A Least-Squares Digital Filter for Repetitive Data Acquisition

SCOTT L. NICKOLAISEN and STEPHEN E. BIALKOWSKI*

Department of Chemistry and Biochemistry, Utah State University, Logan, Utah 84322

Received October 1, 1985

A least-squares digital filter is proposed for impulse-response data acquisition schemes. The filter operates in real time by fitting each signal transient to a filter function. The optimal filter function is found to be the signal waveform for the case of white noise. The signal-to-noise ratio improvement of the filter is proportional to the square root of the number of data points used in the filter-function waveform.

INTRODUCTION

Over the past 2 decades, the theory of digital filtering has been developed to a high degree. Digital reduction of noise in experimental data is now common place and is used extensively in spectroscopic and chromatographic signal processing.[1-6] In most techniques the digital filtering process is applied after the data have been acquired and stored in a digital format. This is due to the fact that filtering is most often performed in the frequency domain, requiring the transformation of data from time to frequency space. There are, however, several well-known filters that do not require transformation of the data into frequency space. Such filters as moving average, cubic spline, and Savitzky–Golay filters are effective at reducing high-frquency noise by time domain signal processing. More sophisticated filters are also being developed and applied to spectroscopic signal processing.[6] These filters are adaptive in that they change within the process of the filter, but they, too, require data transformation to the frequency domain. The time required for such filters to operate is long, but the filters are "matched" to the data, yielding maximum signal-to-noise ratio (SNR) output.

Repetitive impulse-response data are common in chemical analysis. Data collection of the pulsed signal waveform requires instrumentation that synchronously samples the signal. Two common devices for this are gated integrators and multichannel averagers. These devices generally perform simultaneous data collection and signal processing. They do not, however, allow analysis of samples that change in composition over time scales on the same order of the signal repetition rate. Gated detection without averaging results in noisy data, which must be processed with postexperimental smoothing filters. In this type of procedure, much of the experimental information is lost in the finite gate time and the recording process. Transient digitizers serve as a number of gated samplers that record data in a set time sequence. Multichannel averaging is used to perform the equivalent of gated integration on several time sequential data. However, multichannel averaging results in a loss of information for pulse-to-pulse dynamic experiments since several signal transients are averaged.

THEORY

In this work we use a simple least-squares filter to accomplish real-time filtering of individual transients. The important parameter to be determined is the signal magnitude. The filter is constructed so as to result in a maximum SNR in the result. With this filter, the best estimate of the signal magnitude can be made on a pulse-to-pulse basis. We first formulate the least-squares filter in terms of an impulse response function that uses the entire signal waveform to determine the signal data value in real time, thus eliminating the need for postexperimental filtering. We then show that the optimal impulse-response filter function is the expected signal in the limit of uncorrelated or "white" noise. The input signal waveform, $x(t)$, may be represented by

$$x(t) = s(t) + n(t) \tag{1}$$

where $s(t)$ is the noise free signal and $n(t)$ is the noise. The signal in turn is modeled as

$$s(t) = a_0 + a_1 h(t) \tag{2}$$

where the filter function $h(t)$ is a zero mean waveform and a_0 and a_1 are the coefficients to be determined by minimization of the square errors. In particular, a_0 is the time-invariant base-line offset, and a_1 is the magnitude of $h(t)$ in $s(t)$. The combination of eq 1 and 2 is an overdetermined set of equations at each time t, which are linear in their coefficients. Minimization of the errors with the Euclidean norm leads to the determination of the two coefficients[7]

$$a_0 = \sum x(t)/N \tag{3a}$$

$$a_1 = \sum h(t)x(t)/\sum [h(t)]^2 \tag{3b}$$

where the summations are performed over the N data points in the filter-function waveform. The simplicity of these equations is due to the zero mean characteristic of $h(t)$. It should be noted that a_1 is the product of the zero retardation cross-correlation of $h(t)$ with $x(t)$ and a constant. The time domain filter can be extended to cases where the impulse time of $s(t)$ is not the same as that of the filter function by using the correlation at other retardation times or the full correlation function.

