Review of Scientific Instruments

A wavelet transform algorithm for peak detection and application to powder x-ray diffraction data

Cite as: Rev. Sci. Instrum. **82**, 015105 (2011); https://doi.org/10.1063/1.3505103
Submitted: 07 June 2010 . Accepted: 01 October 2010 . Published Online: 13 January 2011

John M. Gregoire, Darren Dale, and R. Bruce van Dover

AIP Publishing

A wavelet transform algorithm for peak detection and application to powder x-ray diffraction data

John M. Gregoire,[1,a)] Darren Dale,[2] and R. Bruce van Dover[1]

[1]*Department of Materials Science and Engineering and Energy Materials Center at Cornell, Cornell University, Ithaca, New York 14853, USA*

[2]*Cornell High Energy Synchrotron Source (CHESS), Cornell University, Ithaca, New York 14853, USA*

(Received 7 June 2010; accepted 1 October 2010; published online 13 January 2011)

Peak detection is ubiquitous in the analysis of spectral data. While many noise-filtering algorithms and peak identification algorithms have been developed, recent work [P. Du, W. Kibbe, and S. Lin, Bioinformatics **22**, 2059 (2006); A. Wee, D. Grayden, Y. Zhu, K. Petkovic-Duran, and D. Smith, Electrophoresis **29**, 4215 (2008)] has demonstrated that both of these tasks are efficiently performed through analysis of the wavelet transform of the data. In this paper, we present a wavelet-based peak detection algorithm with user-defined parameters that can be readily applied to the application of any spectral data. Particular attention is given to the algorithm's resolution of overlapping peaks. The algorithm is implemented for the analysis of powder diffraction data, and successful detection of Bragg peaks is demonstrated for both low signal-to-noise data from theta–theta diffraction of nanoparticles and combinatorial x-ray diffraction data from a composition spread thin film. These datasets have different types of background signals which are effectively removed in the wavelet-based method, and the results demonstrate that the algorithm provides a robust method for automated peak detection. © *2011 American Institute of Physics*. [doi:10.1063/1.3505103]

I. INTRODUCTION

The wavelet transform is a broadly applicable analysis tool that is analogous to the more familiar Fourier transform. Fourier analysis is commonly used to express spectral data in terms of frequency components and associated phases. While transformation into this frequency-phase space is useful for many data analysis practices, the most intuitive transformation space for peak identification is a peak width-position space. In the formalism of wavelet analysis, this type of transform can be built by appropriate choice of a mother wavelet $w(x)$ such that peak width and position are accessed through dilation and translation of this wavelet,

$$w_{a,b}(x) = a^{-1/2}\, w\left(\frac{x-b}{a}\right). \tag{1}$$

As indicated in Eq. (1), dilation and translation are described by the wavelet scale parameter a and wavelet position parameter b, respectively. Given a 1D function (e.g., spectral data) $f(x)$ and assuming both w and f are real-valued functions, the wavelet transformation of f is given by

$$T(a,b) = \int w_{a,b}(x) f(x)\,\mathrm{d}x. \tag{2}$$

A common choice of mother wavelet is the Lorentzian of Gaussian (LoG) wavelet,

$$w_{1,0}(x) \propto (1-x^2)\exp[-x^2/2], \tag{3}$$

which is illustrated in Fig. 1. The LoG wavelet transform of example data with Gaussian and Lorentzian peaks is depicted in Figs. 2 and 6.

Wavelet analysis has been employed in data processing algorithms for several branches of the physical sciences, including x-ray crystallography. Wavelet-based data compression and filtering algorithms are analogous to their counterparts built upon Fourier transformations. In x-ray diffraction analysis, wavelet-based data compression has been applied to powder patterns to both improve computation efficiency[1] and identify crystalline phases via comparison with wavelet transforms of pure phase patterns.[2] Wavelet denoising algorithms have been implemented in data processing algorithms for several varieties of spectral data, including powder x-ray diffraction.[3] As in Fourier-based filtering, the denoised data are typically transformed back into their original (1D) coordinate space for further analysis.

Important properties of the data f can also be acquired through direct analysis of its wavelet transform. For example, the variation in $T(a,b)$ as a function of the wavelet scale a provides "multiresolution" analysis, which has been incorporated into analysis of x-ray differential correlation functions by Ding *et al.*[4] In this work, LoG wavelets with different widths probe ordering on different length scales, and analysis of the wavelet transformed data provides insights into the structure of silica glass *vis-a-vis* crystalline counterparts.

Algorithms for the identification of peaks in 1D data through analysis of the 2D wavelet transform surface have recently been developed. These algorithms demonstrate high sensitivity and low false detection rate for peak identification in low signal-to-noise spectra.[5–7] Du *et al.*[5] developed such an algorithm for analysis of mass spectrometry data using the LoG wavelet transform. A similar algorithm has been implemented using the derivative-of-Gaussian (or Ridger) wavelet to analyze low signal-to-noise data from contactless electrophoresis measurements.[6] These algorithms exhibit some

[a)] Current address: School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.

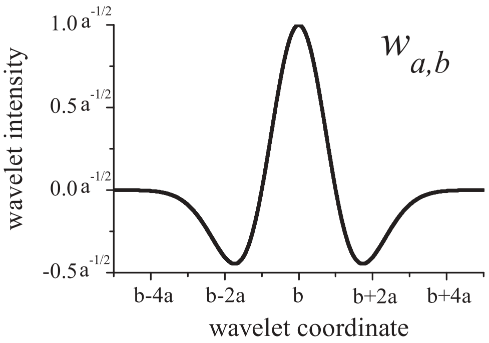

FIG. 1. The Lorentzian of Gaussian wavelet plotted in terms of its scale parameter a and position parameter b.

limitations in the analysis of powder x-ray diffraction data, particularly in the identification of overlapping peaks.

