues, and (c) its corresponding denoised and unbiased time domain representation using four eigenvalues out of 10 000.

distribution of eigenvalues in Fig. 3, we can see that the first eigenvalue corresponds to as much as 23 dB higher than that of the 20th (or 200 times as a ratio). For practical applications this can be considered a reasonable ratio. Moreover, after the 20th eigenvalue, the KLT spectrum shows a gently diminishing nature, vanishing to zero, confirming the spectral property of the Brownian process.

3 Random fluctuations in spectroscopic measurements

By using tunable quantum cascade lasers and invoking a pump–probe scheme, we recently demonstrated acquisition of spectral fingerprints of substances from a variable standoff distance.[1] The experimental details have been reported previously.[1] In brief, the experimental system is composed of a tunable IR laser (Daylight Solutions, Inc., Poway, CA) for the pumping action and a low power visible or near IR laser for the probing action. The pump laser is of the quantum cascade type having a spectral tunability of 9.26 to 9.8 μm (1079.91 to 1020.41 cm^{-1}) and modulating at a frequency within the thermal response of the sample, *e.g.* 50 Hz. The probe laser, a 632.8 nm HeNe, measures the surface's absorption of the pump light *via* lock-in amplification at the modulation frequency.

In this study we used Parafilm wax (Pechiney Plastic Packaging Co., Chicago, IL), a proprietary formulation of approximately 50% paraffin wax and 50% polyolefins, having no major absorption peaks in the quantum cascade laser (QCL) range. However, the high sensitivity of the pump–probe measurement allowed us to detect small absorbance patterns that changed between unheated and melted wax. Unheated wax film was affixed to a n-type silicon (Si) wafer substrate and spectroscopically measured with the pump–probe method and a commercial Fourier Transform Infrared (FTIR) spectrometer (Spectrum GX, Perkin Elmer); the results of the comparative spectral analysis are shown in Fig. 6. The averaged (11 measurements) pump–probe data is corrected for wavelength-dependent QCL output, but is otherwise unfiltered. Background-correction for the Si wafer substrate is included in the FTIR data.

The intensity corrected absorption spectra presented in Fig. 6 can be considered as matched well to the reference spectra obtained from FTIR spectrometer. The pump–probe measured data are also complemented by the high level of random fluctuations that obstruct the systematic pattern whose components should be identified. The noise sources include the environmental temperature changes and air movement in the laboratory, the mechanical vibrations of the instrument components, and stray light on the photodetector. Recognizing the systematic pattern in the presence of random fluctuations is often challenging requiring sophisticated data processing techniques.

3.1 Spectral and correlation analysis

Standoff detection techniques can be particularly sensitive to the environmental and background noises leading to the power of the noise being higher than the power of the useful pattern.

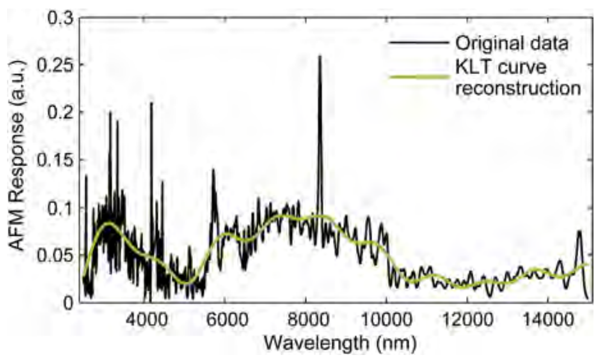

Fig. 5 Photoacoustic force microscopy and KLT reconstructed data using 20 eigenvalues out of 2000. Probe response in the plant cell corner middle lamella region of biomass cross section, indicates lignin content. The reconstructed data, resolving all the salient spectral features while excluding the random spikes, was computed employing only 20 eigenvalues resolve.

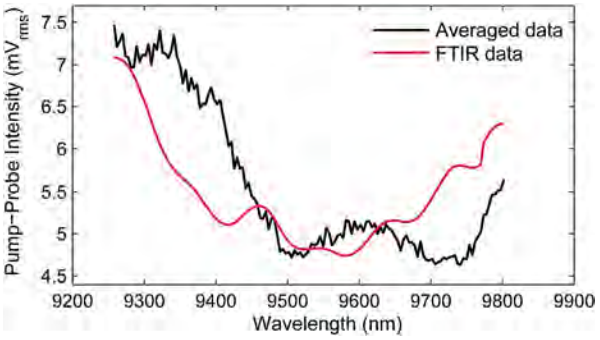

Fig. 6 Photothermal absorption spectroscopy. Averaged pump–probe spectra of 11 measurements for unheated wax film is compared with an FTIR measurement.

