Journal of Neuroscience Methods, 40 (1991) 71–86

NSM 01298

Forward-backward non-linear filtering technique for extracting small biological signals from noise

S.H. Chung [1] and R.A. Kennedy [2]

[1] *Research School of Biological Sciences, Australian National University, GPO Box 4, Canberra ACT 2601 (Australia)* and [2] *Department of Systems Engineering, Research School of Physical Sciences and Engineering, and Interdisciplinary Engineering Program, Faculty of Science, Australian National University, GPO Box 4, Canberra ACT 2601 (Australia)*

(Received 17 January 1991)
(Revised version received 22 April 1991)
(Accepted 29 April 1991)

Key words: Digital signal processing; Non-linear filtering; Noise reduction; Signal extraction; Ion channels; Patch clamp; Channel currents; Post-synaptic currents

A novel and computationally efficient, non-linear signal processing technique for reducing background noise to reveal small biological signals is described. The signal estimate is formed by weighting the outputs of a set of causal (forward) and anti-causal (backward) predictors. The weights used to combine the predictors are adaptively determined at each data point to reflect the performance of the respective predictor within a short analysis window. The method is specifically designed for revealing fast transient signals dominated by noise, such as single-channel or post-synaptic currents. Markovian and exponentially decaying signals embedded in the amplifier noise were extracted using this method and compared with the original signals. The results of such simulations demonstrate the advantage of this non-linear method over low-pass filtering. Brief pulses imbedded in a broad-band amplifier noise can be reliably recovered using our non-linear filtering technique. Moreover, the kinetics of a single channel and the time constant of exponentially decaying signals can be measured with acceptable accuracy even when the signals are dominated by noise.

Introduction

Extracting a real signal from a limited set of imperfect measurements is a problem that commonly occurs in science. When a desired signal is small relative to the background noise, the signal-to-noise ratio can be improved by low-pass filtering, of which a moving average and a low-pass Butterworth filter are examples. However, this type of linear processing tends to obscure and distort fast transients of the signal. For example, even in the absence of noise a moving average processing of a signal consisting of a step change shows the characteristic distorted ramp output – a phenomenon commonly labelled edge blurring.

Alternative signal processing methods have been developed in an attempt to improve the noise suppression without the signal distortion, e.g., (i) Kalman filtering (Anderson and Moore, 1979), (ii) Hidden Markov Model techniques (Rabiner, 1989; Chung et al., 1990), and (iii) maximum likelihood sequence estimation using the Viterbi algorithm (Forney, 1973). However, these techniques have shortcomings in that they tend to be computationally expensive and they rely on an explicit model of the hidden signal. Indeed, the effectiveness of such procedures can hinge critically on an accurate specification of the characteristics of the signal which in reality may be poorly understood. For example, Kalman filtering and its optimality is based on knowledge of an accurate linear model describing the dynamics of the desired signal. For the biological signals that we wish to consider determining the underly-

ing signal generation mechanism is the ultimate objective, so an accurate model is not known a priori. Furthermore, the measurements we are processing reveal that a linear dynamical model would be a poor representation of the mechanism behind the signal because of the jump transients observed.

We introduce here a novel non-linear digital filtering technique, that is computationally inexpensive, assumes no explicit single idealised model describing the signal and is designed to give the minimum distortion of fast transients. The results of the method display comparable performance to classical Hidden Markov Model techniques but with a reduction of up to two orders of magnitude in computation and memory requirements. To obtain an estimate of the signal, our technique utilizes several causal predictors, running forward in time, with a number of mirror-image anti-causal predictors, running backward in time. The final smoothed estimate is obtained as a data-dependent weighted combination of the outputs of all the backward and forward predictors and is designed to avoid the introduction of the artefacts which distort the estimation of original signal.We emphasize that the weights which combine the predictors' outputs are a function of the data and therefore are time varying or adaptive. The technique has been developed specifically for, but its application is not limited to, analyses of single-channel currents recorded from membrane patches and post-synaptic potentials or currents recorded intracellularly.

In the first section of the paper, we briefly describe the theoretical basis of the non-linear noise rejection scheme. A fuller and more rigorous account of the theory is given in the Appendix. To validate the technique, we have applied our technique to reveal small, computer-generated signals embedded in noise generated by a resistor/patch-clamp amplifier/filter system. The results of these simulations, detailed in the second section, show the superiority of our technique compared to low-pass filtering or other noise filtering operations. These results indicate that the technique can be particularly useful for extracting brief channel openings, which are obliterated when a conventional analogue low-pass filter is used before sampling during measurement. Moreover, the single channel kinetics or the amplitude and time constants of post-synaptic currents can be deduced with an acceptable degree of accuracy even when signal amplitudes are relatively small compared to the background noise.

Theoretical background

1. Non-linear filtering

The technique that we develop, which we shall refer to as non-linear filtering, is suitable for removing noise from an underlying signal having jump or abrupt changes. Such a noisy signal might arise from instrument noise contaminating a laboratory measurement of some jumpy process for example. In this section we give an overview of the theoretical underpinnings of our filtering scheme and provide some insights into its rationale. We also compare at an intuitive level non-linear filtering with the standard technique of linear filtering.

2. Motivation

Suppose our desired signal is represented by the sequence of values

$$x(0),\ x(1),\ x(2),\ \ldots$$

which we write as $x(k)$, with k denoting time. Now let

$$y(k) = x(k) + n(k);\ k = 0,\ 1,\ 2,\ \ldots$$

represent our sequence of measurements $y(k)$ of our desired signal $x(k)$, where $n(k)$ is some unknown, unwanted, additive noise. The standard means for suppressing this noise and thereby estimating the unknown desired signal $x(k)$ is to filter the measurements $y(k)$. That is, we combine a group of measurements of $y(k)$ and (usually) perform some averaging computation, e.g., let

$$\hat{x}(k) = \tfrac{1}{3}y(k-3) + \tfrac{1}{3}y(k-2) + \tfrac{1}{3}y(k-1),$$
$$k = 0,\ 1,\ 2,\ \ldots \qquad (1)$$

be one such estimation procedure for $x(k)$. We refer to the time interval $(k-3,\ k-2,\ k-1)$ as

the window of the estimator $\hat{x}(k)$ above. (This type of "moving average" estimator is easily generalized to other window lengths.) What we now argue is that such a linear estimation procedure is ineffective for the sort of low-level jump signals that we wish to estimate.

We begin by understanding when the above linear estimator (Eqn. 1) gives a satisfactory estimate of $x(k)$ – this will be important in recognising why our non-linear filtering can give improved noise rejection with less signal distortion. For acceptable performance of the above estimation scheme (Eqn. 1) we need the signal $x(k)$ to be reasonably constant over the window of the filter. That is, we require $x(k-3) \approx x(k-2) \approx x(k-1)$. This condition ensures that under filtering the signal values combine constructively. Contrast this with the situation where the underlying signal has a jump within the window of the estimator e.g., $x(k-3) \approx x(k-2)$ but $x(k-1) \approx -x(k-2)$. Now the signals under averaging will tend to combine destructively, thereby on average reducing the performance of the filter relative to the level of noise suppression. In short, linear filtering is only effective if the underlying signal is slowly or smoothly changing or the mechanism behind the signal variations is known to be accurately modelled by a set of linear difference equations. However, the technique generally is not very suitable for signals having jump changes – the major reason being one cannot get an accurate "smooth" linear model for such a process which is fixed over any extended time interval.

3. Forward–backward processing

The means to avoid the edge blurring mechanism is conceptually quite simple and forms the basis of the heuristic justification of our technique. The principle that we employ is to avoid at all costs filtering or averaging across jump changes.