In the present work, we develop an algorithm based on that of Ref. 5. The wavelet-based peak detection algorithm is broadly applicable, and we describe an implementation of the algorithm that is well suited for peak identification in noisy x-ray powder patterns. In particular, we demonstrate the utility of the algorithm in analysis of datasets with different types of noise. In a dataset obtained with a conventional powder diffractometer, we identify low signal-to-noise Bragg peaks in powder patterns of Pt–Zn nanoparticles. In addition, we demonstrate successful analysis of diffraction patterns of a Pt–Ru composition spread thin film that were acquired using a recently developed high energy diffraction experiment. In these powder patterns, the background signal is not amenable to filtering by conventional algorithms, but the Bragg peaks are efficiently analyzed via the algorithms described in this paper.

In both cases, we demonstrate that noise filtering and background subtraction are inherent in the wavelet-based peak detection algorithm. With the parameters of the algorithm appropriately defined for the given noise level and anticipated range of peak widths, the peak detection algorithm comprises an automated, noninteractive method for the robust extraction of peak properties from any spectral data.

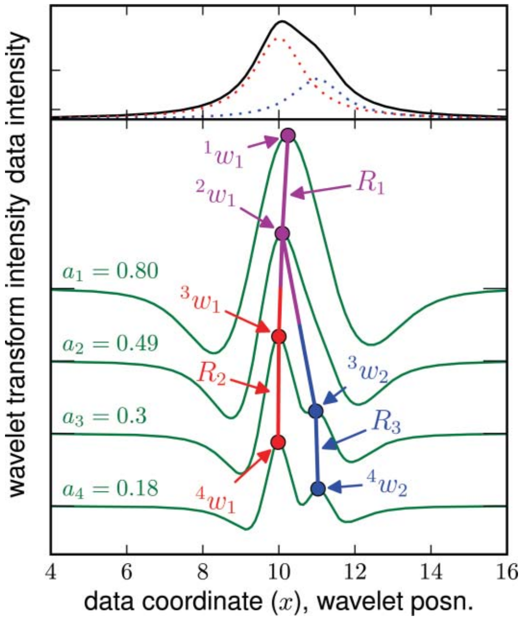

FIG. 2. (Color online) Peak identification of data (top, black) consisting of strongly overlapping Lorentzian peaks (top, dashed). The LoG wavelet transform at four scale parameters (bottom) is shown along with the local maxima and ridges, as defined in Sec. III B.

II. EXPERIMENTAL

The synthesis and diffraction measurements of Pt–Zn nanoparticles were carried out as described by Muira *et al.*[8] and employed a Scintag theta–theta diffractometer. The Pt–Ru metal thin film was sputter deposited in a combinatorial magnetron sputter deposition system described elsewhere.[9] Diffraction images of the thin film on single crystalline silicon substrate were acquired in transmission geometry using a 60 keV x-ray source. The details of the experiment and of the algorithm for processing the diffraction images into powder patterns are described in Ref. 10.

III. WAVELET PEAK IDENTIFICATION ALGORITHM

While a variety of mother wavelets could be used to identify peaks in diffraction spectra, the current work only involves the use of LoG family of functions $w_{a,b}$ given by Eqs. (3) and (1). The data f are assumed to be a 1D set of measurements performed on a grid of positions X with regular spacing δx in the measurement coordinate x. The wavelet transform is essentially a discrete convolution of data f with $w_{a,b}$, and we note that artifacts may arise in the wavelet transform calculation when the wavelet scale approaches the measurement interval ($a \lesssim 10\,\delta x$) or the wavelet position approaches the edge of the dataset (e.g., $b < \min\{X\} + 5a$). In the Appendix, we outline a simple algorithm for making robust wavelet transform calculations under these conditions.

The algorithm for extracting peak positions from the wavelet transform surface $T(a, b)$ is based on that of Ref. 5. The guiding principle of the algorithm is that given a peak in f with position x_{peak} and width σ_{peak}, the wavelet transform $T(a, b)$ will have two important properties. At fixed $a \approx \sigma_{\text{peak}}$, $T(a, b)$ will have a peak-shaped profile with local maximum near $b = x_{\text{peak}}$. At fixed $b \approx x_{\text{peak}}$, $T(a, b)$ will be a slowly varying function with appreciable intensity when $a \approx \sigma_{\text{peak}}$. Peaks are identified by finding the local maxima in $T(a, b)$ with respect to b, clustering these local maxima by their proximity with respect to b, and examining the cumulative magnitude of $T(a, b)$ at these local maxima with respect to a (see Sec. III B).

A. Design parameters of transformation

Given a mother wavelet, the only design parameters of a wavelet transform are the choices of a and b. For convenience, the set B of values for the position parameter is taken to be a subset of X. Transformations that use an equal spacing in the set A of scale parameters are loosely defined as continuous wavelet transforms and are used in the peak searching algorithms of Refs. 5 and 6.

We find that equal spacing in A results in oversampling of the large scale features of the data. The variation in $T(a, b)$ with respect to both parameters generally increases with decreasing a, and thus logarithmic spacing in the set A is more suitable for both data interpretation and computational efficiency. Lossless data compression algorithms based on discrete wavelet transforms commonly employ dyadic spacing in the scale parameter.[11] While this $\log_2$ spacing is adequate for wavelet-based peak detection in some datasets, we find that a finer spacing enables more straightforward identification of peaks. For peak identification in diffraction patterns, we employed $\log_{1.18}$ spacing in A (4.2 cycles per octave).

