ELSEVIER

Chemometrics and Intelligent Laboratory Systems 41 (1998) 31–42

Chemometrics and intelligent laboratory systems

A structural approach for smoothing noisy peak-shaped analytical signals

A. Antonacopoulos [a,*], A. Economou [b,1]

[a] *Department of Computer Science, University of Liverpool, Peach Street, Liverpool L69 7ZF, UK*
[b] *Department of Instrumentation and Analytical Science, UMIST, PO Box 88, Manchester M60 1QD, UK*

Received 16 March 1997; accepted 25 March 1998

Abstract

This work describes an approach for smoothing noisy peak-shaped analytical data based on the identification of the structural form of the signal. The data series was described first as a succession of ‘peak structures’ and then as a succession of ‘meta-peak structures’. This description enabled convenient identification of the characteristic peaks arising from an analytical measurement and their separation from the noise components in the data. The method was applied to both voltammetric and spectroscopic data featuring different distributions of noise in the frequency domain. It was demonstrated that the suggested structural approach is successful in identifying the characteristic peaks with precision and, subsequently, in smoothing the test signals. The smoothing operation relying on the structural approach is fast, and, in contrast to traditional smoothing techniques, the fine detail of the signals is retained and no artefacts are generated as a result of the smoothing operation.

Keywords: Smoothing; Analytical measurement; Voltammetric and spectroscopic data; Noise; Structural description

1. Introduction

By definition, trace analysis techniques (such as electrochemical stripping analysis and atomic absorption spectroscopy) are specifically designed to determine low concentrations of analytes. However, in most cases, the limit of detection is dictated by the noise in the measurement step rather than limitation in the instrumentation side of the relevant analytical technique. As a result, the signal-to-noise (S/N) ratio (where S is the magnitude of the signal and N is half the peak-to-peak magnitude of the noise) becomes the main figure of merit for assessing the detection (and quantification) capability of a technique: the S/N ratio must exceed a certain numerical value in order for the presence and the amount of a species to be reported at a certain confidence level. Given that most trace analytical techniques operate at, or close to, their limiting sensitivity, the magnitude of the signal, S, is usually difficult to increase. However, many methods have been devised to minimise the contribution of the noise component, N. The noise-reduction methods range from hardware modifications (grounding and analogue filters) to on-line signal processing (real time FFT filtering, signal averag-

[*] Corresponding author.
[1] Present address: Laboratory of Analytical Chemistry, Chemistry Department, University of Athens, Panepistimiopolis, 157 71 Athens, Greece.

0169-7439/98/$19.00 © 1998 Elsevier Science B.V. All rights reserved.
PII S0169-7439(98)00040-9

ing) and off-line data processing (digital filtering, curve fitting) [1–3].

This work reports a novel off-line method for smoothing peak-shaped analytical signals. The method is based on the description of the experimental data series, first in terms of 'peak structures' and then in terms of 'meta-peak structures'; 'meta-peak structures' corresponding to analytical peaks are distinguished from structures corresponding to noise by observing certain differentiating criteria. Based on this discrimination, smoothing of the signal could be implemented. The suggested smoothing procedure was successfully applied to electrochemical and spectroscopic measurements that suffer from significant noise interference.

2. Theory

In this section, the rationale and the theory behind the proposed peak-identification and smoothing method are outlined.

The problem can be formulated as follows: Given a series of noisy measurements, identify and measure one or more characteristic peaks (if present) and smooth out the spurious peaks, i.e., the background noise. (A characteristic peak is one that signifies the presence of one of the substances to be measured).

Although the primary concern is the measurement of the characteristic peak(s), it is also desirable that the noise is smoothed out in the series resulting from the processing. Particular attention must be paid, however, so that the smoothing process does not affect the characteristic peaks. The true measurement details must be preserved as they are in the original data.

It becomes clear that the required course of action is: (a) to identify the characteristic peaks; and, (b) smooth the rest of the data. Therefore, the desired smoothing will be achieved and, most importantly, the measurement of the characteristic peaks will be unaffected. In the following subsections, the theory and process of each of these two phases will be presented and analysed.

2.1. Peak identification

The most important problem is the identification of the characteristic peaks in the data contaminated with noise. The distribution of noise across the frequency spectrum (white noise, sinusoidal noise and $1/f$ noise) and its intensity may be variable. However, a human observer is capable of thinking in terms of abstractions and has no great difficulty in classifying parts of the data into characteristic peak structures and noise structures. This is a classic example in which the structural differences between characteristic peaks and noise can be exploited in distinguishing between the two.

In order to identify the structural differences between characteristic peaks and noise, it is necessary to define what a 'peak structure' refers to.

> A *peak structure* (PS) is defined as the description of a peak-shaped consecutive set of data points in terms of a *left local minimum* (L), a local maximum (P) and a *right local minimum* (R). Each of these three reference points is a data point. All the other contributing data points (if any) lie between each pair of reference points (L and P, or P and R) and therefore: $\forall x \in \mathrm{PS} | x \neq \mathrm{L} \wedge x \neq \mathrm{P} \wedge x \neq \mathrm{R}$: $\mathrm{L} < x < \mathrm{P} \vee \mathrm{P} > x > \mathrm{R}$.

The whole series of the experimental data points can be represented as a series of consecutive PS's. Each PS_i ($1 \leq i \leq N$, N being the number of structures) is associated with a triplet (L_i, P_i, R_i). Each of these three reference points is a data point, which is represented by a pair (x_j, y_j) where $1 \leq j \leq M$, M being the number of data points.

The PS's are identified by sequentially examining each of the experimental data points and identifying local maxima and local minima.

The advantage of describing the data points as PS's is twofold. Firstly, a kind of abstraction is achieved which will allow more meaningful reasoning about the data. Secondly, the amount of data is reduced thus, increasing the efficiency of the method.