In our scheme (see Appendix) we combine predictors based on different non-overlapping windows of data. As an illustration of the principle suppose we have an unbiased forward predictor given by

$$\hat{x}^f(k) \triangleq \tfrac{1}{3}(y(k-3)+y(k-2)+y(k-1))$$

and an unbiased backward (anti-causal) predictor given by

$$\hat{x}^b(k) \triangleq \tfrac{1}{3}(y(k+1)+y(k+2)+y(k+3))$$

each estimating the desired signal $x(k)$ but based on different non-intersecting windows of noisy data. Our filtering scheme combines these two individual estimates, to produce a final non-linear filtering estimate, by weighting them as follows:

$$\hat{x}(k) = f(k)\hat{x}^f(k) + b(k)\hat{x}^b(k) \tag{2}$$

where $0 \leq f(k) \leq 1$, and $0 \leq b(k) \leq 1$ are the forward and backward weights respectively that satisfy

$$f(k) + b(k) = 1 \tag{3}$$

to ensure that there is no bias in the final estimate. Note that the weights which are yet to be explicitly defined are not constants but functions of the data and thus time making the scheme non-linear and time varying. If the weights were indeed constant then the overall scheme in (Eqn. 2) would be linear and edge blurring would generally occur.

The idea behind the selection of the weights $f(k)$ and $b(k)$ is to select automatically the forward predictor, $f(k) \approx 1$, if the jump changes occurs in the data set $\{y(k), y(k+1), y(k+2), y(k+3)\}$, or automatically select the backward predictor, $b(k) \approx 1$, if the jump changes occurs in the data set $\{y(k-3), y(k-2), y(k-1), y(k)\}$. Note that with no jump change occurring in either data interval then both predictors will present acceptable signal estimates, and therefore the relative weightings of predictors is not crucial. Now we move on to give the formulation for our general estimation scheme which uses banks of predictors.

4. Processing using banks of predictors

It is easy to incorporate, rather than a single forward moving estimator, a bank of K forward moving unbiased predictors, under a common weighting rule. The idea of using a bank of predictors is that predictors with short window lengths work well on fine structure of signals whilst predictors with long windows do better on broad signal features. So we can expand on the

simple scheme (Eqn. 2) as follows. Our estimate at time k takes the form (see Appendix)

$$\hat{x}(k) = \sum_{i=1}^{K} \left[f_i(k)\hat{x}_i^f(k) + b_i(k)\hat{x}_i^b(k) \right]$$

where

$$f_i(k) \propto \pi_i \left(\sum_{j=0}^{M-1} \left[y(k-j) - \hat{x}_i^f(k-j) \right]^2 \right)^{-p} \quad (4)$$

$$b_i(k) \propto \pi_i \left(\sum_{j=0}^{M-1} \left[y(k+j) - \hat{x}_i^b(k+j) \right]^2 \right)^{-p} \quad (5)$$

are the weights which are normalized by

$$\sum_{i=1}^{K} f_i(k) + \sum_{i=1}^{K} b_i(k) = 1,$$

and M is a parameter which determines the size of the window of data over which the predictors are to be compared (the analysis window), typically $4 \leq M \leq 20$ depending on the data. The constants π_i denote a priori probabilities reflecting any prior knowledge as to which predictor indexed by i should be preferred on average in processing the noisy signals. These constants represent additional degrees of freedom to customize the set of predictors which in the most conservative setting can be made equi-probable (i.e., $\pi_i = (2K)^{-1}$ for all i). Indeed, for many of the simulations presented in this paper these prior probabilities were chosen to be equal (and sum to unity), e.g., if there are 4 forward and 4 backward predictors then $\pi_i = 0.125$ for all i. The parameter p (which may be chosen independently from M) is a weighting factor whose effect is described in the next section. In our simulations later we display the effect of values between $p = 1$ and $p = 100$, with typically higher values giving better results.

The proportionalities expressed in Eqns. 4 and 5 would need special treatment in any practical implementation when any predictor gave the measured signal as its prediction for all samples within the analysis window. (Such a case may arise when there is no measurement noise which we consider next.) This would imply that the right hand side of Eqn. 4 would represent a division by zero. However, as the $f_i(k)$ and the $b_i(k)$ are merely probabilities (see Appendix) there is no actual explosion of these coefficients (under the normalization). In a programme one would simply check for a zero before dividing in the calculation of $f_i(k)$, say, and in this event make $\hat{x}(k) = \hat{x}_i^f(k)$ (similarly if there was a potential division by for a backward coefficient calculation). (If more than one coefficient lead to a division by zero this is not a problem because the predicted value of $x(k)$ would be identical.)

Implicit in Eqns. 4 and 5 is that a noise-free stepped signal processed with at least one forward predictor of the form of a moving average (as in Eqn. 1) of length less than the duration of the shortest feature of the signal, along with its mirror image backward version, will leave the signal undistorted. Also as with many other filters, the estimation error of the non-linear filter, measured as the sum of differences between estimated and original signals, will decrease progressively as the signal-to-noise ratio increases. Finally, as a guide to the limitations of the non-linear filter, the shortest pulse that would be reliably be recovered though not without some distortion would be not less than 4 samples.

Simulations

1. Experimental set-up

The reliability and the limitation of our non-linear filter was assessed by an extensive simulation study using known signals buried in noise. Unless stated otherwise, the background noise was simulated by recording the output of a patch-clamp amplifier (Axopatch 1C, Axon Instruments) with a 10 GΩ resistor across the input. The output was either filtered at 2 kHz and sampled at 5 kHz or filtered at 20 kHz and sampled at 40 kHz. The noise, when filtered at 2 kHz and 20 kHz (−3 dB, Bessel), was Gaussian with standard deviation 0.10 pA and 0.39 pA, respectively. To this noise was added either: (i) rectangular pulses of various durations and amplitudes, or (ii) step responses that decayed exponentially with various time constants. The noisy traces containing the signals were then processed

with the non-linear filter to evaluate the increase in signal-to-noise ratio. Although short segments of the results are illustrated in the following figures, typically we processed traces containing 20 000 to 50 000 points, using a workstation computer (Sun IV).

2. Characteristics of the non-linear filter

We processed a 20 000-point segment of the computer-generated noise containing no signal with the non-linear filter using 3 forward and 3 backward predictors. The predictors were moving averagers of window lengths 4, 8, and 16. The

