Article

Applied Spectroscopy

Applied Spectroscopy
2016, Vol. 70(11) 1861–1871

DOI: 10.1177/0003702816671065
asp.sagepub.com

SAGE

Automatic Spike-Removal Algorithm for Raman Spectra

Yao Tian[1] and Kenneth S. Burch[2]

Abstract

Raman spectroscopy is a powerful technique, widely used in both academia and industry. In part, the technique's extensive use stems from its ability to uniquely identify and image various material parameters: composition, strain, temperature, lattice/excitation symmetry, and magnetism in bulk, nano, solid, and organic materials. However, in nanomaterials and samples with low thermal conductivity, these measurements require long acquisition times. On the other hand, charge-coupled device (CCD) detectors used in Raman microscopes are vulnerable to cosmic rays. As a result, many spurious spikes occur in the measured spectra, which can distort the result or require the spectra to be ignored. In this paper, we outline a new method that significantly improves upon existing algorithms for removing these spikes. Specifically, we employ wavelet transform and data clustering in a new spike-removal algorithm. This algorithm results in spike-free spectra with negligible spectral distortion. The reduced dependence on the selection of wavelets and intuitive wavelet coefficient adjustment strategy enables non-experts to employ these powerful spectra-filtering techniques.

Keywords

Raman, wavelet transform, despike

Date received: 13 August 2015; accepted: 16 March 2016

Introduction

Raman microspectroscopy has been applied to wide-ranging areas including chemistry,[1] physics,[2–11] materials science,[12] and biology[13] because of its simple instrumental structure and wide range of information provided by Raman spectra. However, Raman measurements are often limited to low signal levels and thus long integration times are necessary. In principle this would only lead to increased statistical noise, which could ultimately be averaged out. However, for dispersive spectrometers, charge-coupled devices (CCD) are widely employed because of their unique advantages such as high quantum efficiency, great sensitivity, high dynamic range, linear response to photons, small thermal/readout noise, and high reliability. Unfortunately, these detectors are highly vulnerable to cosmic rays, resulting in extremely large spikes in the data, where the signal level on a single/few pixels becomes many orders of magnitude larger than the measured spectra. The majority of cosmic rays are muons and most of the remainder is protons and neutrons.[14] These high energy particles collide and interact electromagnetically with materials on a CCD chip causing ionization and atomic or collective excitations and thus cannot be easily blocked by shielding.[15] Such events typically generate a charge of at least several thousand electrons on a single pixel or over a few consecutive pixels of a CCD detector,[16] leading to spurious, comparatively narrow spikes.[17] Moreover, given the nature of the highly stochastic collision process, these spikes are distributed randomly both in time and space causing further complexity for spike-removal. To solve this problem, many approaches have been taken, including both software and hardware based implementations. For example, an image curvature correction method was employed to improve the optical hardware of a spectrometer.[18] However, this is more instrumentally complex and costly. Therefore, for most Raman applications, software approaches are more widely applied.

Many algorithms have been proposed for different applications based on either a single scan or multiple scan mechanism. The single scan methods include smoothing, weighted

[1]Department of Physics, and Institute of Optical Sciences University of Toronto, Toronto, Ontario, Canada
[2]Department of Physics, Boston College, Chestnut Hill, Massachusetts, USA

Corresponding author:
Kenneth S. Burch, Department of Physics, Boston College, 140 Commonwealth Ave, Chestnut Hill, MA 02467-3804, USA.
Email: ks.burch@bc.edu

moving window filtering,[19] wavelet transform based filtering[16,20] and attempts to fit spikes with predefined profiles.[21] These algorithms rely on the assumption that the amplitudes of spikes are much higher and/or linewidths narrower than real Raman features. Beyond the sometimes limited validity of this assumption, these algorithms also require a deep knowledge of spectral filtering such that proper threshold settings can be chosen to minimize distortion. Therefore, taking the random nature of the cosmic rays into account, multiple scan methods, to some extent, can overcome the above drawbacks through a comparison between consecutive scans. For example, the upper-bound spectrum method,[22] second difference comparison,[23] and time domain comparison[24] have received a lot of attention and some of them are even fully automated. Nevertheless, these methods also suffer from the difficulty of properly choosing the numerous parameters involved or failure to explore the different local frequency characteristics between spikes and real features. As such, these methods have not found widespread use due to the difficulty of implementing them, inability to properly detect all spikes, and/or their tendency to distort the spectra.

To overcome these difficulties, we have devised a new algorithm that combines wavelet transforms with data clustering methods to automatically detect and remove spikes from Raman spectra. Specifically, based on the randomness of cosmic rays, spike detection was ensured by a clustering of wavelet coefficients method. By analyzing the clustering behavior, the erroneous coefficients can be reset to the most probable value. A multiresolution analysis is also employed to enable separation of spikes and real features by different local frequency characteristics. This approach ensures the preservation of real Raman features of both broad and narrow profiles, as well as insensitivity to spike amplitude. Furthermore, this automatic detection enables easy implementation, provides a more intuitive threshold setting, and reduces dependency on the particular wavelets employed. Although our method is not fully automated, the algorithm is capable of removing spikes in different contexts and preserving the real Raman features well. As such our method is easily implemented by those not familiar with spectral filtering algorithms.

Theory

In this section, we outline the theory behind the two techniques used in our algorithm.

Wavelet Transform

Fourier transforms are well known for their exceptional ability to reveal the frequency composition of a series $x(t)$ and thus remove periodic noise. However, because of its extended and periodic basis functions (sine and cosine), localized features in $x(t)$ will strongly overlap in the frequency domain after transform. Thus Fourier filters are unable to remove localized features, and as such are not appropriate for separating spectral spikes from Raman features, both of which are generally localized and not periodic. To solve this problem, some authors have turned to wavelet transforms.[25]

In wavelet transforms one represents localized and non-stationary signals by a set of functions called wavelets. Wavelets are a series of functions that are all localized, quickly decaying, and can be translated and scaled to form a complete basis. To gain more insight into wavelet transforms, let us consider the following formula which is used to obtain continuous wavelet transform coefficients,

$$c(p,q) = \int x(t)\Psi_{p,q}(t)\mathrm{d}t \tag{1}$$

$$\Psi_{p,q}(t) = \frac{1}{\sqrt{|p|}}\Psi\left(\frac{t-q}{p}\right) \tag{2}$$

Here $x(t)$ is the original signal, $\Psi_{p,q}(t)$ is a series of wavelet functions generated by the scaling and translation of the mother wavelet $\Psi(t)$, p characterizes the range and local frequency (here local frequency represents how fast the signal changes locally), and q translates the center of the wavelet. An example is shown in Figure 1 where one can see that as the scaling factor (p) decreases, the wavelets become more localized with a decreasing period of oscillation. Thus, as p is reduced, higher local frequencies are sampled, and only a smaller section near $x(q)$ is included in Eq. 1. In other words, $c(p, q)$ is capable of extracting the local frequency component of $x(t)$. The frequency window (bandwidth) is controlled by the scaling factor p whose reduction leads to the extraction of higher frequency/more local components. For a continuous transformation, p and q can take any positive value. From the perspective of reduction of redundant information and for real applications, discrete wavelet transforms are preferred where the values taken by p, q normally have the following relation

$$p = 2^j \quad q = 2^j - k \quad k, j \in \mathbb{Z} \tag{3}$$

where wavelets $\Psi_{p,q}(t)$ are denoted by $\Psi_{j,k}(t)$ instead.[26]