Some key observations about PS's can be made at this point. First of all, PS's corresponding to noise comprise the significant majority of peak structures. Another observation is that noise PS's have more similarities between them than they have with what are identifiable as characteristic peaks. The word *identifiable* is used here because it is observed that in many cases a characteristic peak is a composite of more than one PS's.

From the above it is clear that the PS's corresponding to noise must be identified and ignored but at the same time the PS's that comprise each characteristic peak must not be considered as noise. The latter PS's must be 'fused' into a new PS that corresponds to the characteristic peak. This requirement leads to the following definition:

> A *meta-peak structure* (MPS) is defined as the description of a peak-shaped set of P reference points (of consecutive PS's) in terms of a *left local minimum* (ML), a *local maximum* (MP) and a *right local minimum* (MR). Each of these three reference peaks is a P reference point of a PS. All the other contributing P's (if any), lie between each pair of reference peaks (ML and MP, or MP and MR) and therefore: $\forall x \in \text{MPS} | x \neq \text{ML} \wedge x \neq \text{MP} \wedge x \neq \text{MR}$: $\text{ML} < x < \text{MP} \vee \text{MP} > x > \text{MR}$.

The whole series of P's can be represented by a series of MPS's. Each MPS_i ($1 \leq i \leq K$, K being the number of structures) is associated with a triplet (ML_i, MP_i, MR_i). Each of these three reference peaks is a peak reference point P of a PS, which as mentioned earlier itself corresponds to a data point. An illustration of the relationship between an MPS, its PS's and the data points is shown in Fig. 1. In practical terms, in the identification of the series of the MPS's, the series of P's (from the PS's) are taken as input, while for the identification of PS's the input is the original data. In effect, the following set transformations take place: {data} → {PS} ⇒ {P} → {MPS}. The function ' → ' stands for a 'many to one' relation.

The procedure for deriving the MPS's is very similar to that used for the identification of the PS's. In this case, the P's in the series of PS's are sequentially examined and local maxima and local minima are identified.

The advantages of having described the data in terms of the series of MPS's are significant. An even higher abstraction is achieved in that the unnecessary details of local variations are suppressed while important information (location of characteristic peaks) is preserved. Data reduction has also been achieved. The PS's corresponding to noise have been fused into a smaller number of larger and smoother MPS's. Most importantly though, each characteristic peak is now described by an MPS as its constituent PS's have been fused. Therefore, a clear distinction between characteristic peaks and noise is effected because each structure is described by an MPS. This fact enables direct comparison of MPS's to identify each MPS corresponding to a characteristic peak. More-

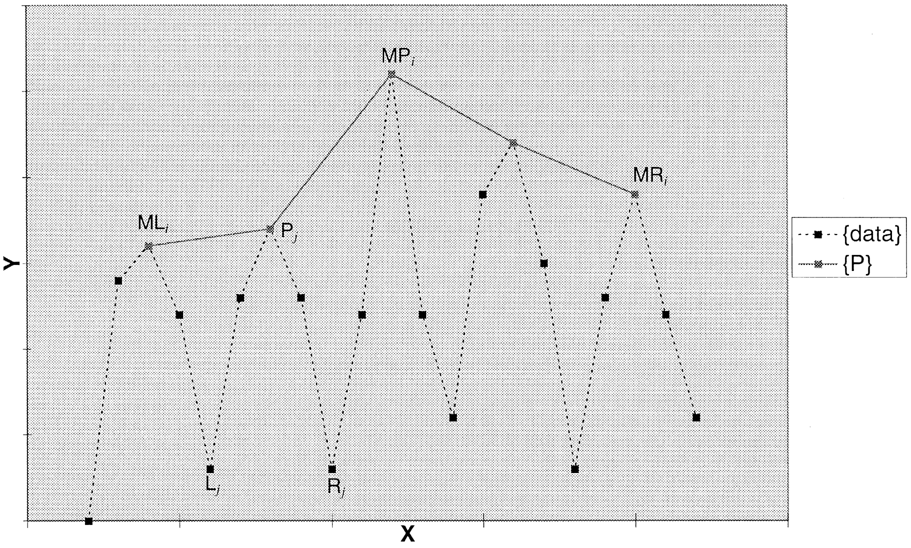


Fig. 1. An example of an MPS with corresponding PS's and data points.

over, no loss of information is incurred as the location and height of peaks remains the same as it is in the original data. There is a 1:1 correspondence between the MP value of an MPS and the underlying value of the original data point.

Having achieved a functional description of the data in terms of the series of MPS's, the next step is to identify and express the structural differences between a typical noise MPS and an MPS describing a characteristic peak. A measure of difference between MPS's that will separate the two types of MPS's is therefore required. This measure will comprise one or more *features* which will be employed in the classification of the MPS's into the two required categories. The choice of classification features is an important aspect in the effectiveness of the method, in terms of both accuracy and speed of execution. Consequently, a small number of features is desirable, each of which should be derived from an MPS using simple and fast operations.

It is observed that the *shape* of an MPS is a fundamental characteristic that differentiates between the two classes of MPS. A straightforward and in most cases sufficient description of shape can be expressed by width and height features:

The *width* of an MPS is defined as the scaled horizontal distance between ML and MR. More specifically, $\mathrm{MPS_w} = \mathrm{MR}_x - \mathrm{ML}_x$

The *height* of an MPS is defined as the greater of the (vertical) heights of the two legs of the structure: $\mathrm{MPS_h} = \max\{(\mathrm{MP}_y - \mathrm{ML}_y), (\mathrm{MP}_y - \mathrm{MR}_y)\}$.