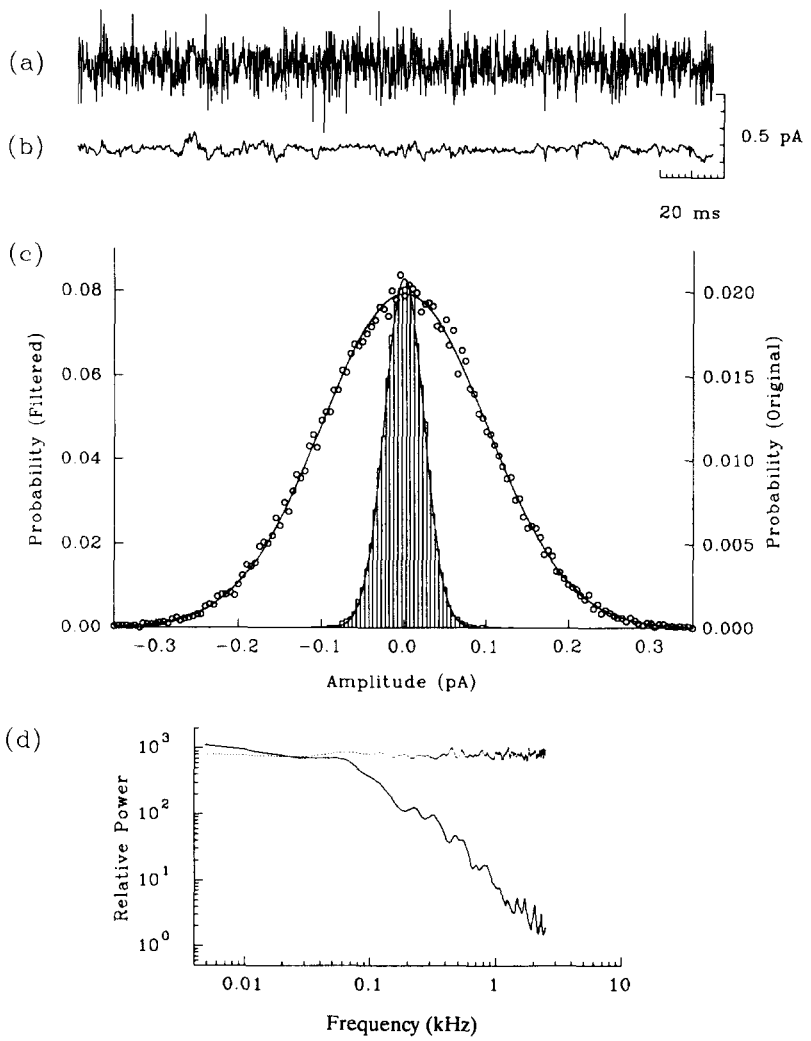

Fig. 1. Characteristics of original and filtered noise. A 20 000-point random noise was generated by a computer, with the mean and the standard deviation of 0 and 0.1 pA. The digitizing interval was assumed to be 200 μs. A sample segment of this noise, shown in (a), is similar to the noise recorded from a patch-clamp amplifier, passed through a low-pass filter with the cutoff frequency of 2 kHz. The noise was then processed with the non-linear filter with 3 pairs ($\pm$4, $\pm$8 and $\pm$16) of predictors. (In this and subsequent figures, the forward and backward predictors of, for example, 16-points in window length are abbreviated as $\pm$16.) A segment of the processed trace is shown in (b). The amplitude distributions of both original (open circles) and processed data (bars) could be fitted with a Gaussian, as shown in (c). The power spectra, shown in (d), of the original (dotted line) and processed data (continuous line) were obtained by the Maximum Entropy Method, using 64 autoregressive coefficients.

prior probabilities were selected in proportion of the window lengths, i.e., 1:2:4. A comparison of the sample trace of the original record with a corresponding trace of the processed record, shown in Fig. 1a, reveals the effectiveness of the method in reducing random noise. The amplitude probability density curves of the original noise and the processed noise are superimposed in Fig. 1b. The standard deviation of the processed noise was reduced to 28% of the original value, from 98.8 fA to 27.8 fA. The 3rd (skew) and 4th (kurtosis) moments of the distribution, however, remained unchanged. The power spectra of the original noise and non-linear filter output, obtained by the Maximum Entropy Method (Childers, 1987), reveals that our non-linear filter has noise suppression performance comparable to a low-pass filter in the frequency domain (see Fig. 1c). However, we caution that our signal processing structure is a non-linear device (because the weights are data dependent). That is, the common frequency domain intuition is based on linear filtering experience, so the time domain response of our non-linear structure need not closely resemble the comparable bandwidth linear filter particularly near signal jump changes.

3. *Parameters of the non-linear filter*

In using the non-linear filter, there are three main sets of parameters which can be adjusted to

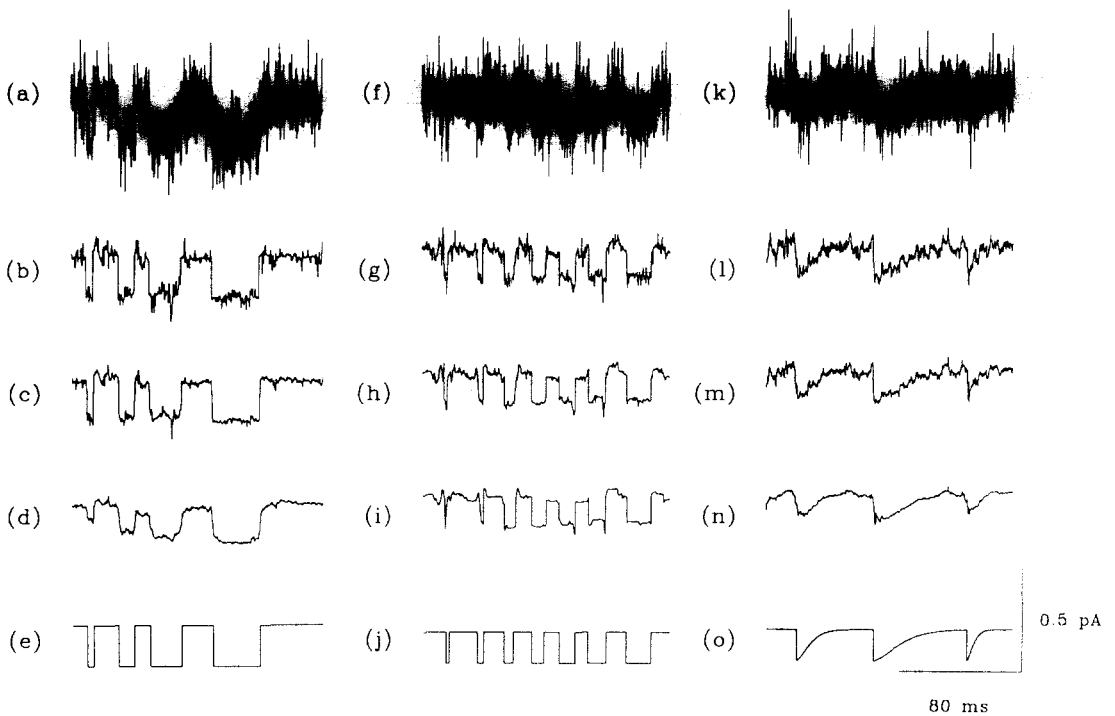

Fig. 2. Adjustment of non-linear filter parameters. Three different ways of using the non-linear filter are illustrated. To the noise obtained from a patch-clamp amplifier (filtered at 2 kHz, sampled at 5 kHz), the signal sequences shown in the bottom row (e, j and o) are added. The noise containing the signals (top row: a, f and k) are processed with the non-linear filter adjusting different filter parameters. The segments exhibited in the left-hand column (b, c and d) were processing by adjusting the relative weights assigned to each forward and backward predictor, namely, the term p in Eqns. 4 and 5. The numerical values of p used for b, c and d were 100, 5 and 1, respectively, while keeping the lengths of 5 pairs of predictors constant (± 4, ± 8, ± 16, ± 32 and ± 64). In the middle column, the filtered output (g) was processed again (h) and this procedure was repeated once again (i). Repeated iteration of noisy data suppressed the noise progressively but also distorted the signal characteristics. Three pairs of predictors (± 4, ± 8 and ± 16) with the relative weight of $p = 100$ were used. In the right-hand column, the progressive reduction in the noise was achieved by adding longer lengths of forward and backward predictors, while keeping the relative weight at 10. The predictors used were ± 4 and ± 8 for (l), ± 4, ± 8, ± 16 for (m) and ± 4, ± 8, ± 16, ± 32 for (n).

reject noise but introduce minimal distortion to the signal:

(i) The relative weights assigned to each forward and backward predictor can be adjusted by varying the negative exponent p in Eqn. 4 and Eqn. 5. When a large value of p is selected, the relative weight given to each predictor is accentuated. With the exponent of 100, for example, one forward or backward predictor of a given length will predominate at a given instant, and the weight given to the others will tend to be zero or near zero. Conversely, with a small p the difference in the weight assigned to each predictor will be less pronounced. Fig. 2 (a–e) illustrates the effects of varying the weighting factor on performance. As the weighting factor was decreased, the reduction in the noise became more drastic (Figs. 2b, c and d), but the amplitude and the edges of the signals (Fig. 2e) became distorted with the lowest weighting factor used (Fig. 2d).

(ii) The processing can be applied repeatedly, i.e., in a multi-pass fashion, to achieve a desired reduction in the background noise. Using 3 pairs of forward and backward predictors, a noise-contaminated record (Fig. 2f) was processed once (Fig. 2g), twice (Fig. 2h) and three times (Fig. 2i). The gain in the signal-to-noise ratio achieved by the repeated iteration was accompanied by the distortion of the original signal (Fig. 2j).