This natural ability of wavelets to separate out information at various degrees of local variation makes them a natural choice for removing cosmic ray induced spikes in Raman spectra. Specifically, real features are typically broader than spikes leading to a natural separation into different local frequency bands. Therefore, most of the real Raman features will appear in lower local frequency components (i.e., higher j) than spikes. If features in a series $x(t)$ have different local frequency characteristics, a

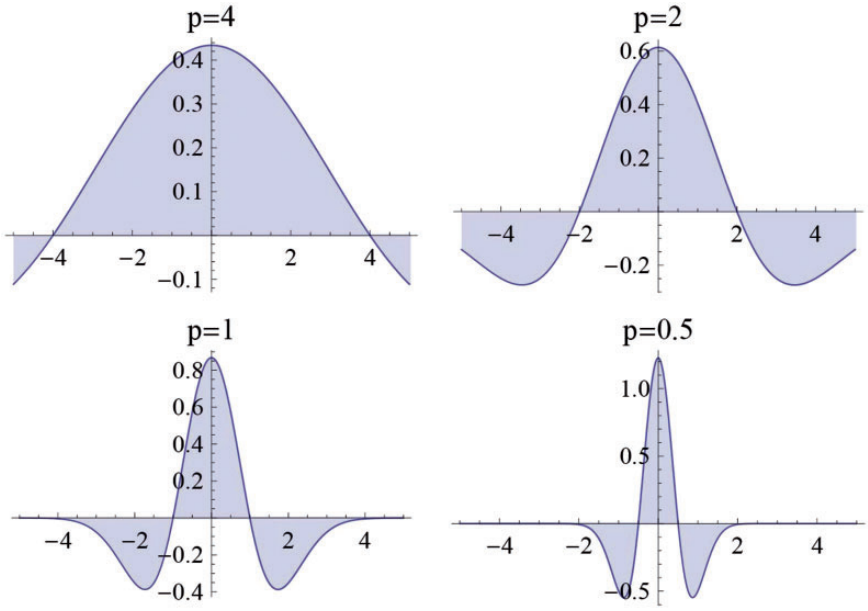

Figure 1. Mexican hat wavelet functions with four different scaling factors.

multiresolution analysis can be achieved by discrete wavelet transform.

Multiresolution analysis was first proposed by Mallat[27] in his research of computer vision. The basic idea is to decompose signals under different scales (local frequencies). Through analysis and comparison of the decomposition coefficients, one can obtain useful information. Mathematically speaking, $x(t)$ is firstly projected onto a series of orthogonal subspaces that are spanned by the wavelet functions $\Psi_{j,k}(t)$ ($j \in [0, j_n]$ where j_n is the number of steps in wavelet transform). To satisfy the requirement of completeness, $x(t)$ is also projected to the space of spanned by the scaling functions $\Phi_{j_n,k}(t)$ which is the orthogonal complementary space of the spaces spanned by $\Psi_{j,k}(t)$. Each of these spaces has its unique local frequency characteristics. The coefficients obtained by projecting $x(t)$ on $\Psi_{j,k}(t)$ ($\Phi_{j_n,k}(t)$) are named as $d_{j,k}$ ($a_{j,k}$). d and a are abbreviations for detail and approximation due to the local frequency characteristics of g and h explained below. The same as continuous wavelet transform, as j increases, the more coarse (low local frequency) information is represented. More details about the multi-resolution analysis and the relation between Ψ and Φ can be found in the supplemental materials. Readers interested in the mathematical construction procedure of wavelet function Ψ and scaling function Φ are referred to Daubechies et al.,[26] where the detailed mathematical background is discussed. From the perspective of signal filtering, the wavelet functions $\Psi_{j,k}(t)$ are determined by a series of high-pass filters (g) and the complementary scaling functions $\Phi_{j,k}(t)$ are related to low-pass filters (h).[26] If the original spectrum can be considered as the approximation at level 0 by the a_0 coefficients, then the Mallat algorithm tells us the hierarchical coefficients $a_{j+1,k}$ ($d_{j+1,k}$) can be obtained by deconvolution of $a_{j,k}$ with filters h (g) and subsequent down-sampling.[27] Thus, $d_{j,k}$ ($a_{j,k}$) represents the information of the original signals in different local frequency bands. High local frequency information is stored in low level (small j) detail coefficients and vice versa. Once the coefficients are obtained, one can then analyze the signal at different resolution levels. This shows a clear advantage for the wavelet approach, namely the local distortion can be removed without distorting the signal resulting from non-local components. However, a key difficulty in this approach is finding a reliable method for adjusting the coefficients at each level, such that only the spikes are removed. The solution to this is described in the section via a data clustering technique.

K-Means Clustering

In our multiresolution analysis, a data clustering technique is employed to distinguish wavelet coefficients originating from spikes from real signals. The most widely used clustering algorithm is k-means clustering. K-means clustering is designed to partition n data points into K clusters in which each point belongs to the cluster with the nearest mean. The goal of the algorithm is to choose the cluster centroid (c_l) to minimize total intra-cluster variance, while maintaining maximum distinction between the clusters. Thus the cost function in Euclidean space, which represents the variant, is just

the sum of the square distance of each data point to its corresponding centroid, given by Bishop et al.[28]:

$$J = \sum_{l=1}^{K} \sum_{i=1}^{l_n} |x_i^l - c_l|^2 \quad (4)$$

where c_l is the centroid for cluster l and l_n is the number of data points in the cluster and x_i^l is a data point in cluster l. A more detailed description can be found in Smola and Schölkopf.[29] Nonetheless, a difficulty with this approach is that the best number of clusters is not known and has to be determined for each application of the algorithm. Another downside of k-means algorithms are that the cost function is not concave, leading to the production of local minimums, and the outcome strongly relies on the initial guess. So as to achieve a global minimum, one typically runs the k-means algorithm multiple times with different initial guesses and chooses the one with minimum J.