To ensure consistency, both $\mathrm{MPS_w}$ and $\mathrm{MPS_h}$ are subsequently scaled with the maximum width and height correspondingly in the series.

Other shape related candidate features are given below.

The *width ratio* of an MPS is defined as the ratio of the smaller over the larger horizontal distance between the MP and the other two ends. More specifically, if $\mathrm{MPS_{lw}} = \mathrm{MP}_x - \mathrm{ML}_x$ and $\mathrm{MPS_{rw}} = \mathrm{MR}_x - \mathrm{MP}_x$ then

$$\mathrm{MPS_{w\_r}} = \begin{cases} \dfrac{\mathrm{MPS_{lw}}}{\mathrm{MPS_{rw}}}, & \text{if } \mathrm{MPS_{lw}} < \mathrm{MPS_{rw}} \\ \dfrac{\mathrm{MPS_{rw}}}{\mathrm{MPS_{lw}}}, & \text{otherwise.} \end{cases}$$

The *height ratio* of an MPS is similarly defined, given that $\mathrm{MPS_{lh}} = \mathrm{MP}_y - \mathrm{ML}_y$ and $\mathrm{MPS_{rh}} = \mathrm{MP}_y - \mathrm{MR}_y$ as

$$\mathrm{MPS_{h\_r}} = \begin{cases} \dfrac{\mathrm{MPS_{lh}}}{\mathrm{MPS_{rh}}}, & \text{if } \mathrm{MPS_{lh}} < \mathrm{MPS_{rh}} \\ \dfrac{\mathrm{MPS_{rh}}}{\mathrm{MPS_{lh}}}, & \text{otherwise.} \end{cases}$$

As expected from the observations, an MPS corresponding to a characteristic peak has in the vast majority of cases greater values for $\mathrm{MPS_w}$ and $\mathrm{MPS_h}$ than an MPS corresponding to noise. A very good classification of the MPS's into the two classes can therefore be obtained using only these two features. In the $\mathrm{MPS_w} - \mathrm{MPS_h}$ feature space a decision boundary is identified in the form of a straight line separating instances of the two classes. An MPS whose feature vector ($\mathrm{MPS_w}$, $\mathrm{MPS_h}$) lies to the left of and below the decision boundary is labelled as noise. After experimental validation, the equation of this straight line is chosen to be $y = 1 - x$.

It should be noted that other combinations of features could also be used such as the $\mathrm{MPS_w}$ with $\mathrm{MPS_{h\_r}}$ (noise MPS's tend to have uneven $\mathrm{MPS_{lh}}$ and $\mathrm{MPS_{rh}}$). However, the chosen approach is both computationally cheaper and there is less variability in the feature values.

In exceptional cases, it is noted that an MPS corresponding to noise has width and height that qualify it to be classified as a characteristic peak. For this to happen, the characteristic peak(s) in that data series must be quite low, at a comparable level to that of the highest noise MPS. The nature of the noise structure causing the exception is different from the background noise found in the measurement. It is attributed to a transient in the data series. Although this initial part of the measurement data can be easily eliminated without loss of significant data, the proposed structural approach can be refined to identify such MPS's and reject them.

The observed structural differences between a characteristic peak MPS and a transient MPS are as follows. Firstly, the transient MPS abruptly gains height and then slowly descends. In terms of the available shape features, this can be expressed with a small $\mathrm{MPS_{lw}}$ value and a quite larger $\mathrm{MPS_{rw}}$ one.

Therefore, such an MPS will have small MPS_{w_r} value. In contrast, MPS's corresponding to characteristic peaks have larger values of MPS_{w_r}.

Secondly, the right part of the transient MPS contains more P points from the PS series than the left part. A new feature is derived from this observation to complement the shape related ones:

The *peak-content ratio* is defined as smallest value of P content over the largest value of P content in an MPS. If in an MPS the MPS_{lpc} is the number of P's between ML and MP, and MPS_{rpc} is the number of P's contained between MP and MR,

$$MPS_{pc\_r} = \begin{cases} \dfrac{MPS_{lpc}}{MPS_{rpc}}, & \text{if } MPS_{lpc} < MPS_{rpc} \\ \dfrac{MPS_{rpc}}{MPS_{lpc}}, & \text{otherwise}. \end{cases}$$

Both the MPS_{w_r} and the MPS_{pc_r} are used to differentiate between characteristic peak MPS's and transient ones. It should be noted however, that these two extra features are calculated and considered only after the first classification, and for the characteristic peak candidates only.

The issue of exceptions is interesting as in the cases that peaks corresponding to transients were encountered as characteristic peak candidates the desired peaks were quite low. This fact might, in the first place, raise questions about the validity of the measured data and the level of confidence associated with the particular measurement. In any case, the inherent flexibility of the structural approach allows for the configuration of possible additional exceptions or indeed any other type of structure that should be differentiated from the rest.

Having presented and discussed the features, the algorithm used for the classification of MPS's into characteristic peaks and noise is outlined:

```
for each MPS
  if ($MPS_h + MPS_w > F_{hw}$) then
    if (($MPS_{w\_r} > F_{w\_r}$) AND ($MPS_{pc\_r} > F_{pc\_r}$))
      then MPS is a characteristic peak
    else MPS is noise
  else MPS is noise
```

The values of the feature decision thresholds have been experimentally determined as $F_{hw} = 1$, $F_{w\_r} = 0.11$ and $F_{pc\_r} = 0.11$.

2.2. Smoothing

The identification of the characteristic peaks enables the correct smoothing of the original measurement data. It is now possible to smooth all noise structures while leaving the characteristic peaks unchanged. Therefore, the objective is to smooth all PS's apart from those whose P is the MP of an MPS identified as a characteristic peak. Therefore, apart from the maximum point of the characteristic peak(s) all PS's will be smoothed including the PS's that comprise the left and right parts of the characteristic peak(s).