While this spacing in the set A is an important wavelet transform design parameter, one must also choose the appropriate range of A. The minimum value of A must be at least as big as the measurement spacing δx, has important implications for peaks splitting, and is an upper limit on the resolution of the position of identified peaks (see the Appendix). The maximum value of A is practically chosen to be larger than the value of a that maximizes the wavelet transform of the widest data peaks of interest. The maximum LoG wavelet transform for a peak with a half-width at half maximum equal to 1 is achieved with scale parameter of 2.9 for a Lorentzian peak and 1.9 for a Gaussian peak. For powder diffraction analysis, a suitable maximum value of A can be determined by considering the maximum anticipated peak width due to both Debye–Scherrer and instrument broadening.

Given the set A that is appropriately chosen for the widths of the peaks in f, we introduce a weighting function $g(a)$ for the mother wavelet. The coefficient $a^{1/2}$ in Eq. (1) ensures energy conservation of the wavelet for any a, an important property for many wavelet-based algorithms such as data compression. For peak detection algorithms, the energy conservation constraint can be lifted, allowing the wavelet to be modified and thus tailored for the shape of the data peaks and the desired sensitivity to the detection of overlapping peaks (the Appendix, Sec. II). That is, the weighting function allows the wavelet transform calculation to be customized for a particular type of data. For the LoG wavelet employed in the present work, the modified wavelet is given by

$$w_{a,b}(x) = g(a)a^{-1/2}\left[1 - \left(\frac{x-b}{a}\right)^2\right] \times \exp\left[-\left(\frac{x-b}{a}\right)^2 \Big/ 2\right], \tag{4}$$

and for our analysis of powder diffraction data, we use $g(a) = 1/a$. We note that the weighting function could equivalently be applied to the wavelet transform, which may be useful when a user is exploring different choices for $g(a)$. Regardless, using Eqs. (4) and (2), a discrete wavelet transform $T(a, b)$ is calculated and used in the following peak identification algorithm.

B. Ridge and peak identification algorithm

The following algorithm provides identification of data peaks by identifying the local maxima in the LoG wavelet transform and grouping the local maxima into ridges. Using the formalism of this algorithm, Fig. 2 depicts the analysis of the wavelet transform for a pair of overlapping peaks.

a. Calculate $T(a, b)$ over the grid of parameters $a \in A = \{a_1, a_2, \ldots, a_n\}$, $b \in B = \{b_1, b_2, \ldots\}$ where A is strictly decreasing.

b. For each $a_i \in A$, find the set M_i of local maxima of $T(a, b)$ with respect to b. That is, find $M_i = \{{}^i m_1, {}^i m_2, \ldots\} \subset B$ such that each ${}^i m_j$ is a local maximum and $T(a_i, {}^i m_j) > \eta_i$, where η_i is a chosen noise threshold.

c. Group the set of elements in $M_1 \cup M_2 \cup \ldots M_n$ into ridges R_g. The initial ridges are defined by $R_j = \{{}^1 m_j\}$ for each ${}^1 m_j \in M_1$, the set of local maxima for the largest wavelet scale.

d. For each wavelet scale $a_i, i = 2, 3, \ldots$ (ordered in decreasing wavelet scale), and for each element ${}^i m_j \in M_i$, append ${}^i m_j$ to one of the existing ridges R_g if the ridge contains an element ${}^{i-1}m$ from the larger wavelet scale such that $|{}^{i-1}m - {}^i m_j| < \delta_i$ and $|{}^{i-1}m - {}^i m_k| \geq \delta_i$ for every ${}^i m_k \in M_i/\{{}^i m_j\}$, where δ_i is a chosen interval in the shift parameter b. That is, local maxima of $T(a_{i-1}, b)$ and $T(a_i, b)$ are included in the same ridge if their separation in the b coordinate is less than δ_i and there are no other local maxima in $T(a_i, b)$ that meet this requirement.

e. If there is no existing R_g that satisfies this condition (or equivalently, no such ${}^{i-1}m$), then a new ridge is initialized as $R_h = \{{}^i m_j\}$. If there are multiple elements of M_i that are sufficiently close (within δ_i) of an ${}^{i-1}m \in R_g$, then similarly initialize new ridges R_{h1}, $R_{h2}, \ldots$, one for each of these elements. We define R_g as the mother ridge of each of these new ridges, i.e., $mother(R_{h1}) = mother(R_{h2}) = \cdots \equiv R_g$. We also define the new ridges as the descendants of R_g, i.e., $descendants(R_g) \equiv \{R_{h1}, R_{h2}, \ldots\}$, and catalog the descendants over multiple generations such that if R_g is a descendant of a ridge R_f, $descendants(R_f)$ becomes $descendants(R_f) \cup descendants(R_g)$. We note that R_g is effectively terminated in this step, as this ridge will not contain elements from finer wavelet scales.

f. (optional) In some cases, it may be necessary to allow ridges to effectively "skip" a number $s \geq 1$ of wavelet scales. That is, if there is no ${}^{i-1}m \in M_{i-1}$ such that $|{}^{i-1}m - {}^i m_j| < \delta_i$ but there is a ${}^{i^*}m \in M_{i-2} \cup M_{i-3} \cup \ldots M_{i-s-1}$ such that $|{}^{i^*}m - {}^i m_j| < \delta_i$, then include ${}^i m_j$ in the ridge containing ${}^{i^*}m$.