(iii) The number of predictors and their lengths can be varied. Three short pairs of the predictors used to extract exponentially decaying signals imbedded in the noise (Fig. 2k) preserved the original features of the signal, but the background noise was not effectively suppressed (Fig. 2l). As the predictors of longer lengths were added (Figs. 2m and 2n), a further reduction in the noise was achieved at the expense of distorting the original signal.

Although there are a large number of possible predictor combinations that can be used for processing a segment of data, the choice of the bank of predictors in practice is straightforward. Various choices of lengths of predictors should be based on the expected durations of signal features. If, for example, we anticipate signals of width 10 samples to be present in the data, then there should be at least one predictor whose length is less than 10 points. Naturally, if longer signal features are present, then correspondingly longer window predictors should be included in the bank. The weights automatically adjust when one or another of the predictors in the bank should make a greater or lesser contribution to the final estimate. Thus, for a faithful preservation of fast signal transient details, such as the abrupt changes observed during a single channel conductance, a large value of the weighting factor p in Eqn. 4 and Eqn. 5 is effective, whereas slow signal variations, e.g., during exponential decay back to the baseline, can best be extracted with a smaller p value.

Similarly, if the record contains long openings with many flickers, the bank of predictors should contain long as well as short ones, and a large value of p should be used to accentuate the jump. Reduction in the length of analysis window, which has been kept to 20 digital points throughout (except in Fig. 4i), has the further effect in faithfully preserving short pulses. For a slowly changing signal, such as the falling phase of excitatory post-synaptic currents, a low p (see Fig. 4n) ensures that the time constant of the decay is minimally distorted.

4. Comparison with low-pass filter

Noise rejection can also be achieved by linear low-pass filtering but it is usually accompanied by unacceptable signal distortion. Using binary rectangular pulse sequences, we have made extensive comparisons between our non-linear filter and a conventional linear moving averager.

The results of one such comparison are illustrated in Fig. 3. To the broad-band amplifier noise (filtered at 20 kHz and sampled at 40 kHz), two brief pulses of 0.63 and 0.88 ms duration and 1 pA in amplitude were added (Figs. 3a, b). Using 3 pairs of predictors, the segment containing the noise and signals was processed repeatedly. The traces illustrated in Figs. 3c, e, g and i were obtained, respectively, after 1, 2, 4 and 5 iterations. The same set of data was digitally filtered using a discrete low-pass, second-order recursive Butterworth filter (obtained by a conventional bilinear z-transformation). The cutoff frequency of the low-pass filter was adjusted so as to achieve

approximately the same reduction in the rms noise as with the output of the successive iterations using our technique. The cutoff frequencies used for the traces shown in Figs. 3d, f, h and j are, respectively, 4 kHz, 2 kHz, 1 kHz and 500 Hz. As expected, the amplitudes and the shape of the imbedded signal became progressively distorted as the cutoff frequency of the low-pass filter was lowered. The output of heavily filtered traces contained mainly filter artefacts, their shapes bearing little, if any, resemblance to the original signals. In contrast, the signals estimated with the non-linear filter faithfully retained the characteristics of the original signals.

The results of simulations shown in Fig. 4 further demonstrate that certain features of the signal which are severely distorted by low-pass filtering are preserved when the data are processed with the non-linear filter. The original records, shown in the first row, were assumed to have been filtered at 20 kHz and correctly sampled at the Nyquist frequency. The data were digitally filtered, again with a second-order recursive Butterworth filter, at 2 kHz (second row) and 1 kHz (third row). The records processed with the non-linear filter, displayed in the fourth row, may be compared with the original signal sequence, shown in the last row, as well as those processed with a low-pass filter. The reduction in the rms noise achieved by the non-linear filter was equivalent to that with a low-pass filter with the cutoff frequency of 1 kHz. The opening of a ficticious channel (Fig. 4e) was instantaneous but the transit to the closed state was achieved in 2 steps, pausing briefly at an intermediate conductance state. From Figs. 4b and c, it would have been difficult to deduce that there was a brief sojourn to an intermediate state. In the middle column,

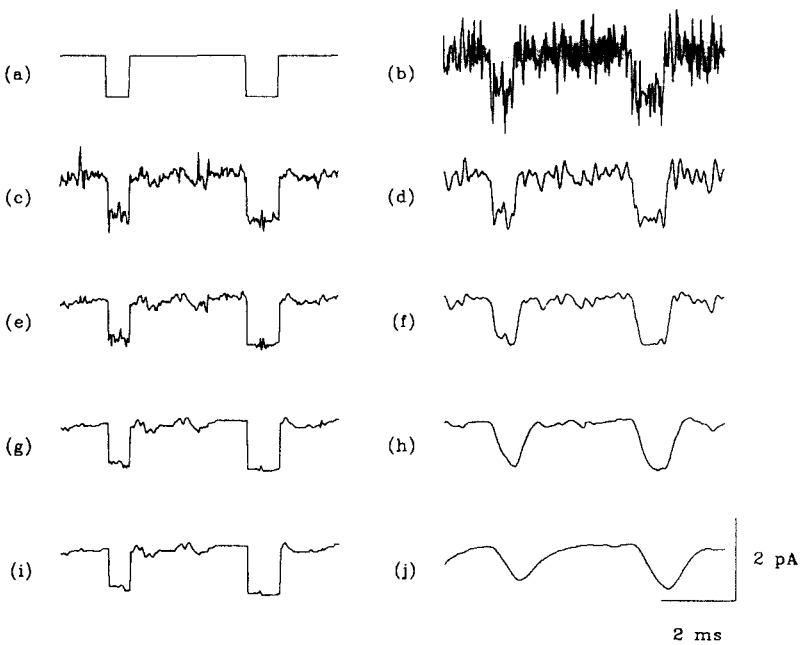

Fig. 3. Comparison between the non-linear filter and low-pass filter. A series of 1 pA pulses, with the durations of 0.63 ms and 0.88 ms, shown in (a), was added to a broad-band noise obtained from a patch-clamp amplifier (filtered at 20 kHz and sampled at 40 kHz). The signal imbedded in the noise (b) was then processed with the non-linear filter, using 3 pairs of predictors (± 2, ± 4, ± 8) and the weighting factor of 100. Sample segments of the processed data after the first (c), second (e), fourth (g) and fifth (i) iterations are shown in the left-hand column. For comparison, the same segments of the data were filtered digitally with a low-pass filter, using a bilinear z-transformation method based on a recursive second-order Butterworth filter model (cutoff rate = -12 dB). The filtered records with the cutoff frequencies of 4 kHz (d), 2 kHz (f), 1 kHz (h) and 500 Hz (j) are shown in the right-hand column.

an opening of a channel was separated by three brief interruptions (Fig. 4j). Inverse filtering of Fig. 4h would not have given an unambiguous indication that two of the closings represented sojourns in a conductance sublevel rather than a partially resolved closing. A ficticious postsynaptic current shown in Fig. 4o rose instantaneously and decayed exponentially thereafter. The estimated rise-time constant from the linearly filtered records (Figs. 4l and m) would have been erroneous, unless the sampling frequency was kept strictly at the Nyquist frequency. If, on the other hand, the sampling frequency was so lowered, the determination of the decay constant would have been subject to a considerable error. In contrast, some of the salient features of the signals are retained in the records processed with the non-linear filter.

5. Amplitude histograms

A first-order, three-state, Markovian signal sequence of length 50000 points was generated from a transition matrix with elements $a_{ii} = 0.97$ and $a_{ij} = 0.015$ for $i \neq j$, and added to the noise. The mean duration of the signal generated by the matrix is 6.6 ms. The traces containing the signal sequence in the computer-generated noise were processed and then the amplitude probability density curves constructed. From the number of these simulations, using various signal amplitudes, we have ascertained that levels of signals buried in noise can accurately be identified provided their amplitude separations are greater than 0.7 of a standard deviation of the noise.