Spike-Removal Algorithm

The basic idea behind our algorithm is to adjust the erroneous wavelet coefficients caused by spikes ($a_{j_n,k}^m$ and $d_{j,k}^m$, m indicates the specific spectra being transformed) to the values that they are most likely to be. As mentioned above, the locations of the spikes are highly random leading to very low probability that two spikes are located at the same position in all spectra. Consequently, most of the coefficients representing the real data should aggregate into a cluster. Erroneous coefficients that appear with low probability can be detected by analysis of clustering behaviors. Then, the wrong one can be adjusted to the average of the data points in an aggregated cluster. We proceed by assuming three or more spectra (h is the total number of spectra recorded) were measured under the same conditions for a given sample. The algorithm consists of four steps:

(1) Wavelet transforms up to level n are performed on each spectrum individually. The optimal n is determined by the relation: $2^{n+3} \approx$ number of data points. The resulting $a_{j_n,k}^m$ and $d_{j,k}^m$ ($j \in [0, n]$, $m \in [1, h]$) are stored in array $l_{j,k}^m$. With

$$l_{j,k}^m = d_{j,k}^m \; j \in (0, n) \; l_{n+1,k}^m = a_{j_n,k}^m$$

where j denotes the level of wavelet coefficients and k represents the translation.

(2) Set a cluster radius r_j for each decomposition level j to allow for variance among the h spectra. Theoretically, the measured spectra should be very similar resulting in no small difference in coefficients. Thus, if one clusters the coefficients, all the coefficients should aggregate into one cluster except the erroneous coefficients caused by spikes. In practice, there will always be some environmental changes or drifts (e.g., sample temperature or laser power) leading to variances between the consecutive measurements. As mentioned in the previous section, Raman features are usually broader than spikes and thus have different local frequency characteristics. Thus it is safe to infer that $l_{j,k}^m$ at higher level j better represents the real features. Hence, the cluster radius at a high level should be set larger to account for the variance among spectra. Detailed discussion about selection of r_j will be given in the subsection "Selection of r_j" where r_j will be set as different functions of j to minimize spectra distortion.

(3) In this step, the erroneous wavelet coefficients $l_{j,k}^m$ caused by spikes are adjusted. These coefficients are auto-detected through a clustering search. As mentioned above, the optimal number of clusters is unknown in advance. Thus an algorithm is employed to search for the best number of clusters so as to detect all the erroneous coefficients and exclude real features. The number of clusters c_{num} starts with one and is iterated until the

Data: The decomposition coefficients $l_{j,k}$, total number of spectra h, current cluster number c_{num}, radius of the largest cluster $Rc_{j_m,k}$, centroid of the largest cluster $C_{j_m,k}$, preset cluster radius r_j

```
for each j, k do
    set c_num =1;
    while True do
        use K-means method to aggregate l_{j,k} into c_num clusters.;
        get the radius Rc_{j_m,k} and centroid C_{j_m,k} of the largest cluster, if there are more than one largest cluster, choose the one with smallest radius;
        if Rc_{j_m,k} <= r_j then
            reset the l^m_{j,k} in other clusters to C_{j_m,k};
            break to next for loop;
        else
            c_num = c_num +1;
            if c_num==h then
                no change will be made,break to next for loop;
            else
                continue the while loop with updates c_num;
            end
        end
    end
end
```

Algorithm 1. The algorithm of Step 3.

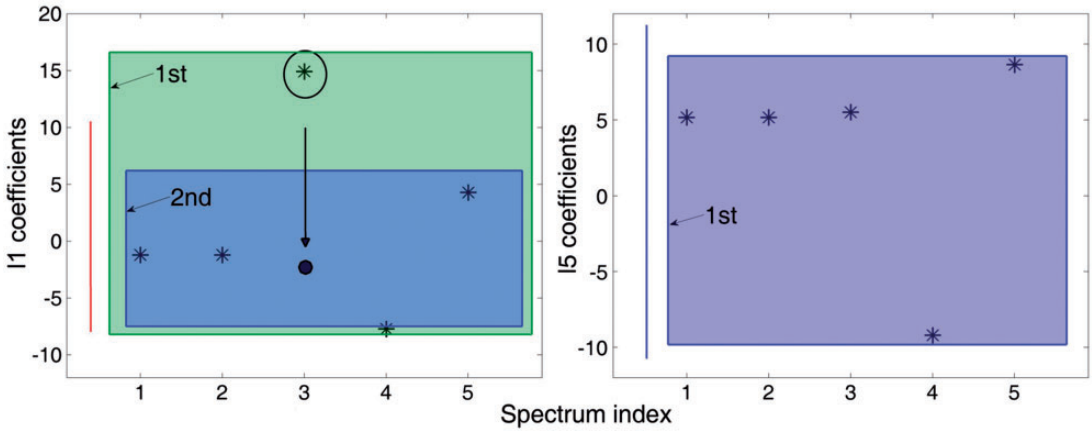

Figure 2. One example of step 3. (a) Clustering performed at $j = 1$. $2*r_1$ is visualized by the red line. In the first trial, the algorithm aggregated all data points into one cluster (the green rectangle). Since the radius (height) of cluster (green) is larger than $2*r_1$, the algorithm continues. In the next trial, the data points were aggregated into two clusters. The radius (height) of the bigger cluster (the blue rectangle) is smaller than $2*r_1$. Thus the algorithm accepts this aggregation and resets the wrong coefficients to the centroid of the bigger cluster (shown by the arrow). (b) Clustering performed at $j = 5$. $2*r_5$ is denoted by the purple line. Since $r_5 > r_1$, one cluster was found and its diameter (the short edge of the purple rectangle) was accepted by the algorithm and no change was made.

appropriate cluster number c_{num} is found. This can be achieved by comparing the radius of the largest cluster to the preset r_j. If the radius is smaller than r_j, all coefficients in the largest cluster will be labeled as the coefficients from real features and those not in the cluster will be detected as erroneous coefficients. Then, all erroneous coefficients will be reset to the centroid of the cluster. If not, the algorithm will try to find $c_{num} + 1$ clusters and repeat the aggregation procedure. On other hand, if no cluster can be found, which means the algorithm is clustering h coefficients to h clusters, in this case, it is likely all coefficients are real but with large variance, so no change will be made. More details of this step can be referred to the pseudocode (Alogrithm 1). In Figure 2, one example is also shown to illustrate this step.

(4) In the end, the processed coefficients $a^m_{j_n,k}$ and $d^m_{j,k}$ will be used to perform an inverse wavelet transform to recover the h spectra with the spikes removed

Experiments

The Raman measurements were performed using a home-built Raman microscope, the details of which can be found in Tian et al.[30] Two single crystal samples, $Cr_2Ge_2Te_6$ and $Sr_3Ir_2O_7$, were measured, chosen for their very small Raman cross-section, low thermal conductivities, and variety of Raman features (two-phonon as well as two-magnon). Thus in order to achieve sufficient signal to noise ratios as well as avoid laser heating, one has to use small laser excitation powers and long exposure times, increasing the possibility of spikes in the spectra. For the clustering purpose, at least three acquisitions taken with the same exposure time, laser power, and under the same physical conditions are required.