g. For each R_g, a peak is identified at position $x = {}^{i_{\max}}m \in R_g$ (where $i_{\max}$ is the largest wavelet scale index represented in R_g) if R_g satisfies the conditions

$$\sum_{{}^i m \in R_g} T(a_i, {}^i m) > \zeta_1, \tag{5}$$

$$\sum_{{}^i m \in R_g \cup mother(R_g) \cup mother(mother(R_g)) \cup \ldots} T(a_i, {}^i m) > \zeta_2, \tag{6}$$

$$\sum_{{}^i m \in R_h} T(a_i, {}^i m) \leq \zeta_1 \; \forall R_h \in descendants(R_g), \tag{7}$$

where ζ_1 and ζ_2 are user-defined thresholds with $\zeta_2 \geq \zeta_1$. For a ridge without descendants or a mother, condition (7) is trivially met and conditions (5) and (6) differ only in the threshold value. The criterion is that the sum of the values of the wavelet transform at the local maxima in the ridge must be larger than a chosen threshold. For a ridge that has a mother ridge, condition (6) includes the local maxima of the mother ridge in the cumulative wavelet transform intensity. For a ridge that has descendants, condition (7) requires that all descendant ridges have failed condition (5), and thus, a ridge is identified as a peak only if every descendant ridge is not identified as a peak.

C. Algorithm discussion

The above algorithm defines ridges and grows them through incremental analysis of finer wavelet scales. If the cumulative wavelet transform of the ridge is sufficiently large, a peak is identified at the position given by the end of the ridge. The creation of descendant ridges from a mother ridge corresponds to the identification of a set of overlapped peaks, and condition (7) precludes double-counting of peaks.

The value of δ_i should be larger than a_{i-1} and should not be so large as to allow ${}^i m_j$ to be associated with multiple local maxima from M_{i-1}. If this arises, ${}^i m_j$ should be associated with the closest member of M_{i-1}. Using larger values of δ_i promotes the detection of overlapping peaks but can also lead to false peak detection. The optimal value will vary with peak shape, and we find that $\delta_i = 1.5a_{i-1}$ is suitable for analysis of powder diffraction patterns with both Gaussian and Lorentzian peaks.

Our chosen weighting function $g(a) = 1/a$ provides increased wavelet transform intensity at smaller wavelet scales, which we found to be desirable for detection of low-intensity and overlapped peaks. An appropriate value for the noise threshold η_i is determined by the noise level in the data and choice of weighting function $g(x)$. To maximize the sensitivity of the algorithm, η_i should be sufficiently small that at a given a_i, there are local maxima in M_i that correspond to false peaks. Then, ζ_1 and ζ_2 are chosen to be sufficiently large that the false peaks are appropriately excluded from the set of identified peaks. While we recommend this heuristic determination of these parameters, a reasonable estimate may be obtained from the maximum wavelet transform sum [see condition (6)] obtained in the analysis of zero-signal (pure background noise) data with η_i set to zero.

The behavior of the algorithm can be tailored in subtle ways with different choices of η_i, ζ_1, and ζ_2, especially with regard to the algorithm's sensitivity to the detection of overlapped peaks. With ζ_2 chosen for the respective noise level, setting $\zeta_1 = 0$ provides maximal sensitivity to overlapped peaks because every descendant ridge that does not have descendants of its own will be identified as a peak, provided the cumulative wavelet transform of the ridge and its ancestry is above the threshold ζ_2. Setting $\zeta_1 = \zeta_2$ provides minimal sensitivity to overlapped peaks as condition (6) becomes obsolete and the descendant ridge alone must contain sufficient cumulative wavelet transform intensity. As described in the Appendix, the local maxima of an overlapped peak may be quite small, and thus for identification of overlapped peak in powder diffraction patterns, we use $\zeta_1 = 0$.

The algorithms of Refs. 5 and 6 include neither the concept of mother and descendant ridges nor cumulative threshold criteria [e.g., conditions (5)–(7)]. In these algorithms, the condition for identification of a ridge as a peak is that the length of a ridge must be above a threshold value. The peak identification conditions of Sec. III B can be extended using this condition and the corresponding condition for the cumulative length of a ridge and its ancestor ridges, which would provide two more parameters for tuning the algorithm for a given application. However, the ridge length condition is very sensitive to the choice of A and provides a discrete thresholding parameter. The continuous threshold parameters ζ_1 and ζ_2 are inherently more versatile and are particularly useful when a small number of wavelet scales are used, which is desirable for computational efficiency.

The peak position is defined as the end of the ridge because the wavelet transform at the smallest value of a has the highest resolution for peak position. While the finer wavelet scale makes the wavelet transform more susceptible to noise, the uncertainty in determination of the peak position is $\lesssim\delta_i$ by virtue of its inclusion in the ridge (although additional systematic errors arise in the case of overlapping peaks). The width of the identified peak can also be estimated by the wavelet scale a_i at which the wavelet transform value of the ridge is maximal, but we note that the relationship between this wavelet scale and the peak width will vary with peak shape, as noted in Sec. III A. For descendant ridges, the estimation of peak width is less straightforward.

D. Further processing and profile fitting

While the above algorithm provides estimates for the position, width, and height of the identified peaks, further refinement of these parameters can be obtained by fitting the data peaks to a functional form (profile). So that the background signal does not need to be modeled in the fitting procedure, the data are processed by traditional algorithms before profile fitting. For the powder patterns from high energy diffraction experiments, we employ an aggressive, noninteractive background-subtraction algorithm which includes Savitsky–Golay filtering and modeling of the baseline with cubic splines. The spline points are chosen such that the area under the baseline curve is maximized with the following two constraints: The height of the baseline curve must be less than f at all X, and the curvature of the background must remain below that of typical Bragg peaks. While this background-subtraction algorithm is very effective at removing most of the intensity outside of the film Bragg peaks, it often creates artifacts in the background-subtracted spectrum which are poorly distinguished from the true Bragg reflections. It is by virtue of the wavelet-based algorithm that the true Bragg peaks are distinguished. Given a functional form for the Bragg peaks, a least squares regression algorithm refines the peak parameters from their initial values provided by the wavelet peak detection algorithm. Further details of this procedure and notes on computational efficiency are included in the Appendix.