The detection and identification of the systematic pattern based on a single measurement vector cannot be considered reliable but can be improved in a multi-observation approach, where decision making is carried out by processing of M observations (reducing the power of random component by a factor M). Here, we aim to process and analyze the data using statistical analysis involving the spectral and correlation analysis, filtering, and Karhunen–Loève orthogonal transform (KLT).

We present our measurements as a discrete series $\mathbf{S}_i(n)$, $n = 1, 2, \ldots, N$, and $i = 1, 2, \ldots, M$, where the wavelength variable, λ, has been substituted by an abstract variable n, which counts the elements in the vector. The classical time series analysis allows the decomposition of the data into four main components: (1) trend – non-periodic tendency of the data over a long period of time, (2) cyclic components – trend's low frequency oscillations, (3) periodic components – narrow band harmonics on a wide range of frequencies, (4) and random noise – unpredictable oscillation with an autocovariance function that asymptotically approaches the autocovariance function of additive white Gaussian noise (AWGN) that is unity if zero lag, and zero for any nonzero lag.[34]

Selected absorption spectra are presented in Fig. 6. Limiting our analysis to the corrected absorption spectra for latex, eleven series were obtained ($M = 11$) with an average $\mathbf{S}_{\text{ave}} = \frac{1}{11}\sum_{i=1}^{11}\mathbf{S}_i$ as shown, along with the FTIR data, in Fig. 6. The average power spectral density (PSD), or periodogram, is found to be $\left|\hat{\mathbf{S}}_{\text{ave}}\right|^2 = \frac{1}{11}\sum_{i=1}^{11}\left|\hat{\mathbf{S}}_i\right|^2$, where $\hat{\mathbf{S}}_i = \text{FFT}\{\mathbf{S}_i\}$. The resulting periodogram exhibit strong peaks at zero frequency as well as at 0.0078 Hz, corresponding to the trend and trend's low frequency oscillations. Then, applying sequentially detrending and moving average (MA) filtering[32] with the edge frequency 0.05 Hz, those components can successfully be removed as displayed in Fig. 7(a) (black curve), showing deep suppression in the range 0.00–0.05 Hz. Afterward, the fundamental frequencies of the higher amplitude harmonics are estimated as (in descending order) – 0.051 Hz, 0.087 Hz, 0.150 Hz, 0.173 Hz, 0.319 Hz, 0.425 Hz, 0.449 Hz, and 0.488 Hz, which are used as the central frequencies of band stop filters designed to reject the mentioned harmonics (prewhitening). The IIR Butterworth band stop filters of 7th order (stable implementation) with the edge frequencies $f_{\min}$ and $f_{\max}$ were then designed, where $f_{\min}$ and $f_{\max}$ is chosen in the range of each peak width.

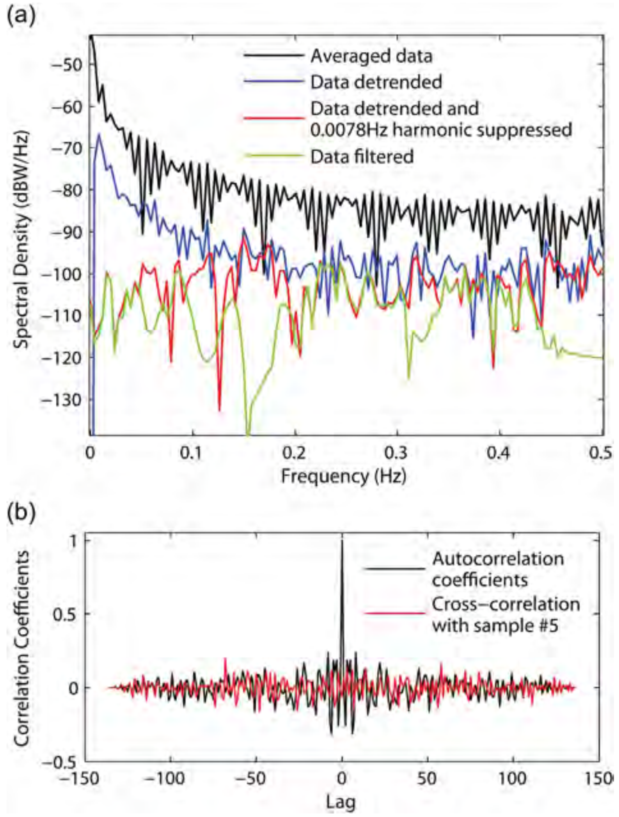

Fig. 7 (a) The effect of the removal of seven higher amplitude harmonics on the power spectral density, or periodogram. (b) Autocorrelation and the cross correlation after filtering of seven higher amplitude harmonics using sample #5.