The algorithm was implemented in Matlab. Matlab 2014a built-in wavelet transform, its inverse and k-means clustering functions were used. The code "Spike Removal for Raman Spectra" can be downloaded at http://www.mathworks.com/.

Results and Discussion

Spikes in Raman spectra are typically very narrow (full width at half-maximum, (FWHM) 1.5–3 cm^{-1});[16] however, the features they interfere with are of various types. In this section, we show the processed results on four data sets with different characteristics. Subsequently, the influence of the selection of r_j and different wavelets on the performance of the algorithm will be discussed.

Primary Tests

The first two sets of spectra were taken on single crystal $Sr_3Ir_2O_7$ at 280 K and 125 K (shown in Figure 3). The measurement was performed under laser power as low as 0.1 mW with 1 μm spot size. Each spectrum was collected with 10 min acquisition. The processed spectral range was set to (–200,1600) cm^{-1} to show both anti-Stokes and Stokes sides, resulting in 1803 data points. The central multiple peaks near 0 cm^{-1} are Rayleigh scattering and artifacts from our notch filters. The Raman spectra of $Sr_3Ir_2O_7$ contain five phonons[31] (centered at 143.0 cm^{-1}, 177.6 cm^{-1}, 268.4 cm^{-1}, 389.4 cm^{-1}, and 586.0 cm^{-1}) and two broad features (710–880 cm^{-1}, 1307–1462 cm^{-1}) due to two-magnon and two-phonon excitations.[32] In the following, the amplitude of the line-widths of both spikes and Raman features are given inside parenthesis. In the first data set (shown in Figure 3a), we focused

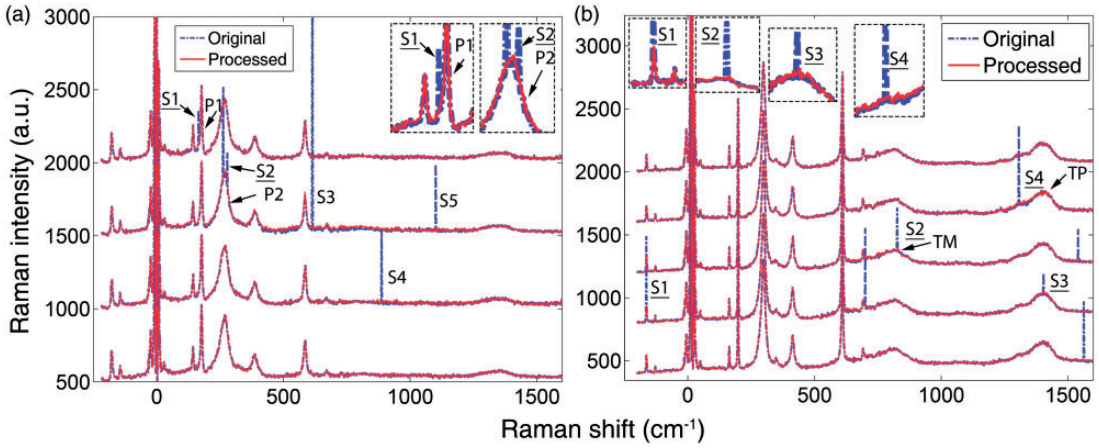

Figure 3. The original (blue dash-dot line) and processed (red line) Raman spectra of $Sr_3Ir_2O_7$. The spectra are offset for clarity. The zoom-in images show details of both spectra. The labels P, S, TM, and TP indicate phonon, spike, two-magnon, and two-phonon, respectively. The correspondences are indicated by the underlined spike indexes. (a) Spectra taken at 280 K. S3 spike is as high as 6000. (b) Spectra taken at 125 K.

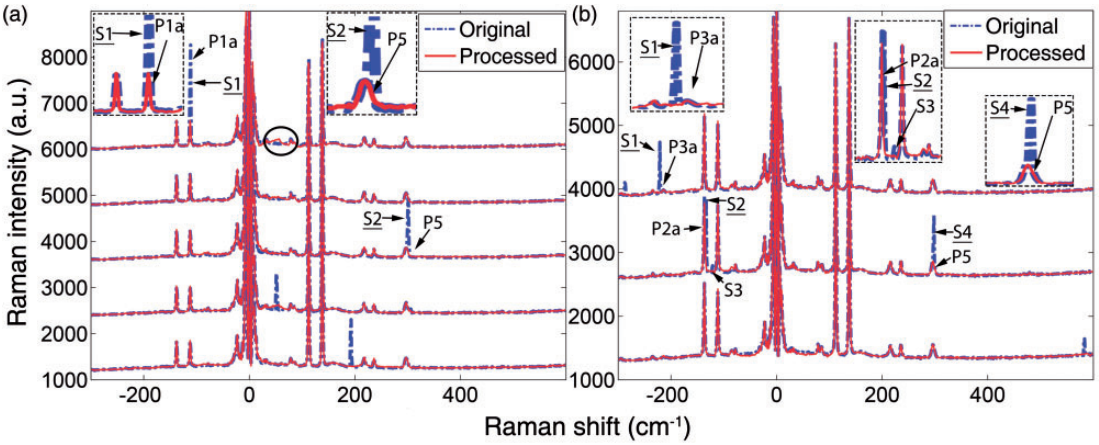

Figure 4. The original (blue dash-dot line) and processed (red line) Raman spectra of $Cr_2Ge_2Te_6$. The spectra are offset for clarity. The zoom-in images show details of both spectra. The correspondences are indicated by the underlined spike indexes. Phonons and spikes are indicated by P and S, respectively. PX_a denote PX (X = 1,2,3) phonon at anti-Stokes side. (a) Spectra taken at 100 K. (b) Spectra taken at 185 K.

on the spikes overlapping with intermediately broad features (FWHM 5-50 cm^{-1}). There are six spikes in total visible in the spectra. Here phonons, spikes, two-magnons, and two-phonons are denoted by P, S, TM, and TP, respectively. The original and processed spectra (obtained with $r_j = 20 \times j$) are shown in a blue dash–dot line and a red solid line, respectively. Judging by the zoom-in image in Figure 3a, we can see the algorithm indeed removes spikes S1 (overlapped with P1 (6 cm^{-1}, 400 counts)) and S2 (overlapped with P2 (40 cm^{-1}, 300 counts)) very well, while there is still some residue left in the processed spectra from S3 to S5. In the next subsection, a strategy of the selection of r_j will be discussed to minimize these residues. Compared to the first set, spikes S2–S4 (1–3 cm^{-1}, 200–600 counts) in the second data set (shown in Figure 3b) interfere with much broader features TM (~100 cm^{-1}, 150 counts) and TP (~80 cm^{-1}, 250 counts). Broad features like this also are often observed in organic samples[33,34] and other two-particle excitations[35,36] and are more likely to be contaminated by spikes due to the broadness. Nonetheless, it is clear that our method was able to eliminate spikes S2–S4 from the spectra with negligible distortion of the original spectra.