IV. RESULTS AND DISCUSSION

A. Wavelet analysis of diffraction data obtained in the theta–theta configuration

The diffraction pattern and data processing of the PtZn nanoparticle diffraction pattern are shown in Fig. 3. Using the algorithm of Sec. III B, nine peaks are identified in the noisy diffraction pattern. For comparison, the known positions of Bragg peaks from the PtZn ordered intermetallic phase are shown, and we note that deviations from these positions are expected due to shifted lattice parameters of the nanoparticles, which are off-stoichiometric and susceptible to lattice constriction due to surface energy effects. The wavelet-based algorithm identified every Bragg peak in the phase except the least intense peak near 75°, whose height in the Joint Committee for Powder Diffraction Standards (JCPDS) pattern is only 0.028 that of the peak at 41°.[12] We note that Fig. 3 contains a ridge of length 3, which likely corresponds to this peak, but the peak was not identified from this ridge due to an insufficient cumulative wavelet transform (condition 6). The only false peak detection is due to the identification of the peak at 66° as two overlapping peaks. We note that if the smallest wavelet scale in Fig. 3 is excluded, the wavelet-based peak detection algorithm correctly identifies the eight Bragg peaks without this additional false peak. Normally, that smallest scale would be excluded on the basis of expected peak width, but we have included this false peak identification in Fig. 3 for illustrative purposes.

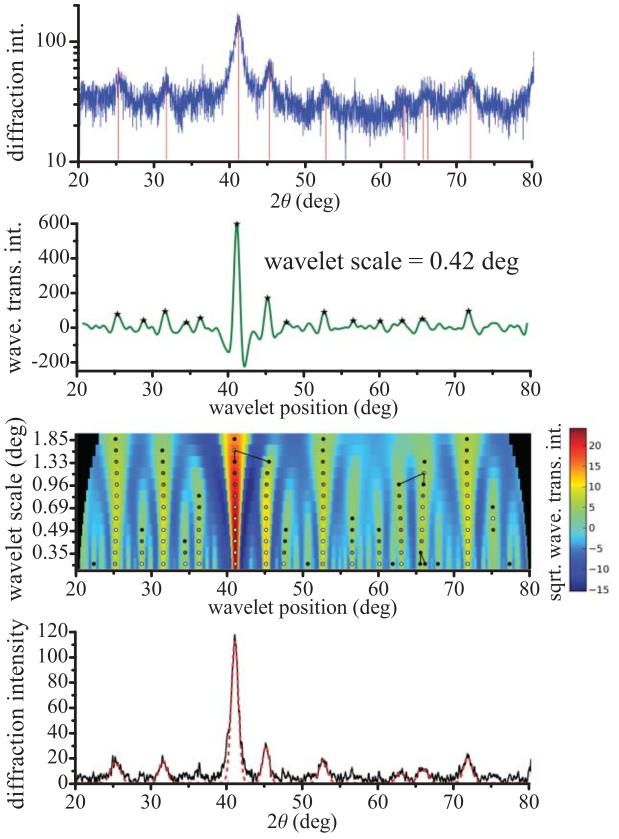

FIG. 3. (Color online) The analysis of the diffraction pattern of PtZn nanoparticles. Top: The diffraction pattern, the locations of the peaks identified by the wavelet-based algorithm (long lines), and the peak locations of PtZn from JCPDS card 03-066-0026 (short lines) (Ref. 12). Top-middle: The wavelet transform at wavelet scale 0.42 deg^{-1} and the local maxima above the noise threshold (stars). Bottom-middle: The wavelet transform at 12 wavelet scales is shown for the entire wavelet position range. The local maxima above the noise threshold are shown as overlaid circles which are colored in grayscale according to the relative wavelet transform intensity within the respective ridge. Mother-descendant relations between ridges are shown by black lines connecting the local maxima. Bottom: The background-subtracted diffraction pattern and fitted peak profile (dashed).

Figure 3 also shows the powder pattern after background subtraction using a Savitsky–Golay filter (Sec. III D). The peaks identified in the wavelet-based algorithm are more apparent in the filtered data, as discussed below in Sec. IV C. The fitted diffraction profile using a Lorentzian peak shape and initial parameters from the algorithm of Sec. III B shows good agreement with the filtered data.

B. Wavelet analysis of data from high energy diffraction experiments

For the Pt–Ru composition spread thin film, diffraction patterns from 20 film locations (compositions) were analyzed and processed using the algorithms of Sec. III B and Sec. III D. The noise thresholds η_i and ζ_2 were set quite high to avoid false peak detection. A summary of the results is presented in Fig. 4. The compositions corresponding to the 20 diffraction measurements are plotted as horizontal arrows on the left ordinate axis, and the diffraction intensity of the postprocessed diffraction patterns is plotted in a logarithmic color scale. The diffraction intensity is interpolated along the composition axis. The diffraction patterns used to create Fig. 4 were processed with the background-subtraction algorithms of Sec. III D, and the positions of peaks identified by the wavelet-based algorithm are plotted as white stars. Bragg peaks near 44 and 76 nm^{-1} are identified by the wavelet-based algorithm but are attributed to an underlayer film (not the Pt–Ru film). Other than these peaks, nearly every identified peak corresponds to a documented reflection from the fcc-Pt (JCPDS card 04-0802) and hcp-Ru (JCPDS card 06-0663) phases, considering the shifts lattice constant due to chemical alloying.[12] That is, the wavelet-based algorithm