The main challenge is how many harmonic components should be suppressed to achieve the complete randomness of the remaining sequence? The hypotheses testing approach can be helpful, which is based on the fact that when an analyzed sequence is a sampling of a purely random process and the sample vector length is very long, then all theoretical lagged correlation coefficients are zero except at lag zero, *i.e.*, the tested random process is AWGN. However, when the series length is not so long (137 in our case), the approximate threshold, or critical level of correlation coefficients that helps to accept or reject the null hypothesis (the sequence represents the random process with zero correlation) should be found. For example, if significance interval is 95%, then the critical level is $-1/N \pm 2/\sqrt{N}$.[35] For $N = 137$ the 95% confidence interval limits are approximately ± 0.18. Using the band-stop filters we suppress one by one the periodical components in descending order of amplitudes and after each suppressed component we performed the hypothesis testing for the 95% significance interval. The resulting correlogram (graph of autocorrelation coefficients level *versus* lags) after suppression of seven larger components is presented in Fig. 7(b) (black curve). The corresponding plot of the periodogram after filtering those components is also presented in Fig. 7(b) (green curve).

Even though two of the correlation coefficients in Fig. 7(b) are outside of the significance limits (at the lag 3, 5, and 8) there is no reason to reject the hypothesis that the sequence data are independently distributed (due to the length not being long). If hypothesis is accepted then the power of random noise is found as $\sigma_{\text{noise}}{}^2 = \dfrac{1}{N-1}\sum_{i=1}^{N}[\mathbf{S}_{\text{ave}}(n) - \mu]^2$, where μ is an expectation value of data after suppression of seven larger components, and the result is $\mu = -9.6858 \times 10^{-8}$, and $\sigma_{\text{noise}}{}^2 = 6.2219 \times 10^{-9}$. Fig. 7(b) also depicts the cross-correlation coefficients *versus* lags between the prewhitening series #5 and the average series with subtracted sample #5. A similar analysis confirms the randomness and independence of the different series after prewhitening. Decomposing the data sequence with spectral analysis allows reconstruction of the original systematic pattern by properly combining the obtained decomposition components. The random noise as well as deterministic instrumental error components can be successfully suppressed, as Fig. 8 shows.

Particularly, Fig. 8(c) shows that after a major part of the noise was filtered the reconstructed data still contains strong fluctuated components. It means that the high correlated instrumental noise still complements the data. The nature of this noise has been described in ref. 1. Fig. 8(a) depicts the data pattern after suppression of instrumental noise, and resulting curve (green) can be considered as a good approximation of the FTIR curve, shown in Fig. 6.