The smoothing in the proposed structural approach is performed by replacing the data points of a PS by a point corresponding to the centre of gravity of the PS. Assuming that there are N data points $\{(x_i, y_i) | i = 1, \ldots, N\}$ included between the L and the R points of a PS to be smoothed, the coordinates of the centre of gravity point (G_x, G_y) are calculated as:

$$G_x = \sum_{i=1}^{N} x_i / N \text{ and } G_y = \sum_{i=1}^{N} y_i / N.$$

This strategy produces a more accurate smoothing, minimising the side-effects caused by large differences in the size of the various noise PS's.

3. Experimental procedure

3.1. Reagents, experimental procedure and data collection

All the chemicals were of analytical grade or better. Ultra pure water (> 18 MΩ) was obtained from an Elgastat Maxima water purification system.

Electrochemical measurements of Ni(II) and Co(II) by adsorptive stripping voltammetry were carried out by making use of the experimental configuration reported earlier [4]. Ni(II) and Co(II) were complexed with dimethylglyoxime and the resulting complex was adsorbed on the surface of a rotating disk glassy carbon electrode covered with a mercury film. The accumulated complex was reduced by scanning the potential of the working electrode to the cathodic direction (in the square wave mode) which caused the appearance of peaks (corresponding to the reduction of Ni(II) and Co(II) in the complex) in the recorded current–potential curve.

Spectroscopic measurements involved the detection of Pb(II) by atomic absorption spectroscopy. Lead was atomised electrothermally in a Perkin Elmer HGA 600 graphite furnace employing a graphite L'vov platform and the absorption of its atoms was monitored at 283.3 nm as a function of time with a Perkin Elmer Model 3100 AAS, resulting in peak shaped absorption-time profiles.

3.2. Software implementation

The software realisation of the analytical peak identification and smoothing method was implemented in the C programming language. For the experiments referred to in this paper, the software was run on an IBM PC compatible based on an Intel 486DX microprocessor operating at 33 MHz. On average, for a typical data set of 512 points, run-times were in the very close region of 0.1 CPU s.