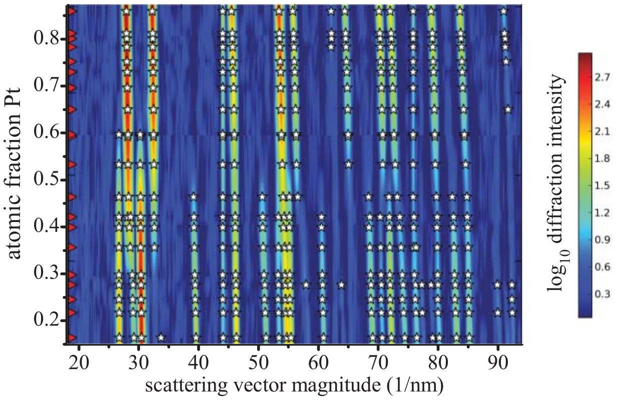

FIG. 4. (Color online) An interpolated diffraction map of the Pt–Ru composition spread thin film. The measurement compositions are indicated by arrows on the left ordinate axis and the identified peaks at these compositions are plotted as stars.

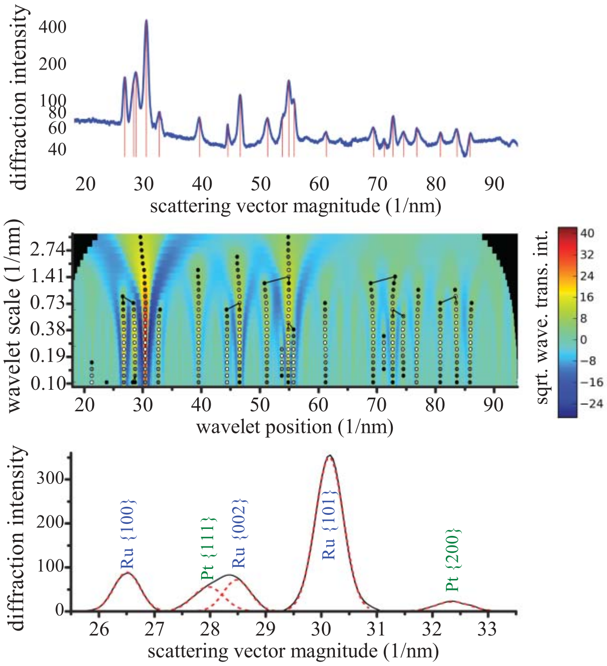

FIG. 5. (Color online) The analysis of the diffraction pattern from the Pt–Ru thin film at 42 at. % Pt. Top: The diffraction pattern and the locations of the peaks identified by the wavelet-based algorithm (vertical lines). Middle: The wavelet transform at 23 wavelet scales is shown with overlaid ridges (see Fig. 3 for other notation). Bottom: A subset of the background-subtracted diffraction pattern and five fitted peaks (dashed).

correctly identifies 338 Pt and Ru Bragg peaks over the entire composition range, including the two-phase region, with the only falsely identified peak occurring at 34 nm^{-1}, 16 at.% Pt.

An example of the data used in the peak searching algorithm is shown in Fig. 5. The wavelet transform and ridges are also shown, indicating the identification of 21 Bragg peaks, many of which are overlapping. Figure 5 also demonstrates the data processing and profile fitting of the five peaks identified in the range 26–33 nm^{-1}. The two most strongly overlapped peaks in this range (Pt {111} and Ru {002}) are resolved in the wavelet-based algorithm only at the smallest wavelet scale in the chosen set of wavelet transform parameters.

C. Background subtraction and false peak detection

The background signal in powder patterns is generally defined as any signal which is not due to Bragg scattering and is typically composed of a slowly varying baseline superimposed with higher frequency Gaussian noise. The baseline is often modeled using low order polynomials or cubic splines and subsequently subtracted from the data. Background subtraction using these "smooth" profiles may make features, such as edges, in the original pattern appear as Bragg peaks in the resulting data. This undesirable occurrence is an example of the general problem that background subtraction may introduce artifacts to the data and lead to false peak identification.

The wavelet transformation naturally includes both of these types of background subtraction, but the algorithm of Sec. III B is less susceptible to false peak detection because the identification of peaks occurs through analysis of $T(a, b)$, not background-subtracted data. As indicated by Eq. (2), the wavelet transform is a window average over the effective support region of the wavelet (see the Appendix). Also, the choice of a wavelet (such as the LoG) with no zeroth or first moment results in a baseline subtraction in $T(a, b)$ which is smooth on the length scale defined by a. Thus, the wavelet transform surface is a background-filtered dataset in which peaks are identified without introduction of artifacts.

The diffraction pattern in Fig. 3 contains a background signal that is typical in theta–theta diffractometry, and this background signal is efficiently removed by both the wavelet transform and the Savitsky–Golay filter. The diffraction patterns obtained in the high energy x-ray diffraction experiment contain a background signal that cannot be removed by a Savitsky–Golay filter because the acquired diffraction images include not only the Bragg diffraction from the thin film but also thermal diffuse scattering from the single crystal Si substrate. While the majority of the intensity from the Si scattering is removed in the image processing (see Ref. 10), the resulting powder patterns may contain residual features, such as edges, due to shortcomings of this image processing. While this type of background is not well modeled by the typical polynomial or cubic spline profiles, Fig. 5 demonstrates that the wavelet transform analysis effectively removes this background and results in reliable identification of the Bragg peaks.