The best estimate of the signal magnitude occurs for an $h(t)$ chosen such that the SNR is a maximum. This maximization results in an optimal linear filter known as a matched filter, which has been used for some time in electrical engineering.[6,8] The maximum SNR occurs when[6]

$$H_{opt}(f) = CS^*(f)/S_{nn}(f) \tag{4}$$

where $H_{opt}(f)$ is the optimal frequency or state function, C is a constant, $S^*(f)$ is the complex-conjugate Fourier transform of the expected signal waveform, and $S_{nn}(f)$ is the noise power spectrum. Division by the noise power spectrum is a "whitening" filter[9] in that the effect of this filter is to transform nonwhite Gaussian noise into white noise. In the frequency domain, data filtering corresponds to the product of the optimal filter function and the frequency-dependent input signal $X(f)$. The inverse Fourier transform of this product is the signal magnitude estimate, $\hat{s}(t)$:

$$\hat{s}(t) = \sum x(t')h_{opt}(t-t') \tag{5}$$

where $h_{opt}(t)$ is the optimal impulse response filter function and the summation is over all t'. For white noise, $S_{nn}(f) = N_0$ is a constant over the measurement frequency interval, and the inverse Fourier transform of eq 4 yields $h_{opt}(t) = Cs(-t)/N_0$. The convolution integral is thus equivalent to the

0095-2338/86/1626-0057$01.50/0

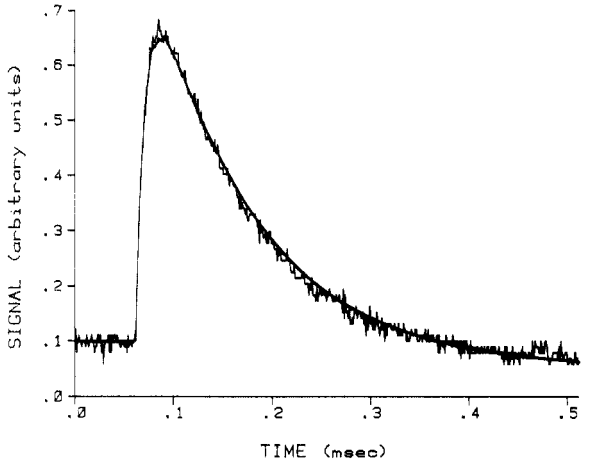

Figure 1. Filter-function (smooth line) fit to a single PDS signal with good SNR.

correlation of the expected signal with the input signal plus noise.[10] Of most significance here is the signal estimate at zero retardation time, which is proportional to a_1:

$$\hat{s}(0) = (C/N_0)\sum x(t)s(t) \qquad (6)$$

By comparing the above result with that of the least-squares estimate of a_1 in eq 2, it may be seen that the filter function is in fact $s(t)$ for white noise. In this case, the filter is a matched filter.[8] The difference in scaling between eq 2 and eq 6 is not important for the determination of relative signals.

EXPERIMENTAL PROCEDURES

The technique used to illustrate the proposed filter was photothermal deflection spectroscopy (PDS) as a gas chromatography (GC) detector. The theory and optical arrangement for this technique have been described in previous pulbications.[11-15] The pump laser was a TEA-CO_2 laser with a pulse repetition rate of 3.75 Hz. The probe laser was a 5-mW He–Ne laser operating at 632.8 nm. A bicell detector was used to monitor the transient probe laser beam deflection. The detector signal was processed with a sum/difference amplifier and an operational divider. This resulted in a signal that was proportional only to the beam position. The processed signal was passed through a high-pass filter, which served as a whitening filter,[13] and collected with an 8-bit transient digitizer. The CO_2 laser energy was monitored and sampled with a 12-bit, programmable gain analog-to-digital converter. The transient digitizer and A/D converter were interfaced with a DEC LSI 11/23 microprocessor.

To determine the filter-function waveform, a mixture of dichlorodifluoromethane in helium at a ratio of 1:6.2 was flowed through the sample cell. The CO_2 laser was tuned to the P16 line of the 10.6-μm transition, and the flow rate was 29 mL/min. The filter function was created by time averaging the signal under these conditions 500 times to make a very high SNR curve. Figure 1 shows a typical filter function and data for a high SNR experiment. The relative scale of these two plots is that determined by the regression procedure described under Theory. Four averaged replicates were summed to make filter functions for 64-, 128-, 256-, and 512-point fits. The transient digitizer was adjusted to 10, 5, 2, and 1 μs/channel, respectively, for each of the above filter-function waveforms. By increasing the number of points in the filter-function waveform, the density of the data points within the signal waveform of a given time duration was increased.