Fig. 5 shows short sample segments of data containing a 3-level (one level being the baseline) Markov signal. The amplitudes of the imbedded

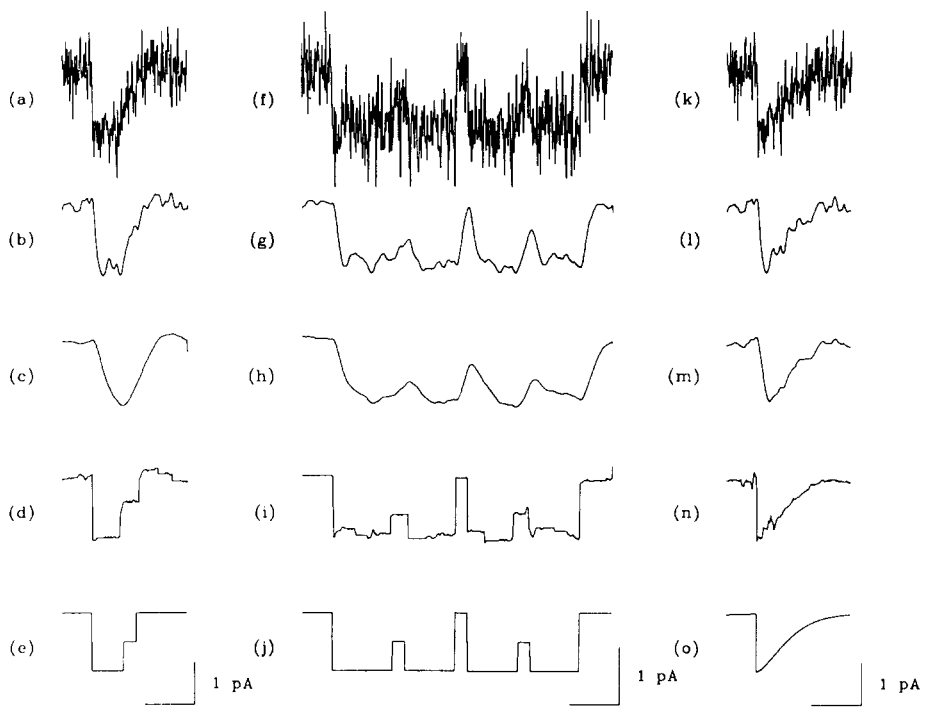

Fig. 4. Further comparison between the non-linear filter and low-pass filter. Each noisy record shown in the first row (a, f, k) was obtained by adding a signal sequence shown in the last low (e, j, o) to a computer-generated noise. The standard deviation of the noise was 0.4 pA, corresponding to a patch-clamp amplifier noise filtered 20 kHz. The records filtered with a low-pass Butterworth filter with the cutoff frequency of 2 kHz and 1 kHz are shown in the second (b, g, l) and third row (c, h, m). In the fourth row (d, i, n), the same records processed with the non-linear filter are displayed. For (d), the records were processed three times using three pairs of predictors (± 4, ± 8, ± 16), with $M = 15$ and $p = 100$. The record (i) was similarly processed, except that the value of p was lowered to 5. For (n), the predictors used were ± 4, ± 8, ± 16 and ± 32, with $M = 20$ and $p = 10$. The time calibration bar represents 2 ms for the sampling frequency of 40 kHz or 0.8 ms for 100 kHz.

signals (in addition to the baseline) in Figs. 5a, b, c and d are, respectively, 0.115 pA and 0.23 pA, 0.10 pA and 0.20 pA, 0.085 and 0.17 pA, and 0.07 and 0.14 pA, whereas the standard deviation of the noise, before the signals were added, was 0.1 pA. The amplitude probability density distributions of the original data points appear as skewed Gaussians, shown as broken lines in Fig. 5. The amplitude histograms constructed from the non-linear filter, in contrast, reveal 3 distinct peaks, corresponding to the baseline and 2 signal levels (solid lines in Fig. 5). Each amplitude histogram

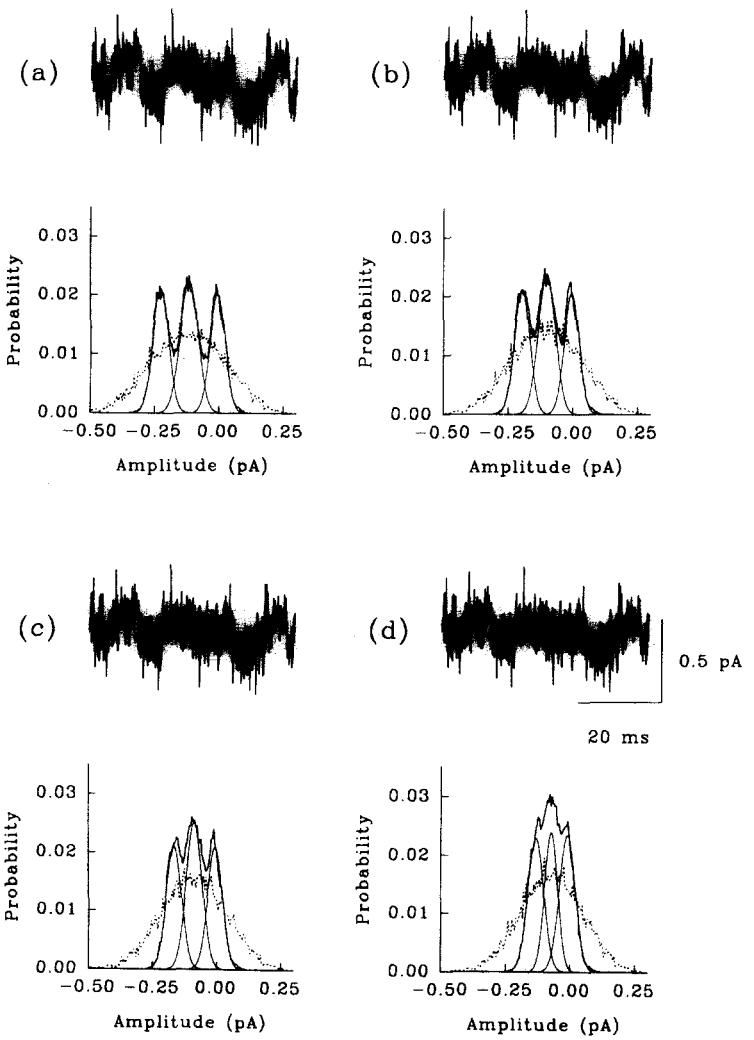

Fig. 5. Determination of signal amplitudes. A 3-state, first-order, Markovian signal sequence, 50000 points long, was generated according to the transition probability matrix of $a_{ii} = 0.97$, and $a_{ij} = 0.015$ for $i \neq j$. The signal sequence was then added to the computer-generated random noise with the standard deviation of 0.1 pA. The amplitudes of the imbedded signals were 0, −0.115, −0.23 pA (a), 0, −0.1, −0.2 pA (b), 0, −0.085, −0.17 pA (c), and 0, −0.07, −0.14 pA (d). The noisy signal sequences were then processed with the non-linear filtering technique, using 4 pairs of predictors (±4, ±8, ±16 and ±32) and the weighting factor of 10. The amplitude probability density curves of the original data (dotted traces) appeared as skewed Gaussian with no obvious peaks. The curves obtained from the processed data, in contrast, showed distinct peaks, corresponding to the amplitudes of the underlying signals (solid lines). Each curve obtained from the processed data was then deconvolved using a maximum likelihood criterion. The errors in the amplitude estimation increased systematically as the amplitudes of the original signals decreased, ranging from 1% in (a) to 5% in (d).

obtained from the filtered data was deconvolved to the 3-Gaussian distributions (shown as continuous curves in Fig. 5) using a maximum likelihood criterion. From the deconvolved curves, the amplitudes of the imbedded signals could be deduced. Also, it could be correctly inferred from the curves that the probability of being in any one of the 3 states was about equal. As the separation between the signal amplitude was further reduced, the peaks in the amplitude histogram became progressively indistinct, and when the signal amplitudes were less than 0.5 standard deviation, the peaks in the histogram could not longer be resolved. From the results of these simulations, we concluded that the limit of resolution in detecting signal levels unambiguously with our technique is about 0.5 of a standard deviation of the noise.