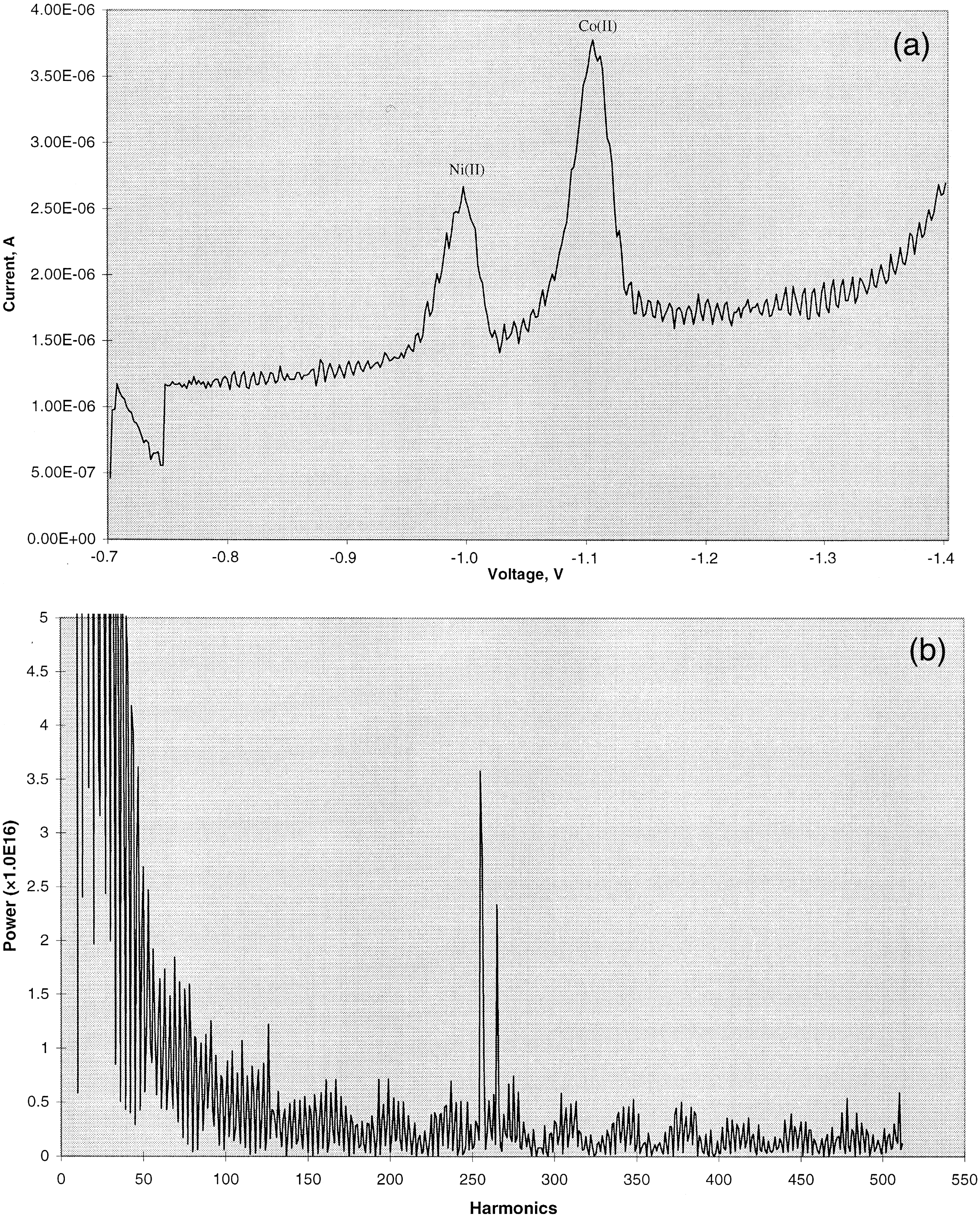
4. Results and discussion

A typical stripping voltammogram for Ni(II) and Co(II) is illustrated in Fig. 2a. The voltammogram features two peaks (arising from the reduction of Ni(II) and Co(II)), exhibits a sloping baseline (due to the charging current), features an S/N ratio of 12 (respective to the smaller Co(II) peak) and is contaminated with a transient at the beginning of the data series (probably due to some electrode effect). The power spectrum of this signal is illustrated in Fig. 2b. This spectrum indicates a strongly localised frequency component at the 255th harmonic which corresponds to the contribution of the square wave potential excitation signal. Contribution of noise at other harmonics, in terms of power, is small compared to the noise component at the 255th harmonic. A typical atomic absorption spectrum for Pb(II) with an S/N ratio of 8 is illustrated in Fig. 3a. Its power spectrum (shown in Fig. 3b) suggests that the noise is uniformly distributed across the frequency spectrum (i.e., it is white in nature). The signals in Fig. 2a and Fig. 3a are two examples of analytical signals with strongly correlated and uncorrelated noise, respectively. Since the given examples provide a satisfactory approximation of the two most widely encountered types of noise, they were selected for demonstrating the efficiency of the suggested smoothing procedure.

The results from the first step of the suggested procedure (i.e., first the identification of the peak structures and then of the meta-peak structures) in the voltammetric signal of Fig. 2a is illustrated in Fig. 2c. It can be seen that the identification of the peak locations was successful. The results of the second step (i.e., the actual smoothing) are illustrated in Fig. 2d. These results indicate that, in contrast to other smoothing procedures, the structural approach did not affect the fine detail in the data while at the same time a significant improvement in the S/N ratio was achieved. An important feature of the algorithm, as suggested from the results in Fig. 2d, is that it is capable of handling signals with sloping baselines which commonly occur in analytical measurements.

The initial results for the smoothing of the atomic absorption signal are illustrated in Fig. 3c (the first step of locating the peak structures is not shown for this signal). Although the initial S/N in this signal was worse than in the voltammetric signal (the concentration was actually close to the limit of determination), the analytical peak was successfully identified and the signal was smoothed, resulting in a significantly improved S/N ratio. Nevertheless, oscillations still exist after this initial smoothing of the signal. However, after the meta-peaks identification, the smoothing procedure can be iterated two or more times. This process results in further improvement of the S/N ratio, as shown in Fig. 3d. Although some noise remains after filtering, the technique is free

Fig. 2. (a) A typical stripping voltammogram for 10 nM Ni(II) and Co(II) adsorbed on a rotating mercury film electrode as their dimethylglyoxime complexes. Square wave scanning potential with frequency 40 Hz and pulse height 10 mV. Adsorption for 60 s at −0.7 V at a rotation speed of 10 Hz. Supporting electrolyte ammonia buffer pH 9. Dimethylglyoxime concentration 0.1 mM. (b) The power spectrum of the signal in (a). (c) The graph of PS's superimposed on the original signal of (a). (d) The smoothed signal of (a) superimposed on the original data.


(a)
Current, A
4.00E-06
3.50E-06
3.00E-06
2.50E-06
2.00E-06
1.50E-06
1.00E-06
5.00E-07
0.00E+00
Ni(II)
Co(II)
-0.7
-0.8
-0.9
-1.0
-1.1
-1.2
-1.3
-1.4
Voltage, V
(b)
Power (×1.0E16)
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
50
100
150
200
250
300
350
400
450
500
550
Harmonics