The common figures of merit for peak detection are the sensitivity and false detection rate, and the performance of a given peak detection algorithm will depend not only on the noise level but also on the nature of the background signal. A comprehensive evaluation of the performance of the wavelet-based algorithm *vis-a-vis* other peak detection algorithms is beyond the scope of the present work. We note that limited studies of this type have been performed in the context of mass spectrometry data.[5,7]

D. Python program and computation time

The algorithms of Secs. III B and III D have been implemented in the PYTHON programming language, and an open-source release of the code is in preparation. While Figs. 4 and 5 demonstrate the effectiveness of the peak detection algorithm, we point out that the algorithm is computationally efficient. As discussed in the Appendix, the values of the wavelets can be calculated from Eq. (4) and saved for analysis of any number of powder patterns. Given these values, the 20 powder patterns from the Pt–Ru thin film were analyzed with the PYTHON implementation of the wavelet transform and peak identification algorithm using a personal computer with a 3.07 GHz quad core processor (Intel Core I7 950). The 332 peaks were identified with an average analysis time per powder pattern of 0.4 s. This computation time is roughly proportional to the number of intensity measurements per powder pattern and the number of wavelet scales (1540 and 23 in this example).

V. CONCLUSIONS

We present algorithms for peak identification and subsequent background subtraction and profile fitting that are applicable to any type of 1D data. The peak identification algorithm involves the analysis of the wavelet transform surface, which is a noise-filtered representation of the data. A description of the general algorithm is given as well as detailed explanations for the "Lorentzian of Gaussian" wavelet analysis of powder diffraction spectra. Using example data from two diffraction experiments with different types of noise, we have demonstrated the ability of the wavelet-based algorithm to identify Bragg reflections in the diffraction patterns with very few false peak identifications. The resolution of strongly overlapped peaks in the wavelet transform is also discussed and demonstrated.

ACKNOWLEDGMENTS

The authors thank Brian Leonard and Francis DiSalvo for providing the PtZn nanoparticle powder diffraction data and for helpful suggestions. This material is based upon work supported as part of the Energy Materials Center at Cornell (EMC2), an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Award No. DE-SC0001086. The authors thank Alexander Kazimirov for assistance with the synchrotron XRD experiments, which were conducted at the Cornell High Energy Synchrotron Source (CHESS) with support from the National Science Foundation and the National Institutes of Health/National Institute of General Medical Sciences under NSF Award No. DMR-0225180.

APPENDIX: FURTHER DISCUSSION OF WAVELET ALGORITHM FEATURES

1. Wavelet modification for analysis of discrete datasets

The discrete convolution of data f with $w_{a,b}$ can be performed either in the coordinate space of the data [the discrete counterpart of Eq. (2)] or in its reciprocal coordinate space through multiplication of the discrete Fourier transforms of f and $w_{a,b}$. In both approaches, artifacts in the calculation result may arise due to the finite spacing in X and "edge" effects due to the termination of the data at the extreme values of X. We implement the following algorithm to circumvent these issues in the direct transform calculation.

The LoG wavelet has no zeroth or first moment, and thus in the support region of the wavelet (nominally $[-5a, 5a]$), constant and linear components of f do not contribute to the transform intensity. However, these desirable properties may be lost in a wavelet transform calculation if a is comparable to δx or b is within $\sim 5a$ of the extremes of X. In many data analysis algorithms, the latter issue is addressed by artificially extending the dataset through interpolation or reflection of f about the boundary. However, we propose that both problems are remedied by modification of the discrete wavelet

$$\mathbf{w}_{a,b}(x) = \begin{cases} \alpha w_{a,b}(x) & \text{if } w_{a,b}(x) \geq 0 \\ \beta w_{a,b}(x) & \text{if } w_{a,b}(x) < 0, \end{cases} \tag{A1}$$

where α and β are chosen such that $\mathbf{w}_{a,b}$ has no zeroth or first moment and has wavelet energy equal to that of the mother wavelet

$$\sum_X \mathbf{w}_{a,b}^2 \delta x = \int w_{a,b}^2 \, \mathrm{d}x. \tag{A2}$$

That is, the basic properties of the wavelet are restored by an asymmetric (vertical) stretch. The energy conservation provided by Eq. (A2) is desirable for any type of wavelet analysis, but significant modification of the wavelet through Eq. (A1) may introduce artifacts in the wavelet transform. The extent of wavelet modification that should be allowed will vary with the desired result of the wavelet transform, but we have found that reasonable results are obtained if the unmodified wavelet has an energy within 12% of the continuous wavelet energy. Otherwise, the measurement grid X is considered to be insufficient for the calculation of $T(a, b)$. While not utilized in the present work, we note that this technique may also be useful for cases of missing data (holes in X due to missing or corrupted measurements) or nongrid data (such as random spacing in X). In these cases, δx in Eq. (A2) must be replaced with the appropriate data point cross sections.