6. Preservation of brief channel events

The duration of channel currents fluctuating in a random manner between the open and closed state can be extremely brief, but it is a general practice that the records containing these channel events are preprocessed before sampling by an analogue Bessel filter with a cutoff frequency of 1 kHz to 3 kHz to reduce the background noise. For channel events that are a fraction of 1 ms such an analogue low-pass filtering will severely distort and potentially remove brief signal states. We claim our discrete non-linear filter can effectively suppress the noise digitally thereby obviating the need for any distorting low-pass analogue preprocessing. By performing all our processing digitally on raw data samples we are able to reveal brief channel events.

Amplifier noise, filtered at 20 kHz and sampled at 40 kHz, was added to a digitally generated, first-order Markovian signal sequence. The mean and the shortest duration of the signals of 0.5 pA in amplitude were, respectively, 1.25 and 0.125 ms. The record containing the Markovian signal sequence in a broadband noise was processed using 3 pairs of predictors. Fig. 6 shows a sample trace of the signal sequence (a) and the noise-contaminated data (b). The filtered output, a segment of which is displayed in Fig. 6c, was quantized into one of two states by defining the

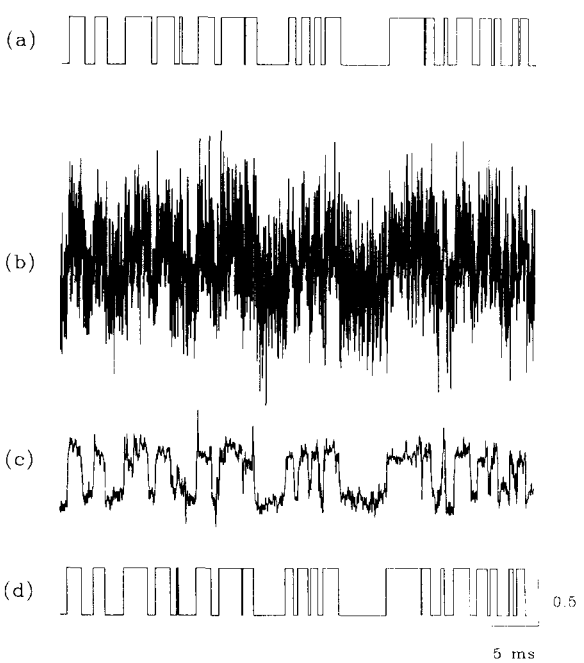

Fig. 6. Extraction of brief pulses buried in the background noise. A binary state, first-order, Markovian signal sequence of 50000 points was generated according to the transition probability matrix, $a_{11} = a_{22} = 0.98$ and $a_{12} = a_{21} = 0.02$, and then added to the same length of the broadband noise obtained from a patch-clamp amplifier (filtered at 20 kHz, sampled at 40 kHz). The amplitude and the expected mean duration of the signal was 0.5 pA and 1.25 ms, respectively (a). The noisy trace containing the signal sequence (b) was then processed once with the non-linear filter using 3 pairs of short predictors (± 4, ± 8, ± 16) and the weighting factor of 10. The output of the non-linear filtering (c) was quantized into a binary signal sequence (d).

appropriate midpoint threshold (the so-called nearest neighbour rule) in Fig. 6d. Brief sojourns to the open state from the closed state or vice versa, normally obliterated by analogue low-pass filtering before sampling, are clearly retained in the data processed with our technique.

To make a quantitative assessment in the efficacy of our processing technique, we have constructed open-time and closed-time histograms of the records containing short pulses of various amplitudes and compared them with the true histograms. The processing technique used for these simulations was identical to those described for Fig. 6. The results of one such simulation are shown in Fig. 7, in which a 2-state signal se-

quence (0.3 pA) was added to the broad-band noise. The histograms obtained from 50 000 points of the processed record are close to those obtained from the original signal sequence. Both sets of the histograms could be fitted with single exponential curves. The estimated mean open-time and closed-time were, respectively, 1.70 and 1.89 ms, compared to the true values of 1.44 and 1.40 ms. The estimated durations of the mean

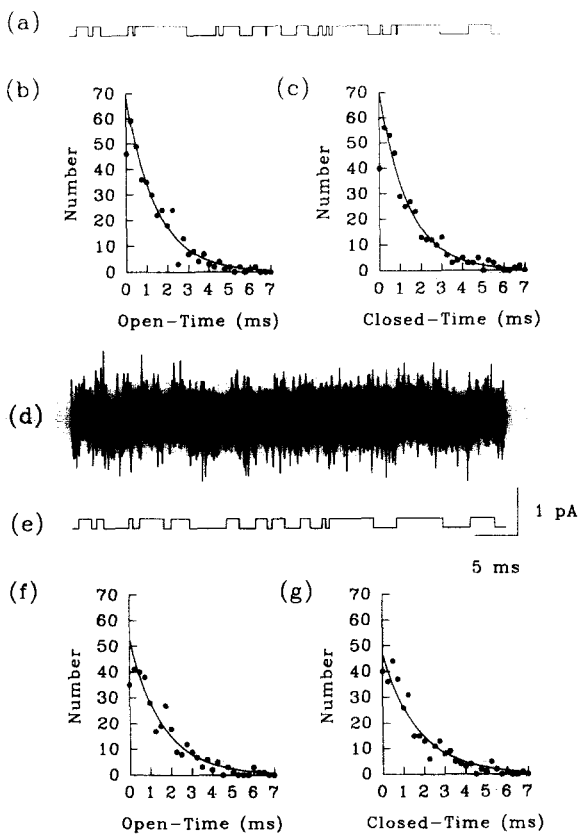

Fig. 7. Estimates of channel kinetics. Using the same procedures as in Fig. 6, the open-time and closed-time histograms of the original signal (b and c) and extracted signal (f and g) were constructed, and fitted with single exponential curves (solid lines). Short segments of the signal sequence, signal imbedded in the noise and signal and estimated signal sequence are displayed, respectively, in (a), (d) and (e). The amplitude of the binary signal used was 0.3 pA. The $1/e$ values of the open-time and closed-time histograms were 1.44 and 1.40 ms for the original signal sequence (b and c), whereas those for the estimated signal sequence were 1.70 and 1.89 ms. When the amplitude of the signal was relatively small, the non-linear filter frequently failed to detect brief sojourns from the closed to the open states and vice versa.

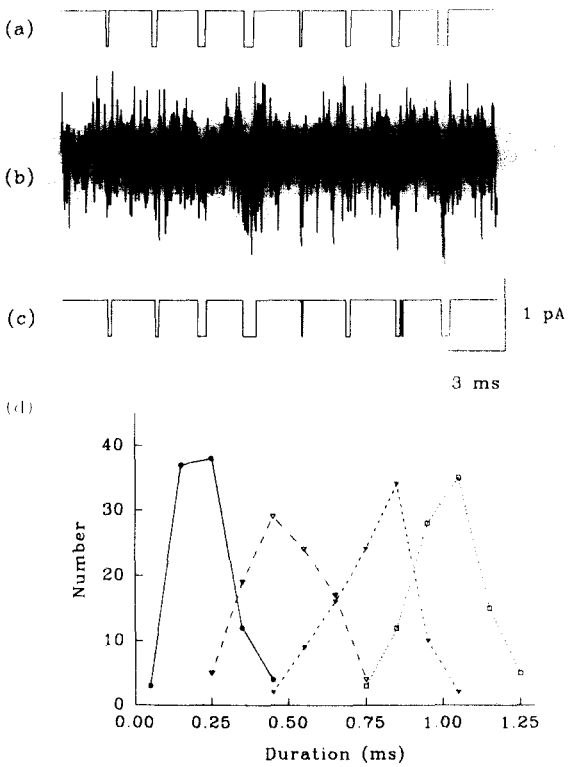

Fig. 8. Errors in estimating durations of brief pulses. A series of pulses lasting 0.25, 0.50, 0.75 and 1.0 ms, shown in (a), was added to the broadband noise, and from the noise-contaminated trace (b) the signal sequence was recovered (c), using the same procedures as in Fig. 6. The amplitude of the signal was 0.5 pA. The durations of the estimated signals were tabulated and plotted in the form of histograms (d).

open- and closed-time tended to be larger than the true values, mainly because brief events lasting less than 0.1 ms were frequently missed, especially when the signal-to-noise ratio was low. The magnitude of the errors decreased systematically as the amplitude of the signal was increased.