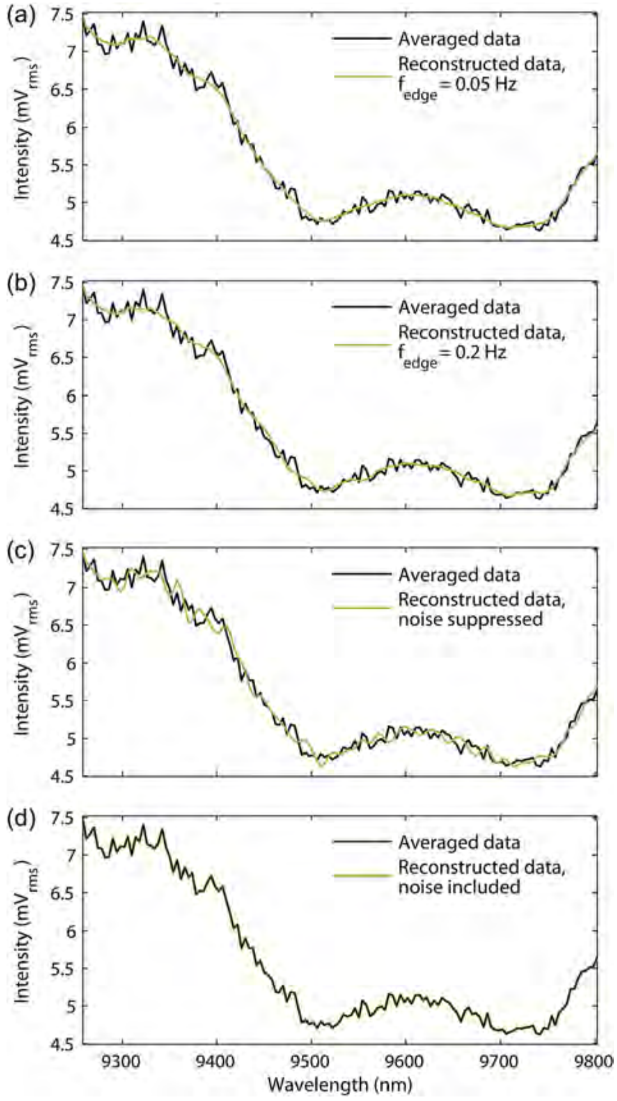

Fig. 8 Data reconstruction using periodogram, (a) including the trend and all harmonics up to a frequency of 0.05 Hz, (b) including the trend and all harmonics up to a frequency of 0.2 Hz, (c) including all the components in the range 0.0–0.5 Hz excluding the random noise, and (d) including all the components in the range 0.0–0.5 Hz including the random noise.

3.2 Denoising with KLT

The spectral and correlation analysis presented in the previous subsection is a simple and non-expensive computational method. However, drawbacks such as low resolution of the spectral components, slow convergence of the Fourier series when the data vector is short, and distortion of periodogram after prewhitening using non-ideal filters, motivates alternative and superior denoising algorithms. Here we show that the KLT overcomes all drawbacks of the FFT.

We begin our demonstration with a small sample size of $M = 11$. Ideally, the number of available realizations should at least equal the number of data points (requiring enormous numbers of measurements and computations), otherwise the matrix $\mathbf{R}$ is singular. However, in our case the process can be handled if the true covariance matrix $\mathbf{R}$ (it is asymptotic covariance matrix when $M \to \infty$) could be predicted. Various types of random processes and their covariance matrices have been analyzed.[36] Analyzing the available data, we hypothesize that the data represents the Wiener process and the appropriate eigenfunctions are just pure sinusoidal functions (noting that such functions also serve successfully for many other random processes, including white Gaussian noise).[37] These observations help us to apply the asymptotic matrix $\mathbf{R}$ of the Wiener process to our KLT algorithm. This matrix forming routine has been described in detail.[36] The KLT denoising procedure over the averaged data sample (Fig. 6) involving different numbers of significant eigenvalues is presented in Fig. 9.

To estimate the threshold between the subspace of the useful data and the subspace of the unwanted noise, we find the eigenvalues of the covariance matrix. Since, $N = 137$, the covariance matrix has a dimension 137 × 137 and a rank $r = 11$, hence it is degenerated because $r < N$.[32] To be able to apply the eigenvalue decomposition, the matrix should have a full rank, *i.e.*, $r = N$. For this, the degenerated covariance matrix was regularized by adding the noise power value into diagonal elements creating intentionally $r = N$ where the degenerated matrix becomes invertible; hence the eigenvalue decomposition can be performed. The resulting distribution of eigenvalues is presented below in Table 1, showing that the eigenvalues have