Another two data sets on single crystal $Cr_2Ge_2Te_6$ were taken at 100 K and 185 K with laser power 0.08 mW.[37] Each spectrum was collected with 15 min exposure and 900 data points were acquired. The spectra processed with $r_j = 15\sqrt{j}$ are shown in Figure 4. As can been seen

from Figure 4, $Cr_2Ge_2Te_6$ contains much sharper phonon lines than $Sr_3Ir_2O_7$ which is obviously more challenging to cope with. For example, spike S1 (2 cm^{-1}, 1500 counts) sits at the top of a narrow feature $P1_a$ (2.8 cm^{-1}, 700 counts) in Figure 4a, while Spike S2 (1.4 cm^{-1}, 800 counts) is very near to the shoulder of $P2_a$ (2.3 cm^{-1}, 1100 counts) in Figure 4b. From the insets of Figure 4a and b, we can see the phonon lines were well recovered. Especially noteworthy is the removal of the tiny S3 spike (Figure 4b). Moreover, in the inset of Figure 4b, we can see the phonon $P3_a$ (5 cm^{-1}, 50 counts) which is almost overwhelmed by spike S1 were well recovered. The spikes (S2 in Figure 4a and S4 in Figure 4b) overlapped with phonons with broader linewidth were also removed. However, as can be seen from the circled region in Figure 4a, there are some distortions to the original spectra. In the next subsection, these artifacts can be resolved through a careful selection of r_j. Nonetheless our method is able to safely remove the spikes while preserving most of the features of the original data.

Selection of $\boldsymbol{r_j}$

Without any doubt, r_j is the most important parameter in the entire algorithm. Since r_j represents the allowance of variance among spectra, a good r_j should be able to keep small differences resulting from environmental changes, while still filtering out spikes. In the previous subsection, we already demonstrated results with different choice of r_j. In this subsection, we study its influence more thoroughly. One representative data set, $Cr_2Ge_2Te_6$ taken at 185 K was used for the investigation. r_j was firstly set as a linear function of $r_s \times j$ with r_s taking four different values 5, 10, 15, and 20. The resulting spectra are shown in Figure 5. For the purpose of quantitative comparison, the spikes were also removed by hand, namely adjusting the data to an approximate value. The cross-correlations[38] between the processed spectra and the corrected spectra were computed and shown in the plot. In theory, if the spectra are of high similarity except for the spikes, one would set r_s as small as possible to remove all local frequency components of the spikes. As shown in Figure 5b, although the spike was removed in all four cases, when we take the difference between the processed spectra and manually corrected spectra, we can see the spectra with minimum r_s indeed has the best performance. Specifically, the low local frequency components (310–340 cm^{-1}) of the spike are not completely diminished in other cases. On the other hand, if the spectral variance is noticeable, an intermediate r_s is needed to avoid unexpected errors. This can be seen by comparing the processed spectra in Figure 5d. For the bottom spectra, because the r_s was set too small, no cluster with radius r_1 smaller than 5 could be found. Consequently, no change was made for d_1 at the location of the spikes, resulting in the spike remaining in the processed spectra. For the other three cases, the algorithm was able to aggregate d_1 with larger r_1 and reset the erroneous coefficients, leading to the spike-removal. Besides, in Figure 5e when one compares the resulting difference between the two green dashed lines where the phonon features are located, as r_s increases the "W" shape in the resulting differences becomes more and more flat. Thus, if r_s is small the resulting difference is larger, and the automatically filtered spectra reveal more rapid fluctuations, indicating the modification of high local frequency components of phonon features. This results from a lack of variance allowance at low levels for high local frequency components. As such, the peak feature is better preserved for larger r_s. On the other hand, just as the case discussed previously, the low local frequency components of the spike are better removed with small r_s which is shown over the spectral range –130 cm^{-1} to –70 cm^{-1} in Figure 5e. So a trade-off between the allowance of variance to keep the real feature and elimination of low local frequency components of spikes has to be made, leading one to choose intermediate r_s. This is confirmed by the highest cross-correlation (shown in Figure 5a) obtained with $15 \times j$.

To further improve the performance and remove the low local frequency components of the spikes (the featureless and flat offset region between –140 and –80 cm^{-1} and between 270 and 340 cm^{-1}, see Figure 5c and e) of the spikes, one can also decrease the r_j for high level clustering. To achieve this, we employ a second strategy where r_j is proportional to $\sqrt{j}$. This strategy will definitely reduce the r_j for high level clustering, however it will also significantly change r_j for the intermediate levels. To compensate that, instead of setting $r_j = 15 \times \sqrt{j}$, we set $r_j = 20 \times \sqrt{j}$. In Figure 6, we show the results with r_j set to $20 \times \sqrt{j}$ as well as the best result obtained with the previous $r_j = 15 \times j$. For convenience, we have abbreviated the r_j settings to S20 (S15) for $20 \times \sqrt{j}$ ($15 \times j$). Judging from the two curves in Figure 6a and c, both settings were able to retain the real features and result in overall similar residues in between the two dashed lines (shown in Figure 6b and d). However, when comparing the region outside of the lines, the residue for S15 is more significant than S20 in terms of both amplitude and oscillation. By comparison, the residue for S20 is more or less flat and featureless. So it can be concluded S20 indeed performed better in terms of eliminating the low local frequency components of the spikes, which is also confirmed by the calculated cross-correlation (shown in Figure 6a).