The GC runs were performed in a Hewlett-Packard Model 5890A gas chromatograph with a Porapak Q 80/100 6 ft × $^1/_8$ in. packed column. The analyte used was chlorodifluoromethane with helium as the carrier gas. The oven temperature was 120 °C, and the carrier gas flow rate was 32 mL/min. The CO_2 laser was tuned to the R16 line of the 9.6-μm transition, and 1.0 mL of the gaseous analyte was injected for each run. For the gated detection data acquisition run, the A/D converter and transient digitizer were switched so that the A/D converter sampled the signal and the transient digitizer sampled the CO_2 laser energy. The A/D converter gate was set to the point of maximum signal with a variable delay. A program was used that determined the base-line value by averaging the value of five pulses at the A/D converter gate with no analyte in the cell and then subtracting this value from the gate value during the GC run. For both the gated detection data acquisition and least-squares filter, the signal value was divided by the CO_2 laser energy to account for variations in its value.

Table I

data acquisition technique	base-line SD	S/N ratio
6-point gated detection	19.90×10^{-2}	50.3
64-point least squares	5.814×10^{-2}	172
128-point least squares	4.345×10^{-2}	230
256-point least squares	3.734×10^{-2}	268
512-point least squares	1.981×10^{-2}	505

RESULTS AND DISCUSSION

Identical GC runs were made by employing gated detection data acquisition and 64-, 128-, 256-, and 512-point least-squares fits with the optimal filter. In the optimal least-squares filter routine, the data value of the signal was considered to be the scaling coefficient of the filter function, a_1, since we were interested only in the height of the signal and not in any base-line deviations manifest in a_0. The height of the dichlorodifluoromethane peak was normalized to a value of 1, and the SNR of each chromatogram was determined by dividing the peak height by the standard deviation of the chromatogram base line. The results are shown in Table I. The SNR improvement between gated detection and the 512-point least-squares fit is approximately 1 order of magnitude. This is equivalent to time averaging 100 times. The advantage of this digital filter over time averaging for noise reduction is obvious, especially for flowing samples where time averaging a large number of times is impossible.

There are two possible sources of error in the proposed filter. The first is the error associated with the filter-function waveform used in the fitting routine. In theory, the error of the filter function could be reduced to the digitization error of the sampling instrument by time averaging a sufficient number of times, but this may be experimentally impractical because of the number of time averages required. In these experiments, the filter function was created by time averaging 500 signal pulses from a high analyte concentration mixture. The model used in the filter limits the SNR improvement. The model SNR of the filter function was measured to be approximately 2500. This is much greater than the SNR of the GC runs performed, so the error within the filter function can be assumed negligible. The second possible source of error is that associated with the actual fitting routine. The variance in the coefficient a_1 in linear least-squares fitting is

$$\sigma_{a_1}^2 = \sigma^2/\sum[h(t)]^2 \qquad (7a)$$

$$\sigma^2 = \sum[a_0 + a_1h(t) - s(t)]^2/(N-2) \qquad (7b)$$

Calculations of eq 7 for a single pulse in the GC runs with a 512-point fit give a relative variance in a_1 of approximately 2×10^{-5}. The contribution of this error to the SNR measured is also negligible. The SNR measured then can be attributed to instrumental and experimental errors such as mode variations and pointing noise in the lasers.[13]

R. R. Ernst has shown that the SNR improvement of white noise in a single scan performed in a total time, T_t, is equal to the SNR improvement achieved through time averaging of a signal response n times in the same total performance time; that is, $T_t = nt_s$ where t_s is the scan time of the signal averaging process. This SNR improvement over a single scan performed in time t_s is proportional to $T_t^{1/2}$.[16] In the least squares filter proposed here, the total performance time can be thought of in terms of the number of points, N_h, in the filter function, $h(t)$. Gated detection uses only some portion of the total signal waveform available in the given signal response cycle, whereas the least-squares filter uses the entire N_h points available to determine the signal value. Thus, the SNR improvement of the least-squares filter over gated detection is given by $(N_h/N_g)^{1/2}$, where N_g is the number of points averaged in the gated detection scheme. The gated detector employed in these experiments used a total of six points to determine the signal value. The maximum number of points used by the least-squares filter was 512, which gives a theoretical SNR improvement of 9.2. This is in good agreement with the actual SNR improvement measured to be 10.0. A plot was also made of the SNR of each GC run vs. the number of points used in the filter-function waveform. The plot was linear, meaning better SNR is obtained by using more data points in the filter function. The limiting factor in the number of points used is the speed with which the computer can perform the necessary operations.

ACKNOWLEDGMENT

This research was supported through a Utah State University Faculty Research Grant award.