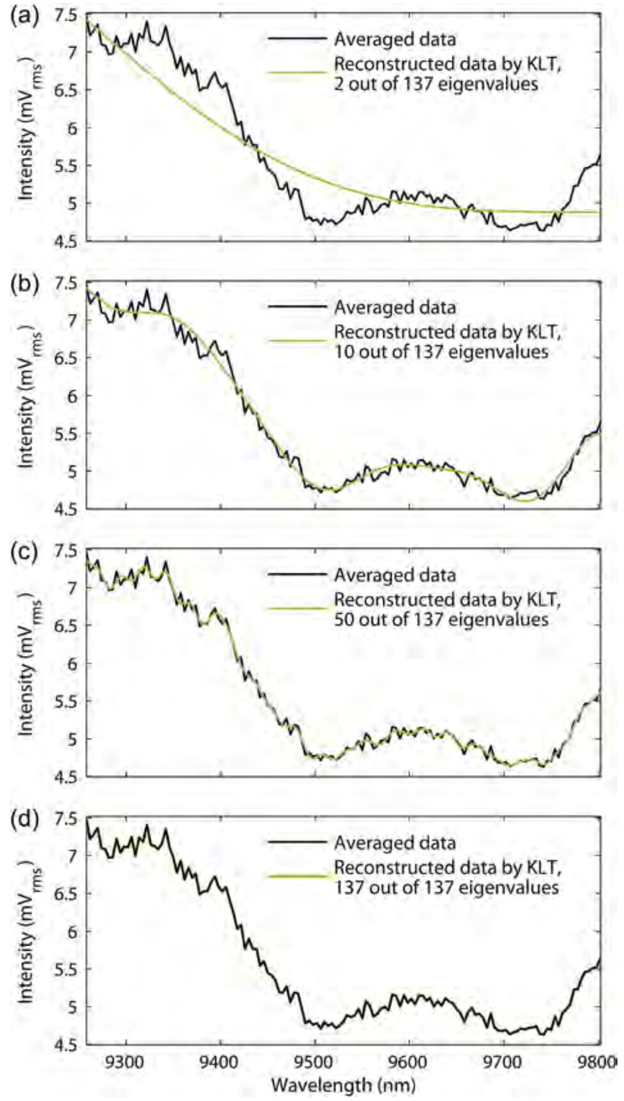

Fig. 9 The effect of the number of eigenvalues on data reconstruction employing KLT. Out of 137 significant eigenvalues, employing (a) 2, (b) 10, (c) 50 and (d) all 137 eigenvalues.

Table 1 Distribution of eigenvalues of sampled correlation matrix

Eigenvalue	Numerical value	Eigenvalue	Numerical value
λ_1	3.0505×10^{-6}	λ_7	5.8191×10^{-7}
λ_2	2.7596×10^{-6}	λ_8	4.4572×10^{-7}
λ_3	1.6693×10^{-6}	λ_9	3.5650×10^{-7}
λ_4	1.3818×10^{-6}	λ_{10}	2.4938×10^{-7}
λ_5	9.8285×10^{-7}	λ_{11}	6.2219×10^{-9}
λ_6	7.7259×10^{-7}	λ_{12}	6.2161×10^{-9}

converged to the noise power we found in the previous subsection. Hence, it is more likely that the first ten eigenvalues could be considered as significant, *i.e.*, $m = 10$. The pattern reconstruction with ten significant eigenvalues presented in Fig. 9(b) is well matched with result depicted in Fig. 8(a). The 11th eigenvalue is significantly less than the 10th, hence, 10 eigenvalues can be certainly considered as belonging to the useful pattern set. We interpret that these figures likely represent the systematic pattern of our measurement.

4 Conclusions

In conclusion, we showed that the measured Brownian or otherwise thermal actuation of microcantilevers and analytes can be advantageously treated with the KLT. The KLT was demonstrated to effectively improve both the spectral identification of the microcantilever resonances that were driven by the Brownian fluctuation, and the infrared spectral data that contained random fluctuation of thermal, mechanical and electronic origin. The result of the analysis illustrates that KLT can be adapted as a powerful data denoising tool for the presented pump–probe infrared standoff spectroscopy and cantilever based sensing applications.

Acknowledgements

This work was in part supported by BESC and in part by the laboratory directed research and development (LDRD) fund at Oak Ridge National Laboratory (ORNL). Viktor Zaharov acknowledges the financial support received from the Department of Energy (DOE) Visiting Faculty Program (VFP). The BESC is a US Department of Energy (DOE) Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. ORNL is managed by UT-Battelle, LLC, for the US DOE under contract DE-AC05-00OR22725.

References

1 R. H. Farahi, A. Passian, Y. K. Jones, L. Tetard, A. L. Lereu and T. G. Thundat, *J. Phys. D: Appl. Phys.*, 2012, **45**, 125101.

2 C. W. Van Neste, L. R. Senesac and T. Thundat, *Anal. Chem.*, 2009, **81**, 1952–1956.

3 R. M. Measures, *Laser Remote Sensing: Fundamentals and Applications*, Wiley-Interscience, 1984.

4 A. Labuda, J. R. Bates and P. H. Grütter, *Nanotechnology*, 2012, **23**, 025503.

5 H. Kawakatsu, S. Kawai, D. Saya, M. Nagashio, D. Kobayashi, H. Toshiyoshi and H. Fujita, *Rev. Sci. Instrum.*, 2002, **73**, 2317–2320.