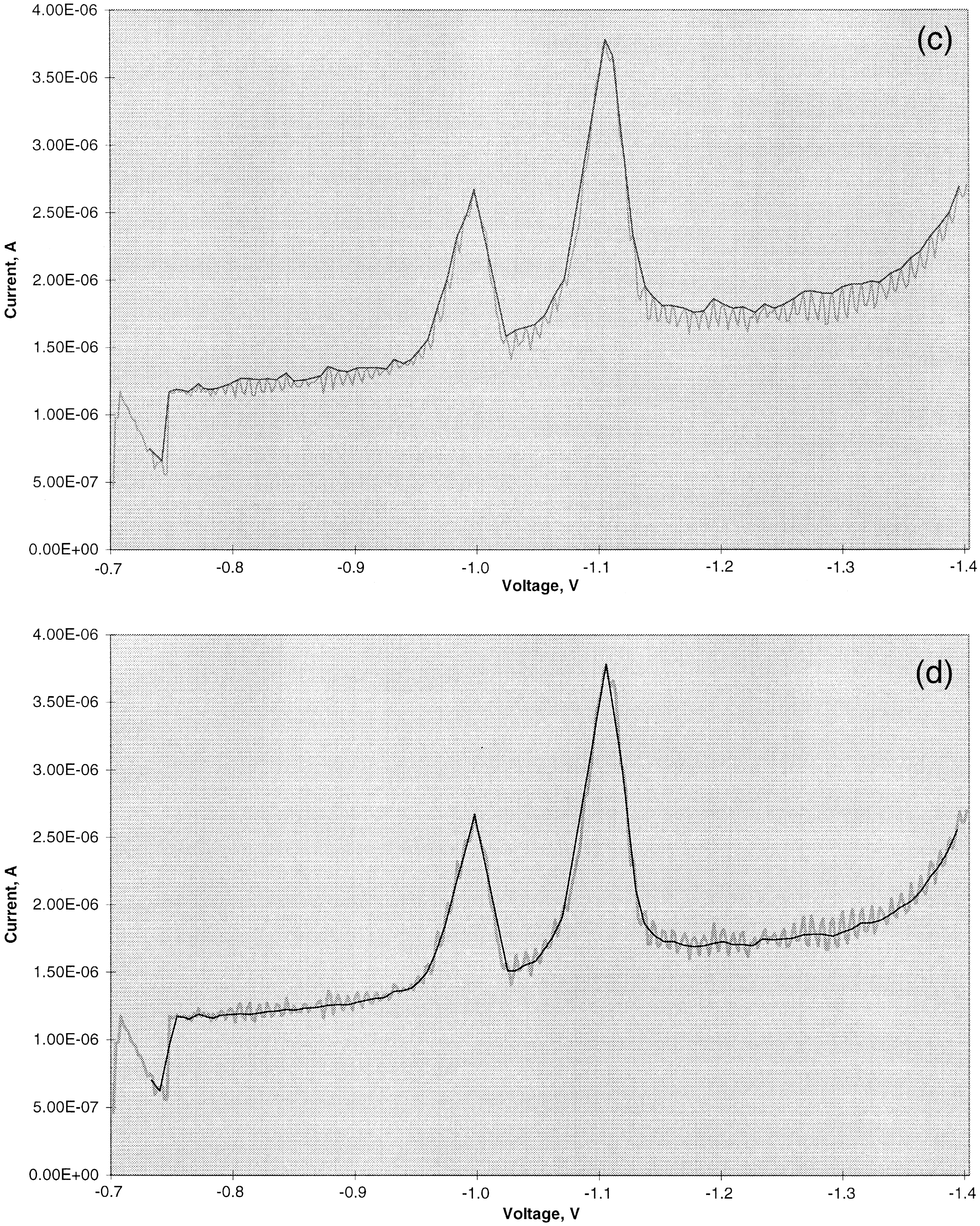


Fig. 2 (continued).

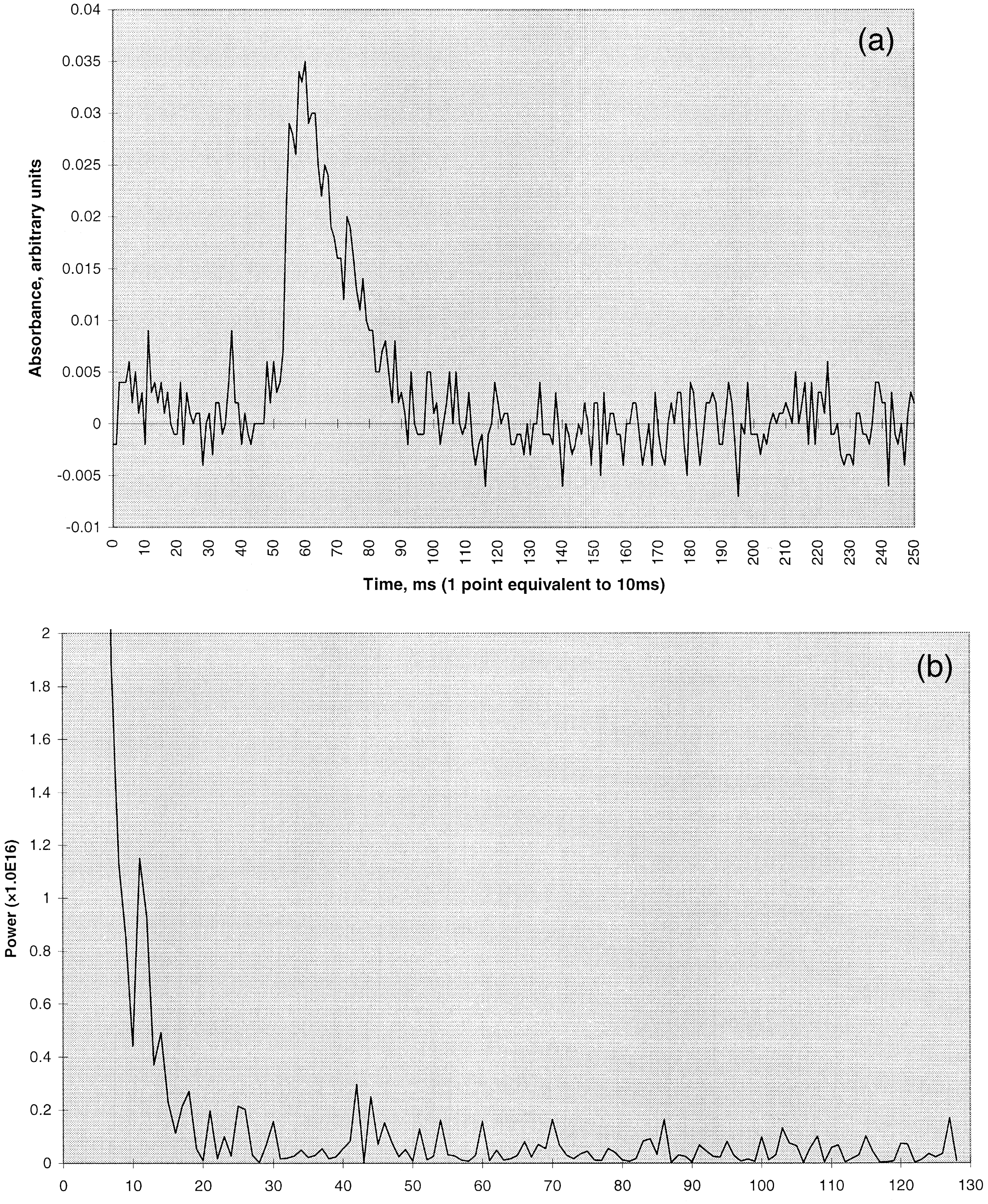
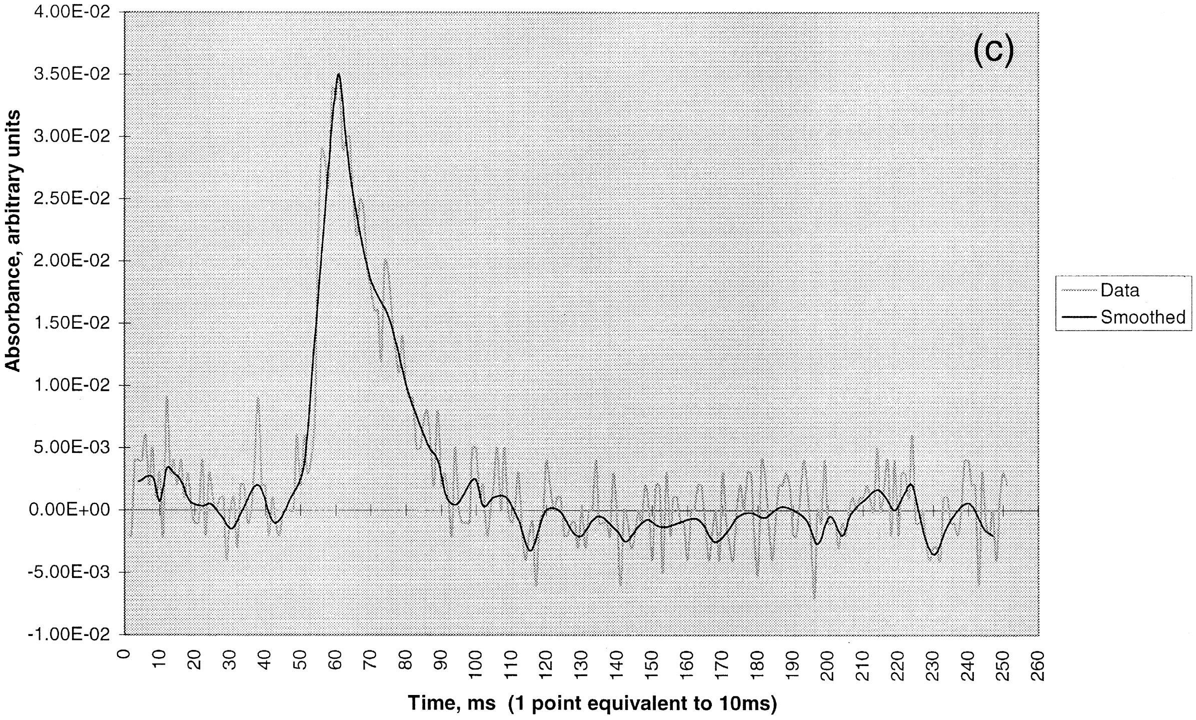