As we mentioned above, sometimes artifacts are generated in the processed spectra (circled region in Figure 4a). These broad artifacts are usually due to the less sufficient tolerance of r_j at high level. Since it is just the opposite to what we met previously where we shrunk the amplitude of r_j at high level, in this case, we need to adjust r_j in the opposite direction and reset r_j from $15 \times \sqrt{j}$ to $10 \times j$. The results are shown in Figure 7. We can see clearly the artifacts are removed while leaving other real Raman

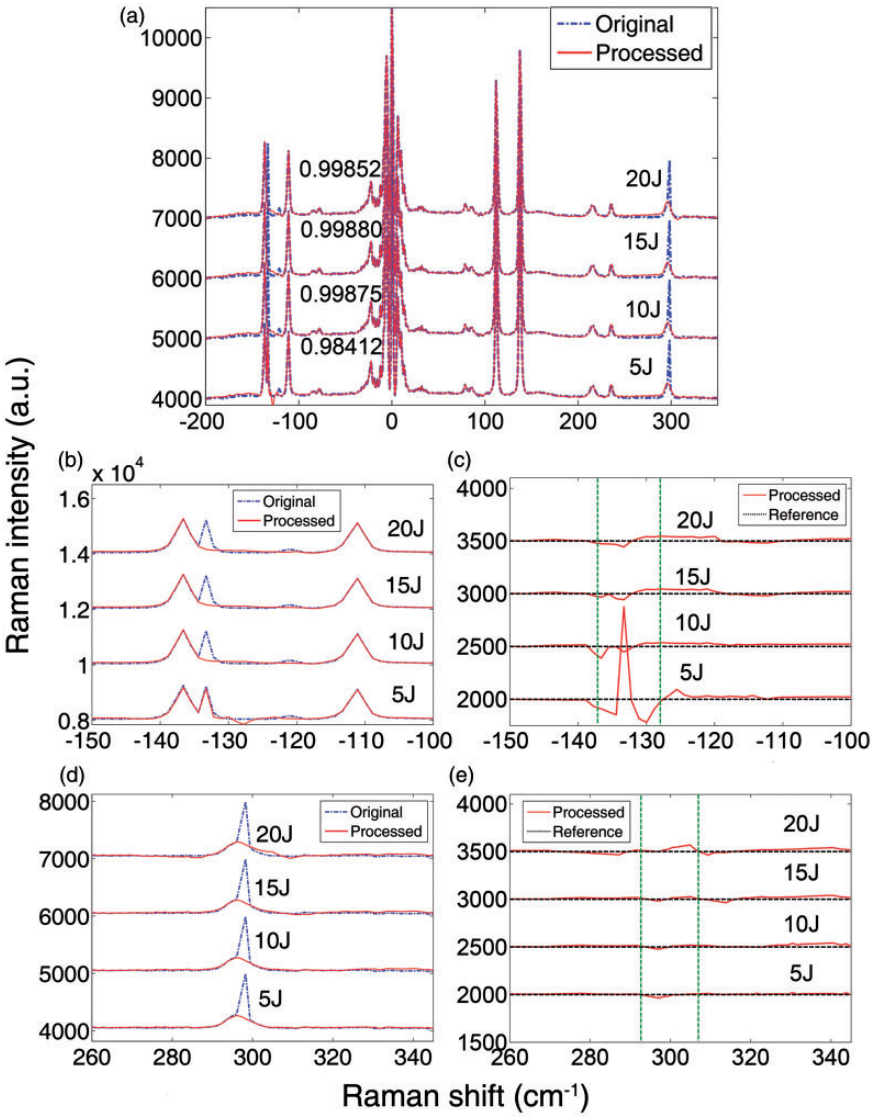

Figure 5. Influence of r_s (shown in the labels) on the resulting spectra. The original (processed) spectra are shown in blue (red). Spectra are offset intentionally for clarity. (a) Wider range spectra processed by four r_s. The cross-correlation coefficients are shown in number. (b, d) The processed spectra are shown in zoom-in images. (c, e) The differences between the processed spectra and manually fixed spectra. The resulting differences are shown in red. The black dashed lines are the reference lines. The green dashed lines indicate the region where phonon features and spikes are located.

features almost intact. This was further confirmed by the larger cross-correlation shown in the figure.

To correctly choose a r_j, one can judge the relative local frequency characteristic of Raman features, if the Raman features are very broad as well as have large variation, a linear setting of r_j should be used. On the other hand, if the feature is sharp as the case in the representative spectra above, a square root setting of r_j should achieve better performance. However, if broad artifacts are generated in the resulting spectra, one could consider switching back to the "linear" strategy and adjusting the coefficients correspondingly. We only show two different strategies to set r_j: linear and square root. In practice, other settings can be explored to fit in different applications.

Selection of Wavelets

Eight orthogonal wavelets were considered (Figure 8). The corresponding wavelet functions Φ and scaling functions Ψ can be found in the supplemental materials. More details about these wavelets can be found in Matlab help documents and the "Wavelet Browser" website developed by Filip Wasilewski.[39] The same representative spectra on $Cr_2Ge_2Te_6$ taken at 185 K as well as the processed spectra

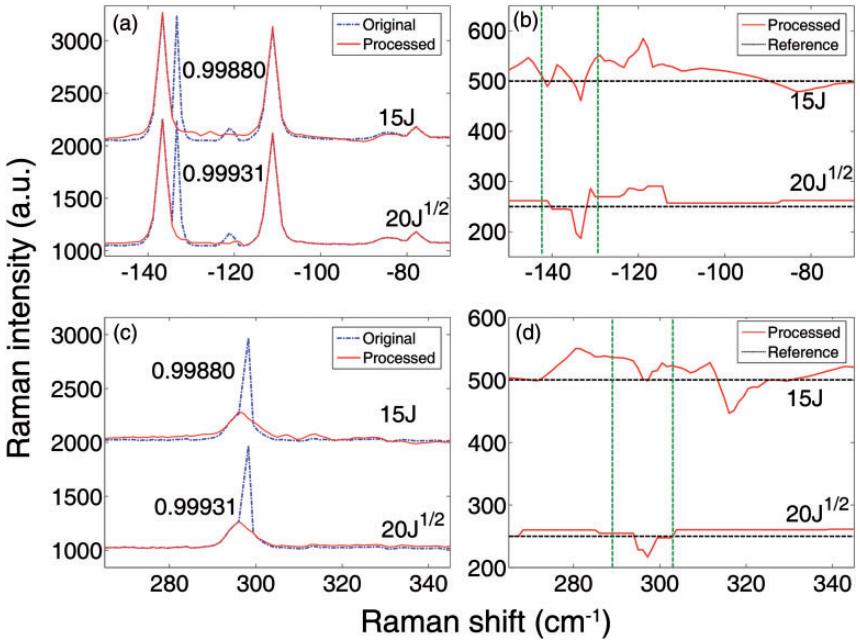

Figure 6. Spectra processed by different threshold setting strategies $20 \times \sqrt{j}$ and $15 \times j$. (a, c) Zoom-in images highlight the position of spikes. The numbers in plot show the cross-correlation coefficients for the entire spectral range. (b, d) Differences between processed spectra and manually fixed spectra. The resulting differences are shown in red. The black dashed line are the reference lines. The green dashed lines indicate the region where phonon features and spikes are located.