REFERENCES AND NOTES

(1) Savitzky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627-1639.
(2) Bromba, M. U. A.; Ziegler, H. *Anal. Chem.* **1983**, *55*, 648-653.
(3) Bromba, M. U. A.; Ziegler, H. *Anal. Chem.* **1983**, *55*, 1299-1302.
(4) Leach, R. A.; Carter, C. A.; Harris, J. M. *Anal. Chem.* **1984**, *56*, 2304-2307.
(5) Baedecker, P. A. *Anal. Chem.* **1985**, *57*, 1477-1479.
(6) Dyer, S. A.; Hardin, D. S. *Appl. Spectrosc.* **1985**, *39*, 655-662.
(7) Bevington, P. R. *Data Reduction and Error Analysis for the Physical Sciences*; McGraw-Hill: New York, 1969.
(8) Papoulis, A. *Probability, Random Variables, and Stochastic Processes*; McGraw-Hill: New York, 1965.
(9) Wozencraft, J. M.; Jacobs, I. M. *Principles of Communication Engineering*; Wiley: New York, 1967.
(10) Brigham, E. O. *The Fast Fourier Transform*; Prentice-Hall: New York, 1974.
(11) Long, G. R.; Bialkowski, S. E. *Anal. Chem.* **1984**, *56*, 2806-2811.
(12) Long, G. R.; Bialkowski, S. E. *Anal. Chem.* **1985**, *57*, 1079-1083.
(13) Long, G. R.; Bialkowski, S. E. *Anal. Chem.*, in press.
(14) Nickolaisen, S. L.; Bialkowski, S. E. *Anal. Chem.* **1985**, *57*, 758-762.
(15) Nickolaisen, S. L.; Bialkowski, S. E. *Anal. Chem.* **1986**, *58*, 215-220.
(16) Ernst, R. R. *Rev. Sci. Instrum.* **1965**, *36*, 1689-1695.

Comparison of Manual and Online Searches of *Chemical Abstracts*

E. AKAHO,* A. BANDAI, and M. FUJII

Faculty of Pharmaceutical Sciences, Kobe-Gakuin University, Ikawadani-cho, Nishi-ku, Kobe, 673 Japan

Received November 7, 1984

Manual and online searches of *Chemical Abstracts* on five selected topics were conducted to compare the cost effectiveness, relevance factor, and search characteristics of the two methods. It was found that the online search was more expensive when the cost calculation was based on the part-timer's salary while it was less expensive when it was based on the professional worker's salary. A universal equation to evaluate the overall cost effectiveness of the search was proposed. The equation takes into consideration such factors as relevance factor, recall factor, cost factor and time factor and gives a value of "1" when the search is most cost effective. The actual application of this equation for the five selected topics gave 0.778 for the online search and 0.736 for the manual search, which indicates that the online search is a little more cost effective than the manual search. Each method of searching has its own merits and demerits, and a practice of using a single method of searching sometimes gives an incomplete search result.

INTRODUCTION

Chemical Abstracts (CA), founded in 1907 by the American Chemical Society, is now the largest and oldest abstract journal in the field of chemistry. It deals not only with the area of pure chemistry but also with the surrounding areas of chemistry such as biochemistry, applied chemistry, and so on. It is surprising to note that a high proportion of articles dealt with by CA is rather biology oriented than chemistry oriented. A considerable number of articles can be retrieved by using biology-oriented questions.

This means that CA has become more a comprehensive information source than before and that its scope of usefulness has been expanded. It means, at the same time, that it has become more complicated than before and that it has become more difficult to search and retrieve appropriate articles.

Online information search and retrieval is widely accepted in various areas of sciences, and its usefulness and cost effectiveness as an alternative to manual searching have been examined from different points of view.[1-8] A small but focused study conducted by Michaels compared the comprehensiveness of searches performed by using *CA Condensates* online with that of the manual searches in keyword indexes of CA.[9] The results were not conclusive, but she pointed out the subject and vocabulary problems that are specific to each mode of searching.

The cost involvement of online retrieval can be discussed on such aspects as computer connect time, telephone fee, staff's salary, equipment cost, etc. There can be attempted to manipulate those variables to establish a universal formula that can be used to calculate the total online cost. A cost of the manual search should be formulated as well. And when those two types of proper formulas are established, the true comparison of the two methods of information retrieval can be done. But whether or not the establishment of those formulas is meaningful is yet to be discussed. Anyhow, what is worth being done at this moment is to obtain as many results of comparative studies as possible. A project was initiated to perform online and manual searchings using *Chemical Abstracts* to compare the two searchings.

METHODOLOGY

Types of Questions Used. Considering the fact that a

0095-2338/86/1626-0059$01.50/0