Fig. 3. (a) A typical atomic absorption signal for the determination of 2 ppb of Pb(II). Sample volume 20 μl. Matrix 0.1 M KNO_3. (b). The power spectrum of the signal in (a). (c) The initial smoothed signal of (a) superimposed on the original data. (d) The final smoothed signal of (a), after an extra iteration on the smoothed data of (c), superimposed on the original data.



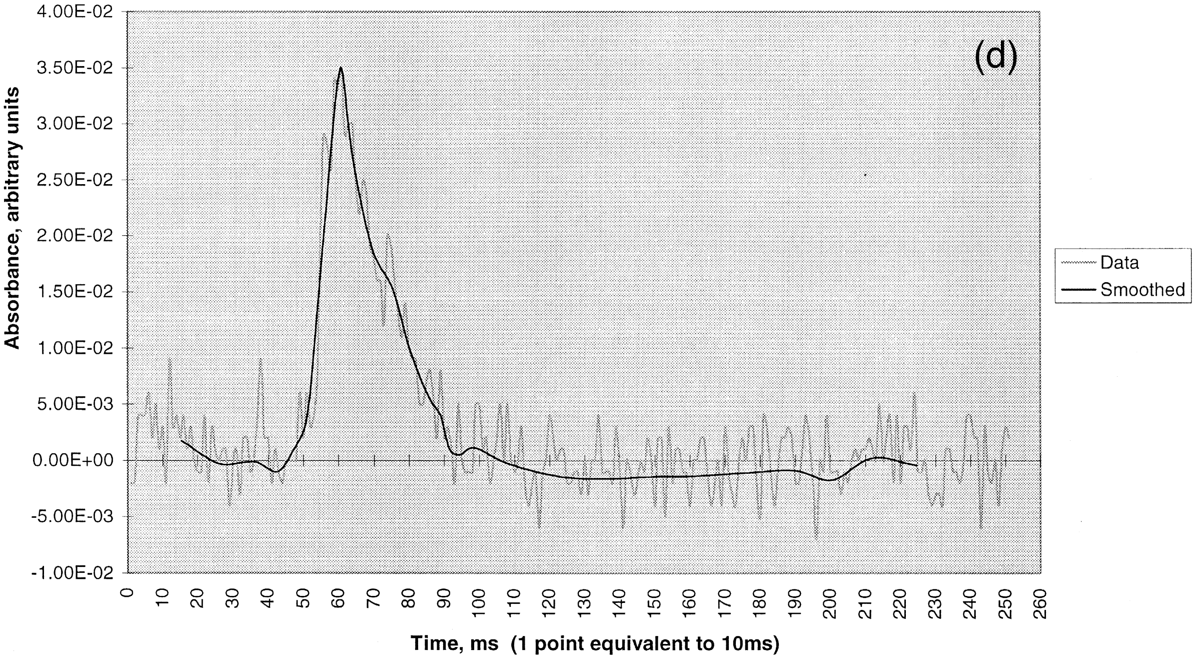


Fig. 3 (continued).