The durations of brief signals buried in the noise also could be estimated with an acceptable degree of accuracy. A series of pulses, lasting 0.25, 0.5, 0.75 and 1.0 ms, was first added to the noise and then processed using our non-linear filter. After quantizing to 2 levels, as before, we tabulated the mean duration of the estimated pulses. Fig. 8 displays: (a) short segments of the original signal sequence, (b) the unprocessed data containing the signal sequence, and (c) the estimated signal sequence. The distribution of esti-

mated intervals are tabulated in Fig. 8d. The estimated signal durations, when the signal amplitude was 0.5 pA, were 0.16 ± 0.03 ms, 0.44 ± 0.07 ms, 0.75 ± 0.06 ms and 0.99 ± 0.06 ms. The corresponding estimates, when the amplitude of the signal was 1.0 pA, were 0.21 ± 0.09 ms, 0.48 ± 0.12 ms, 0.76 ± 0.12 ms and 1.00 ± 0.14 ms.

Discussion

We have described and tested a novel non-linear digital signal processing method for extracting accurate information about channel currents or postsynaptic potentials (or currents) from noisy measurements. The method combines the outputs of a forward and backward bank of predictors by weighting them in a data-dependent, time-varying manner. Unlike other commonly used signal processing techniques, such as Kalman filtering (Anderson and Moore, 1979), Hidden Markov Models technique (Chung et al., 1990) and maximum likelihood sequence estimation (Forney, 1973), the method we present here is computationally inexpensive and simple to implement, requiring the same order of magnitude of computations as running a small bank of simple moving average filters in parallel. For example, with a large weighting factor, the number of computational steps involved in processing a segment of data of the length T with N predictors is about 3 NT. Thus, even without optimizing the code, the processing of a 50 000-point record with 10 predictors on a 7 MIP Sun IV workstation could perform the calculation in less than a minute. The computational steps involved in our non-linear filter can readily be programmed on a modern microprocessor, thereby enabling the data to be processed on-line.

We have demonstrated the power of the non-linear technique for extracting known signals buried in noise that was generated by a model "patch" attached to the input of the patch-clamp amplifier. After processing the noise variance is reduced by an order of magnitude (Fig. 1), while the signal remains relatively undistorted (Fig. 2). The superiority of this method over a low-pass filtering becomes apparent when the the 2 methods are compared (Figs. 3 and 4). Using our method, we have further demonstrated that the amplitudes of small signals (Fig. 6) and open-time and closed-time histograms of a Markovian signal sequence (Fig. 7) can be accurately estimated even when the signal is dominated by the background noise. Similarly, the amplitudes and the decay constants of small exponentially decaying signals contaminated by the noise, mimicking post-synaptic currents recorded from neurones, could reliably be estimated when processed with the non-linear filter.

One of the fruitful ways we envisage of using the novel non-linear filter is for examining single-channel patch-clamp recordings. When the output of the amplifier containing channel currents are processed with a low-pass filter brief sojourns from the closed to open states or from the open to closed states will not be detected, thereby distorting the true distribution of the dwell times in the open and closed states. There does exist, however, a theoretical framework to model brief events which are missed due to the resolution limitations of the measurement process. In this way one can model what is actually observed, by assuming that there exists a threshold length of time under which short events are overlooked. Previously, only approximations to this mechanism were available (Colquhoun and Sigworth, 1983) but recently an exact formulation, or closed form solution, has been presented by Hawkes et al., (1990), using the theory of semi-Markov processes. This theory remains appropriate to model the resolution limitations of any set of measurements (see, also Fredkin and Rice, 1987; Ball and Sampson, 1988). The non-linear filter that we have introduced here is an alternative method of preserving fast transients of the signals or measuring brief channel events directly, even when their amplitudes are small. The practical limitation for detecting short interval pulses rests largely on the frequency response of the amplifier (see Sigworth, 1983).

As the molecular events underlying the opening and closing of a receptor-channel complex are being explored in increasing detail, it is becoming necessary to characterize single-channel currents more precisely than before. The use of our

method, in conjunction with Hidden Markov Model techniques. (Chung et al., 1990), should enable us to characterize the single-channel kinetics more precisely and provide information about details of channel currents which are buried in noise and have hitherto been largely inaccessible.

Appendix

Derivation of smoothing equations

1. Bayesian framework. We develop an optimal predictor structure in a Bayesian framework. The following development will be directed towards finding the optimal signal estimate at time k. The resulting formulae can be then applied for every time instant to obtain a sequence of signal estimate outputs.

In our notation we let $p(\cdot)$ denote posterior probabilites (or densities where appropriate), and $\pi(\cdot)$ denote prior probabilities (or densities). Also let $E\{\cdot\}$ denote expectations and let $n \sim N$ mean the random variable n is distributed according to distribution N.

We implicitly define our signal model through a set of K forward and K backward mutually exclusive hypotheses indexed by $i \in \{1, 2, \ldots, K\}$ along with superscripts f or b (denoting the direction) as appropriate, i.e.,

$$\mathscr{H}_i^f(k): \begin{cases} a\colon y(j) = \hat{x}_i^f(j) + n(j), \\ \quad \forall j \in \{k-M+1, \ldots, k-1, k\} \\ b\colon n(j) \sim N(0, \rho),\ \rho \sim \pi(\rho) \\ c\colon \{y(i), i > k\} \text{ independent of} \\ \quad \{y(i), i \le k\} \end{cases} \tag{A.1}$$

and

$$\mathscr{H}_i^b(k): \begin{cases} a\colon y(j) = \hat{x}_i^b(j) + n(j), \\ \quad \forall j \in \{k, k+1, \ldots, k+M-1\} \\ b\colon n(j) \sim N(0, \rho),\ \rho \sim \pi(\rho) \\ c\colon \{y(i), i < k\} \text{ independent of} \\ \quad \{y(i), i \ge k\} \end{cases} \tag{A.2}$$

where condition b means the noise is an independent and identically distributed zero mean Gaussian random variable but has unknown variance ρ distributed according to some prior probability $\pi(\rho)$. In (A.1), $\hat{x}_i^f(\cdot)$ is the ith forward one step ahead predictor and is assumed measurable or known in the sense

$$\hat{x}_i^f(k) = g_i(y(k-1), y(k-2), \ldots, y(k-L_i)),$$
$$i \in \{1, 2, \ldots, K\}$$

where $g_i(\cdot, \ldots, \cdot)$ is the function describing the ith unbiased estimator, and L_i is the window length. Similarly, $x_i^b(\cdot)$ is the ith mirror image anti-causal predictor whose outputs are measurable in the following sense

$$\hat{x}_i^b(k) = g_i(y(k+1), y(k+2), \ldots, y(k+L_i)),$$
$$i \in \{1, 2, \ldots, K\}.$$

Each of these hypotheses has a prior probability

$$\pi_i \triangleq \pi\big(\mathscr{H}_i^f(k)\big) = \pi\big(\mathscr{H}_i^b(k)\big) \tag{A.3}$$

which is assumed independent of time k and satisfies $2\sum_{i=1}^{K}\pi_i = 1$. Throughout this paper, unless otherwise stated, we make the selection $\pi_i \propto L_i$.