6 J. Kottmann, U. Grob, J. M. Rey and M. W. Sigrist, *Sensors*, 2013, **13**, 535–549.

7 K. Woerle, F. Seichter, A. Wilk, C. Armacost, T. Day, M. Godejohann, U. Wachter, J. Vogt, P. Radermacher and B. Mizaikoff, *Anal. Chem.*, 2013, **85**, 2697–2702.

8 H. F. Ji, T. Thundat, R. Dabestani, G. M. Brown, P. F. Britt and P. V. Bonnesen, *Anal. Chem.*, 2001, **73**, 1572–1576.

9 L. Tetard, A. Passian, R. H. Farahi, B. H. Davison and T. Thundat, *Opt. Lett.*, 2011, **36**, 3251–3253.

10 C. Maccone, *Acta Astronaut.*, 2007, **60**, 775–779.

11 C. Maccone, *Acta Astronaut.*, 2010, **67**, 1427–1439.

12 K. Karhunen, *Ann. Acad. Sci. Fenn., Ser. A1: Math.–Phys.*, 1947, 1–79.

13 M. Loève, *Probability theory. Graduate Texts in Mathematics*, Springer-Verlag, 4th edn, 1978, vol. 2.

14 S. L. M. Freire and T. J. Ulrych, *Geophysics*, 1988, **53**, 778–785.

15 K. M. Al-Yahya, *Geophysical Prospecting*, 1991, **39**, 77–93.

16 Y. Ding, Y. C. Chung, S. V. Raman and O. P. Simonetti, *Phys. Med. Biol.*, 2009, **54**, 3909–3922.

17 V. G. Stamatopoulos, D. A. Karras and B. G. Mertzios, *Meas. Sci. Technol.*, 2009, **20**, 104021.

18 J. Chen, J. Benesty and Y. A. Huang, *IEEE Transactions on Audio, Speech, and Language Processing*, 2009, **17**, 787–802.

19 U. Mittal and N. Phamdo, *IEEE Transactions on Speech and Audio Processing*, 2000, **8**, 159–167.

20 S. K. Jha and R. D. S. Yadava, *IEEE Sens. J.*, 2011, **11**, 35–44.

21 J. J. Gerbrands, *Pattern Recognit.*, 1981, **14**, 375–381.

22 K. B. Andersen, N. O. Christiansen, J. Castillo-Leon, N. Rozlosnik and W. E. Svendsen, *Org. Electron.*, 2013, **14**, 1370–1375.

23 F. Long, A. Zhu, H. Shi and H. Wang, *Anal. Chem.*, 2014, **86**, 2862–2866.

24 M. Pfeifer, A. Ruf and P. Fischer, *Opt. Express*, 2013, **21**, 25643–25654.

25 R. van den Hurk and S. Evoy, *Sens. Actuators, B*, 2013, **176**, 960–965.

26 K. Woerle, F. Seichter, A. Wilk, C. Armacost, T. Day, M. Godejohann, U. Wachter, J. Vogt, P. Radermacher and B. Mizaikoff, *Anal. Chem.*, 2013, **85**, 2697–2702.

27 C. Yao, F. Cheng, C. Wang, Y. Wang, X. Guo, Z. Gong, M. Fan and Z. Zhang, *Anal. Methods*, 2013, **5**, 5560–5564.

28 A. Haché, P. A. Do and S. Bonora, *Appl. Opt.*, 2012, **51**, 6578–6585.

29 T. R. Albrecht, P. Grütter, D. Horne and D. Rugar, *J. Appl. Phys.*, 1991, **69**, 668–673.

30 A. Passian, V. Protopopescu and T. T., *J. Appl. Phys.*, 2006, **100**, 114314.

31 R. Wang, *Introduction to Orthogonal Transforms: With Applications in Data Processing and Analysis*, Cambridge University Press, 2012.

32 S. L. Marple, *Digital Spectral Analysis with Application*, Prentice-Hal, 1987.

33 http://www.mathworks.com/products/matlab/.

34 C. Chatfield, *The analysis of time series*, Chapman & Hall/CRC, 2004.

35 M. G. Kendall, A. Stuart and J. K. Ord, *The advanced theory of statistic*, Griffin, 1977.

36 C. Maccone, *Deep Space Flight And Communication*, Springer, 2009.

37 J. B. Burl, *IEEE Trans. Acoust., Speech., Signal Process.*, 1989, **37**, 99–105.