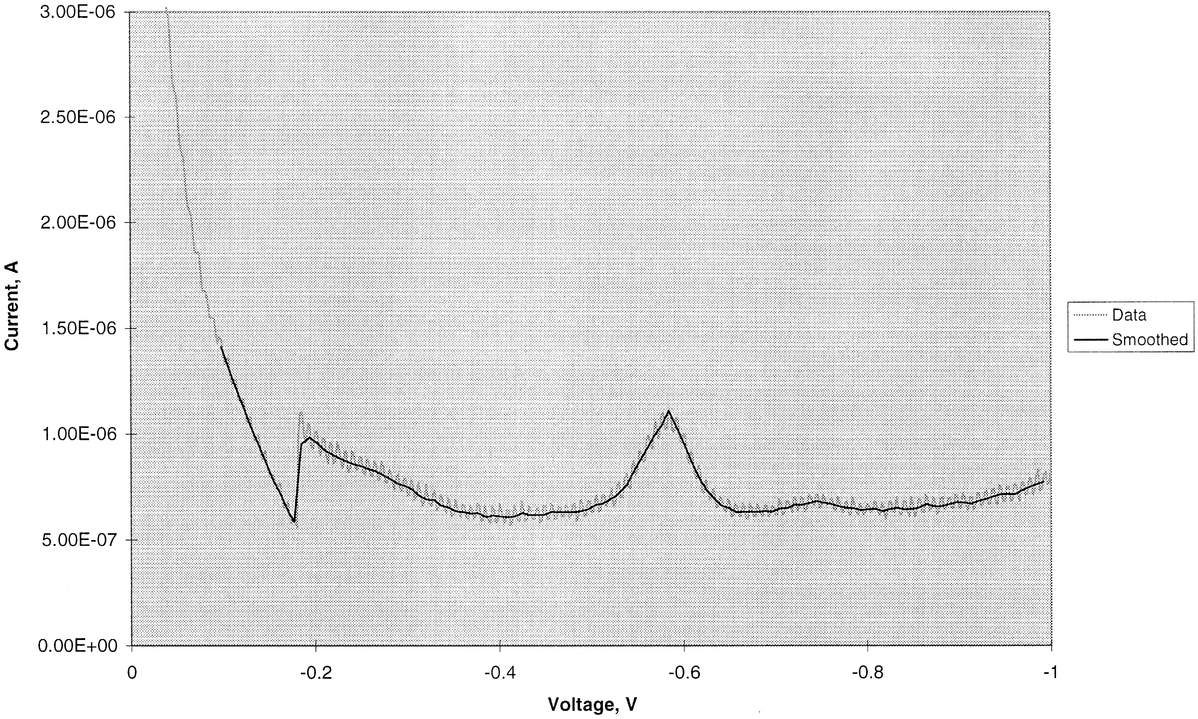


Fig. 4. A typical stripping voltammogram for 10 nM riboflavin adsorbed on a mercury film electrode. Square wave scanning at 40 Hz, pulse height 10 mV. Adsorption for 60 s at 0.0 V and at a rotation speed of 10 Hz. Electrolyte 0.01 M NaOH.

from disturbing artefacts such as the ‘Gibb’s oscillations’ that typically arise when Fourier filters are employed [5].

An important advantage of this particular implementation of structural filtering is the computational speed: for typical applications, only a fraction of a second is required. This is in contrast to other digital frequency domain and time smoothing procedures, which rely on the derivation of Fourier transforms or the time-domain multiplication of signals with filters impulse responses, respectively. Another advantageous property of the proposed strategy is that the original peak heights are maintained, as opposed to most frequency-domain procedures that typically cause a reduction in the peak heights proportional to the filtering efficiency [6].

To demonstrate the efficiency of the method developed in dealing with sudden transients in the signal (such as loss of electrode connection in electrochemical experiments) the method was also applied to such signals; an example was the determination of riboflavin by adsorptive stripping voltammetry (Fig. 4). The transient was correctly identified as such and not assigned as a peak.

5. Conclusions

In this work it has been demonstrated that the suggested structural approach can be successfully applied to the identification and smoothing of analytical peak-shaped signals featuring low S/N ratios (very close to the limiting sensitivity of analytical methods). The method is fast, insensitive to sloping backgrounds, can cope with transients in the data series and retains the fine detail of the underlying signal without affecting the analytical peak heights. The smoothing potential of the method can be increased by its iterative nature while the computational speed makes it preferable to other conventional filtering methods.

References

[1] H.A. Srobel, W.R. Heineman, in Chemical Instrumentation: A Systematic Approach, Chap. 12, Wiley, 1989.

[2] S.D. Brown, T.B. Blank, S.T. Sum, L.G. Weyer, Chemometrics, Anal. Chem. 66 (1994) 315R–359R.

[3] D.E. Smith, The enhancement of electroanalytical data by on-line fast fourier transform, Data Processing in Electrochemistry 48 (1976) 517.

[4] A. Economou, P.R. Fielden, Simultaneous determination of NI(II) and Co(II) by square wave adsorptive stripping voltammetry on mercury film electrodes, Analyst 118 (47–51) (1993) 517A–526A.

[5] J.W. Hayes, D.E. Glover, D.E. Smith, M.W. Overton, Some observations on digital smoothing of electroanalytical data based on Fourier transformation, Anal. Chem. 45 (1973) 277–284.

[6] A. Economou, P.R. Fielden, P.A. Gaydecki, A.J. Packham, Data enhancement in adsorptive stripping voltammetry by the application of digital signal processing techniques, Analyst 119 (1994) 847–853.