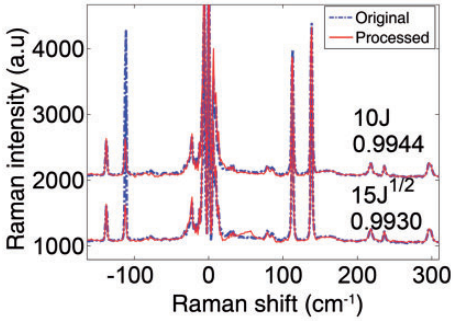

Figure 7. Demonstration of the removal of the artifacts.

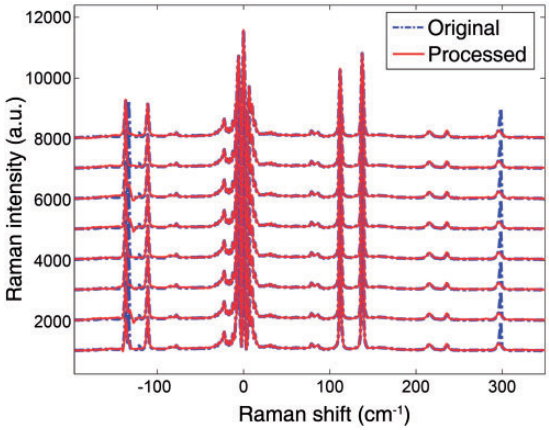

Figure 8. Wavelets dependence of the performance of the algorithm. From the bottom to top, the wavelets are "haar," "db3," "db6," "sym2," "sym3," "sym4," "coif1," and "coif5," respectively.

are shown in Figure 8. The r_j was set to $15 \times j$. As we can see, generally spikes are removed from all spectra regardless the wavelet employed. The reduced dependence on specific choice of wavelet makes our algorithm even more user-friendly. Nonetheless, there is still slightly different behavior in terms of low local frequency residue. This can be explained by the different profiles of the wavelets. When the spikes are projected into the wavelet space using different wavelets, some wavelets can represent spikes and Raman features better than others due to the different vanishing moment of wavelets.

To compare the performances and check the integrity, we computed the cross-correlation between the processed spectra and manually processed spectra for all four data sets. The calculated results and the one for original spectra are listed in Tables 1 and 2. In general, all wavelets are capable of spike-removal and have very close performances less than 0.1–0.2%. Nonetheless, "sym2" has the highest cross-correlation for the first three sets and "coif1" works best for the last data set. Thus, it can be inferred that "sym2" is more suitable for the spike-removal algorithm.

Table 1. Cross-correlation between processed and manually fixed spectra.

Wavelets	CGT_100k	CGT_185k	SIO_280k	SIO_125k
haar	0.996607	0.998906	0.0.996439	0.999322
db3	0.996882	0.998647	0.994785	0.999153
db6	0.994724	0.998510	0.995738	0.998831
sym2	**0.997576**	**0.998980**	**0.997146**	0.999159
sym3	0.996882	0.998647	0.994785	0.999153
sym4	0.995907	0.998508	0.994316	0.999185
coif1	0.996714	0.998718	0.995952	**0.999420**
coif5	0.995491	0.998288	0.994563	0.998753

CGT (SIO) is abbreviated for $Cr_2Ge_2Te_6$ ($Sr_3Ir_2O_7$). The MAX cross-correlation coefficients in the columns are in bold.

Table 2. Cross-correlation coefficients between original and manually fixed spectra. The spike as high as 6000 counts leads to the small correlation coefficient in SIO_280k.

	CGT_100k	CGT_185k	SIO_280k	SIO_125k
Original	0.959753	0.985569	0.873641	0.989123

Conclusion and Future Improvement

Cosmic ray induced spike-removal in Raman spectra is not a simple task due to the complex shape and large amplitude of spikes. A novel algorithm based on wavelet transform and data clustering has been proposed and validated using a wide range of experimental data. The spike detection and removal is performed by a multiresolution data clustering of wavelet coefficients. The processed coefficients can then be used for reconstruction through the inverse wavelet transform. The procedure has advantageously utilized the localization property of wavelets, which not only enables good separation of real features and spikes in wavelet space but also results in little dependence of the specific choice of wavelets. The algorithm is simple, easy to implement, uses widely available functions, has a little dependency on the specific wavelets employed, as well as intuitive threshold setting, allowing for usage by non-experts in spectra filtering. Nonetheless one small drawback still remains: one still has to set a reasonable r_j which limits its potential to be integrated for full automation. Future improvement will be focused on a more intelligent clustering strategy to remove this bottleneck.

Acknowledgments

The authors thank Huiwen Ji, Robert J. Cava at Princeton University for providing the $Cr_2Ge_2Te_6$ crystal. They also thank Stephen Wilson at University of California, Santa Barbara for providing the $Sr_3Ir_2O_7$ crystal.

Conflict of Interest

The authors report there are no conflicts of interest.

Funding

Work at the University of Toronto was supported by NSERC, CFI, and ORF. KSB acknowledges support from the National Science Foundation (grant DMR-1410846).