We now derive the optimal fixed interval smoother under this general signal model hypotheses. Let the fixed interval of data be centered at time k and of width $2M-1$ (the analysis window), i.e.,

$$\mathscr{Y}_M(k) \triangleq \{y(k-M+1), \ldots, y(k), \ldots, y(k+M-1)\}.$$

Also define

$$\mathscr{Y}_M^f(k) \triangleq \{y(k-M+1), \ldots, y(k)\}$$

and

$$\mathscr{Y}_M^b(k) \triangleq \{y(k), \ldots, y(k+M-1)\}$$

which will be used later. Note these two latter data sets appear in the hypotheses (A.1) and (A.2). We regard the desired signal $x(k)$ and its noisy version $y(k)$ as being phenomena either consistent with $\mathscr{Y}_M^f(k)$ or consistent with $\mathscr{Y}_M^b(k)$. Note that with at most one jump signal change

occurring somewhere in the analysis window $\mathscr{Y}_M(k)$ then at least one of the consistency conditions will be true.

Then the condition mean estimate of the signal $x(k)$ based on the window of data $\mathscr{Y}_M(k)$ implements the optimal Bayesian estimate for a wide range of criteria and by Bayes' rule is given by,

$$\hat{x}(k) \triangleq E\{x(k) \mid \mathscr{Y}_M(k)\} = \sum_{i=1}^{K} \left[f_i(k) E\{x(k) \mid \mathscr{Y}_M(k), \mathscr{H}_i^f(k)\} + b_i(k) E\{x(k) \mid \mathscr{Y}_M(k), \mathscr{H}_i^b(k)\} \right]$$

where

$$f_i(k) \triangleq p(\mathscr{H}_i^f(k) \mid \mathscr{Y}_M(k)) \text{ and}$$

$$b_i(k) \triangleq p(\mathscr{H}_i^b(k) \mid \mathscr{Y}_M(k)) \tag{A.4}$$

are maximum a posteriori weights. A consequence of the conditions in hypothesis $\mathscr{H}_i^f(k)$ (A.1) is the following identity

$$E\{x(k) \mid \mathscr{Y}_M(k), \mathscr{H}_i^f(k)\} = \hat{x}_i^f(k).$$

Similiarly by hypothesis $\mathscr{H}_i^b(k)$ in (A.2)

$$E\{x(k) \mid \mathscr{Y}_M(k), \mathscr{H}_i^b(k)\} = \hat{x}_i^b(k).$$

Now we compute the maximum a posteriori weights (A.4) for our fixed interval smoother (Niedźwiecki and Kennedy, 1990). From (A.3) and (A.4) we can write using Bayes' rule

$$f_i(k) \propto p(\mathscr{Y}_M(k) \mid \mathscr{H}_i^f(k)) \pi(\mathscr{H}_i^f(k)) = \pi_i p(\mathscr{Y}_M(k) \mid \mathscr{H}_i^f(k)) \tag{A.5}$$

and an analogous equation holds for $b_i(k)$. (In the following we will just concentrate on the forward equations noting the backward equations are entirely analogous.) Continuing we get

$$p(\mathscr{Y}_M(k) \mid \mathscr{H}_i^f(k)) \propto p(\mathscr{Y}_M^f(k) \mid \mathscr{H}_i^f(k)). \tag{A.6}$$

which follows from the independence assumption: condition c in (A.1).

2. Explicit expression for the weights. We need to give an explicit form for the maximum a posteriori weights in terms of measurable quantities. The first barrier is the removal of the uncertainty implicit in our additive zero mean Gaussian noise model which has randomly distributed noise variance ρ. Write

$$p(\mathscr{Y}_M^f(k) \mid \mathscr{H}_i^f(k)) = \int_0^\infty p(\mathscr{Y}_M^f(k) \mid \mathscr{H}_i^f(k), \rho) \pi(\rho \mid \mathscr{H}_i^f(k)) d\rho. \tag{A.7}$$

Now we evaluate (A.7). Firstly, we assume a family of improper prior distributions (parametrized by the weighting factor p) for the noise variance ρ which we take as being independent of the particular hypothesis, i.e.,

$$\pi(\rho \mid \mathscr{H}_i^f(k)) \triangleq \pi_p(\rho) \propto \rho^{M/2-p-1}. \tag{A.8}$$

The remaining term in (A.7) is given by the standard chain rule

$$p(\mathscr{Y}_M^f(k) \mid \mathscr{H}_i^f(k), \rho) = \prod_{j=0}^{M-1} p(n(k-j) = y(k-j) - \hat{x}_i^f(k-j)) = \frac{1}{(2\pi\rho)^{M/2}} \times \exp\left\{ -\frac{1}{2\rho} \sum_{j=0}^{M-1} \left[y(k-j) - \hat{x}_i^f(k-j) \right]^2 \right\} \tag{A.9}$$

where $y(k-j) - \hat{x}^f(k-j)$ is the one step ahead prediction error. Note (A.9) follows from the Gaussian noise assumption, although any distribution for the noise can be assumed here. Also crucial here is that $x_i^f(\cdot)$ is a predictor in the sense that it is not a function of $\{y(i), i \geq k\}$.

Combining (A.5)–(A.9) and noting

$$\int_0^\infty x^{p-1} e^{-ax} dx = \Gamma(p) a^{-p}$$

where $\Gamma(\cdot)$ is the Gamma function (and letting $x = \rho^{-1}$), gives the desired Eqn. 4; showing that the maximum a posteriori weights take a simple form. Similarly the backward weights are given by Eqn. 5.

Acknowledgements

Throughout the course of this study, Mrs. Jennifer Edwards provided excellent technical assistance, for which we are grateful. We thank Mr. B. Keys for writing several computer programs used for this study. This work was supported in part by grants from the Australian Telecommunications and Electronics Research Board, Australian Research Council, ANU Centre for Information Science Research, and the National Health and Medical Research Council of Australia.

References

Anderson, B.D.O. and Moore, J.B. (1979) Optimal Filtering, Prentice, Eaglewood Cliffs, NJ.

Ball, F. and Sampson, M. (1988) Aggregated Markov processes incorporating time interval omission. Adv. Appl. Prob., 20: 546–572.

Childers, D.G. (1987) Modern Spectrum Analysis. IEEE Press, New York.

Chung, S.H., Moore, J.B., Xia, L., Premkumar, L.S. and Gage, P.W. (1990) Characterization of single channel currents using digital signal processing techniques based on Hidden Markov Models. Phil. Trans. R. Soc. Lond. B, 329: 265–285.

Colquhoun, D. and Sigworth F.J. (1983) Fitting and statistical analysis of single channel records. In: B. Sakmann and E. Neher (Eds.), Single Channel Recording, Plenum, New York, pp. 191–263.

Crouzy, S.C. and Sigworth F.J. (1990) A practical approximate solution to the dwell time omission problem. Biophys. J., 58: 731–743.

Forney, G.D. (1973) The Viterbi algorithm. Proc. Inst. Elec. Electron. Eng., 61: 268–278.

Fredkin, D.R. and Rice, J.A. (1987) Correlation functions of a function of a finite-state Markov process with application to channel kinetics. Math. Biosci., 87: 161–172.

Hawkes, A.G., Jalali, A. and Colquhoun, D. (1990) The distributions of the apparent open times and shut times in a single channel record when brief events cannot be detected. Phil. Trans. R. Soc. Lond. A, 332: 511–538.

Niedźwiecki, M. and Kennedy, R.A. (1990) Non-linear non-causal noise rejection schemes based on competitive smoothing. In: L. Torres, E. Masgrau and M.A. Lagunas (Eds.), Signal Processing. V. Theories and Applications. Elsevier, New York, pp. 133–136.

Rabiner, L.R. (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc. Inst. Elec. Electron. Eng., 77: 257–285.

Sigworth, F.J. (1983) Electronic design of the patch clamp. In: B. Sakmann and E. Neher (Eds.), Single Channel Recording, Plenum, New York, pp. 3–35.