2. Resolution of overlapping peaks by wavelet analysis

Figure 6 demonstrates the resolution of strongly overlapping peaks in example data containing a pair of Gaussian and a pair of Lorentzian peaks. In both cases, the pair of peaks are separated by less than their FWHM and have a height ratio of 2:1. The sum of the peaks serves as the data f, which has a single local maximum, but the wavelet transform $T(0.25, b)$

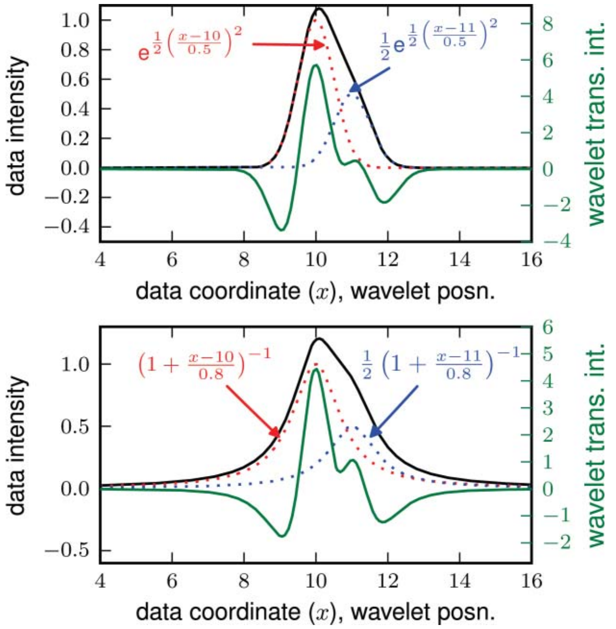

FIG. 6. (Color online) The LoG wavelet transform of strongly overlapping Gaussian (top) and Lorentzian (bottom) peaks. The sum of the overlapping peaks is shown (black line) with its wavelet transform (right axis) at scale parameter $a = 0.25$. The plots (dotted lines) and functions of the constituent peaks are also shown.

has two local maxima near the positions of the original peaks. For any pair of peaks with finite separation, $T(a, b)$ will have two local maxima with respect to b at sufficiently small a. However, the resolving power offered by wavelet transform analysis has two practical limitations. First, the lowest value of a that may be used for robust analysis is limited by the measurement grid X. Also, for peak detection with the algorithm of Sec. III B, the value of $T(a, b)$ at the local maxima must be larger than the specified noise level. This value is always greater than zero, and we note that for overlapping peaks with a large height ratio, $T(a, b)$ may be negative at one of the local maxima. Still, wavelet-based detection offers automated identification of overlapping peaks that is not afforded by searching the data for local maxima. In addition, the wavelet-based peak resolution algorithm is quite straightforward compared to the typical alternative method, which involves the identification of one of the peaks, profile fitting of that peak, and subsequent analysis of the residual.

3. Computational efficiency

a. Wavelet transform

For a given data analysis application, such as peak detection in powder patterns from a particular instrument, the wavelet parameters sets A and B can be used for the analysis of every dataset. If the measurement coordinate set X is the same for every dataset, the arrays of wavelet values can be calculated *a priori* using Eqs. (4) and (A1). With these arrays, each wavelet transform coefficient $T(a, b)$ is attained by computationally efficient array multiplication and summation. For example, using a personal computer with a 3.07 GHz quad core processor, the calculation of the arrays of wavelet values takes several minutes but only needs to be performed once. The entire algorithm of Sec. III B is then performed in <1 s.

b. Profile fitting

Profile fitting for peak parameter extraction typically involves three fit parameters for each peak (assuming symmetric peaks). Profile fitting algorithms commonly involve simultaneous optimization in each parameter, allow for a variable number of peaks, and include additional parameters for simultaneous fitting of the baseline profile. For the fitting of many peaks in a large set X, such profile fitting can be computationally expensive. The wavelet peak identification algorithm provides a fixed number of peaks with near-optimal starting parameters. Thus, the inclusion of the algorithm of Sec. III B as a first pass in a profile fitting scheme can offer increased computational efficiency.

In our profile fitting algorithm we further exploit the wavelet peak identification information by segmenting the dataset for profile fitting. With the peak positions and widths estimated from the wavelet-based algorithm, sets of overlapping or near-overlapping peaks are readily identified. Segments of the data coordinate X are then defined such that the segment containing each peak includes the interval of X within three widths of that peak. The data in each segment are then fitted using the appropriate subset of the identified peaks, and the ranges of X that are not in any such segment are not included in any profile fitting. In this scheme, several different profile fitting routines must be performed but each routine includes a relatively small number of fitting parameters and significantly reduced dataset size. For example, the five fitted peaks in Fig. 5 were fit in a single segment.

[1] L. Smrcok, Z. Kristallogr. **214**, 430 (1999).

[2] S. Bates, Adv. X-ray Anal. **42**, 251 (2000).

[3] L. Smrcok, M. Durik, and A. Jorik, Powder Diffr. **14**, 300 (1999).

[4] Y. Ding, T. Nanba, and Y. Miura, Phys. Rev. B **58**, 14279 (1998).

[5] P. Du, W. Kibbe, and S. Lin, Bioinformatics **22**, 2059 (2006).

[6] A. Wee, D. Grayden, Y. Zhu, K. Petkovic-Duran, and D. Smith., Electrophoresis **29**, 4215 (2008).

[7] A. Cruz-Marcelo1, R. Guerra1, M. Vannucci1, Y. Li, C. C. Lau, and T. Man, Bioinformatics **24**, 2129 (2009).

[8] A. Miura, H. Wang, B. M. Leonard, H. D. Abruña, and F. J. DiSalvo, Chem. Mater. **21**, 2661 (2009).

[9] J. M. Gregoire, R. B. van Dover, J. Jin, F. J. DiSalvo, and H. D. Abruña, Rev. Sci. Instrum. **78**, 072212 (2007).

[10] J. M. Gregoire, D. Dale, A. Kazimirov, F. J. DiSalvo, and R. B. van Dover, Rev. Sci. Instrum. **80**, 123905 (2009).

[11] R. K. Young, *Wavelet Theory and Its Applications* (Kluwer Academic Publishers, Boston, 1993).

[12] *Powder Diffraction File* (JCPDS International Centre for Diffraction Data, PA, 2004).