References

1. R.L. McCreery. Raman Spectroscopy for Chemical Analysis, volume 225. Chichester: John Wiley & Sons, 2005.
2. T.P. Devereaux, R. Hackl. "Inelastic Light Scattering from Correlated Electrons". Rev. Mod. Phys. 2007. 79(1): 175.
3. A. Bera, K. Pal, D.V.S. Muthu, S. Sen, P. Guptasarma, U.V. Waghmare, A.K. Sood. "Sharp Raman Anomalies and Broken Adiabaticity at a Pressure Induced Transition from Band to Topological Insulator in Sb_2Se_3". Phys. Rev. Lett. 2013. 110: 107401.
4. R. Loudon. "Raman Effect in Crystals". Adv. Phys. 1964. 13(52): 423–482.
5. A.C. Ferrari, J.C. Meyer, V. Scardaci, C. Casiraghi, M. Lazzeri, F. Mauri, S. Piscanec, D. Jiang, K.S. Novoselov, S. Roth, A.K. Geim. "Raman Spectrum of Graphene and Graphene Layers". Phys. Rev. Lett. 2006. 97: 187401.
6. W.H. Weber, M. Roberto, editors. Raman Scattering in Materials Science. New York: Springer, 2000.
7. L.J. Sandilands, J.X. Shen, G.M. Chugunov, S.Y.F. Zhao, S. Ono, Y. Ando, K.S. Burch. "Stability of Exfoliated $Bi_2Sr_2Dy_xCa_{1-x}Cu_2O_{8+\delta}$ Studied by Raman Microscopy". Phys. Rev. B. 2010. 82(6): 064503.
8. S.Y.F. Zhao, C. Beekman, L.J. Sandilands, J.E.J. Bashucky, D. Kwok, N. Lee, A.D. Laforge, S.W. Cheong, K.S. Burch. "Fabrication and Characterization of Topological Insulator Bi_2Se_3 Nanocrystals". Appl. Phys. Lett. 2011. 98(1)1911.
9. C. Beekman, A. Reijnders, Y. Oh, S.W. Cheong, K.S. Burch. "Raman Study of the Phonon Symmetries in $BiFeO_3$ Single Crystals". Phys. Rev. B. 2012. 86(2): 020403.
10. H.W. Lo, A. Compaan. "Raman Measurement of Lattice Temperature during Pulsed Laser Heating of Silicon". Phys. Rev. Lett. 1980. 44(2): 1604–1607.
11. Y. Tian, G.B. Osterhoudt, S. Jia, R.J. Cava, K.S. Burch. "Local Phonon Mode in Thermoelectric Bi_2Te_2Se from Charge Neutral Antisites". Appl. Phys. Lett. 2016. 108(14): 041911.
12. C.S. Kumar. Raman Spectroscopy for Nanomaterials Characterization. New York: Springer Science & Business Media, 2012.
13. A.T. Tu, A. Tu. Raman Spectroscopy in Biology: Principles and Applications. New York: Wiley, 1982.
14. D. Groom. "Cosmic Rays and Other Nonsense in Astronomical CCD Imagers". Experimental Astronomy. 2002. 14(1): 45–55.
15. H. Choi. Cosmic-Ray Interactions In Charged-Couple Devices In the DMTPC 4-Shooter Detector. [Ph.D. Thesis]. Boston, MA: Massachusetts Institute of Technology, 2013.
16. F. Ehrentreich, L. Sümmchen. "Spike Removal and Denoising of Raman Spectra by Wavelet Transform Methods". Anal. Chem. 2001. 73(17): 4364–4373.
17. I.R. Lewis, H. Edwards. Handbook of Raman Spectroscopy: From the Research Laboratory to the Process Line. Boca Raton, FL: CRC Press, 2001.
18. J. Zhao. "Image Curvature Correction and Cosmic Removal for High-Throughput Dispersive Raman Spectroscopy". Appl. Spectrosc. 2003. 57(11): 1368–1375.
19. Y. Katsumoto, Y. Ozaki. "Practical Algorithm for Reducing Convex Spike Noises on a Spectrum". Appl. Spectrosc. 2003. 57(3): 317–322.

20. A. Maury, R.I. Revilla. "Autocorrelation Analysis Combined with a Wavelet Transform Method to Detect and Remove Cosmic Rays in a Single Raman Spectrum". Appl. Spectrosc. 2015. 69(8): 984–992.
21. W. Hill, D. Rogalla. "Spike-Correction of Weak Signals from Charge-Coupled Devices and Its Application to Raman Spectroscopy". Anal. Chem. 1992. 64(21): 2575–2579.
22. D. Zhang, K.N. Jallad, D. Ben-Amotz. "Stripping of Cosmic Spike Spectral Artifacts Using a New Upper-Bound Spectrum Algorithm". Appl. Spectrosc. 2001. 55(11): 1523–1531.
23. H.G. Schulze, R.F. Turner. "A Two-Dimensionally Coincident Second Difference Cosmic Ray Spike Removal Method for the Fully Automated Processing of Raman Spectra". Appl. Spectrosc. 2014. 68(2): 185–191.
24. S. Mozharov, A. Nordon, D. Littlejohn, B. Marguardt. "Automated Cosmic Spike Filter Optimized for Process Raman Spectroscopy". Appl. Spectrosc. 2012. 66(11): 1326–1333.
25. Y. Meyer, D.H. Salinger. Wavelets and Operators, Volume 1. Cambridge: Cambridge University Press, 1995.
26. I. Daubechies. Ten Lectures on Wavelets, Volume 61. Philadelphia, PA: SIAM, 1992.
27. S.G. Mallat. "A theory for multiresolution signal decomposition: the wavelet representation". IEEE Transactions on Pattern Analysis and Machine Intelligence. 1989. 11(7): 674–693.
28. C.M. Bishop. Pattern Recognition and Machine Learning, Volume 4. New York: Springer, 2006.
29. A.J. Smola, B. Schölkopf. "A Tutorial on Support Vector Regression". Statistics and Computing. 2004. 14(3): 199–222.
30. Y. Tian, A.A. Reijnders, G.B. Osterhoudt, I. Valmianski, J.G. Ramirez, C. Urban, R. Zhong, J. Schneeloch, G. Gu, I. Henslee, K.S. Burch. "Low Vibration High Numerical Aperture Automated Variable Temperature Raman Microscope". Rev. Sci. Instrum. 2016. 87(4): 043105.
31. M. Moretti Sala, M. Rossi, A. Al-Zein, S. Boseggia, E.C. Hunter, R.S. Perry, D. Prabhakaran, A.T. Boothroyd, N.B. Brookes, D.F. McMorrow, G. Monaco, M. Krisch. "Crystal Field Splitting in $Sr_{n+1}Ir_nO_{3n+1}$ (n = 1, 2) Iridates Probed by X-ray Raman Spectroscopy". Phys. Rev. B. 2014. 90: 085126.
32. M.F. Çetin. Light Scattering in Spin Orbit Coupling Dominated Systems. [Ph.D. Thesis]. Braunschweig, Lower Saxony, Germany, Braunschweig University of Technology, 2012.
33. A. Rygula, K. Majzner, K.M. Marzec, A. Kaczor, M. Pilarczyk, M. Baranska. "Raman Spectroscopy of Proteins: A Review". J. Raman Spectrosc. 2013. 44(8): 1061–1076.
34. S. Rath, M. Hsieh, P. Etchegoin, R.A. Stradling. "Alloy Effects on the Raman Spectra of $Si_{1-x}Ge_X$ and Calibration Protocols for Alloy Compositions Based on Polarization Measurements". Semicond. Sci. Technol. 2003. 18(6): 566–575.
35. P. Parayanthal, F.H. Pollak. "Raman Scattering in Alloy Semiconductors: "Spatial Correlation" Model". Phys. Rev. Lett. 1984. 52(20): 1822.
36. L.J. Sandilands, Y. Tian, K.W. Plumb, Y.-J. Kim, K.S. Burch. "Scattering Continuum and Possible Fractionalized Excitations in α-$RuCl_3$. Phys. Rev. Lett. 2015. 114(14): 147201.
37. Y. Tian, M.J. Gray, H. Ji, R. J. Cava and K.S. Burch. 2D Materials 3, 025035. 2016.
38. A.V. Oppenheim, R.W. Schafer, J.R. Buck. Discrete-Time Signal Processing, Volume 2. Englewood Cliffs, NJ: Prentice-Hall, 1989.
39. F. Wasilewski. Independent IT consultant; website name: Wavelet browser by pywavelets. Website Address: http://wavelets.pybytes.com/. Wavelet plots [accessed 6 Dec 2015].