Spike Detection Using Continuous Wavelet Transform

Zoran Nenadic, Member, IEEE, and Joel W. Burdick, Member, IEEE,
California Institute of Technology, Pasadena CA

Abstract

This paper combines wavelet transforms with basic decision theory to develop a new unsupervised method for robustly detecting and localizing spikes in noisy neural recordings. The method does not require the construction of templates, or the supervised setting of thresholds. We present extensive Monte Carlo simulations, based on actual extracellular recordings, to show that this technique surpasses other methods in a wide variety of recording conditions. Moreover, the simplicity of the method allows for real-time execution.
Spike Detection Using Continuous Wavelet Transform

I. INTRODUCTION

Electrical recordings of action potentials have become an indispensable tool in neuroscience. The shapes and amplitudes of these action potentials, or spikes, are highly stereotyped [1]. Since such regularity obviously carries very little information, it has long been argued that the information capacity of a spike train is largely dominated by the temporal variability of individual spikes within the train [2]. It is further debated whether the nervous system cares about the precise timings of individual spikes (time coding) or the total number of action potentials fired in a certain time window (rate coding) [3]. Either way, for experimental investigations it is ultimately important to accurately and robustly detect and localize the occurrence of individual spikes within the extracellular recording signal. Essentially all studies of the information content of experimental recordings start from this processed data. Errors in detecting the number and location of spikes will necessarily propagate through all subsequent analyses.

Extracellularly recorded spike trains are inevitably corrupted by noise. The noise sources are quite varied: the recording hardware, the ambient recording environment, the superimposed activity of multiple neurons, and the spatially averaged activity of distant cells also known as the local field potential. Perhaps most importantly, the activity of distant neurons may appear as noise which is highly correlated with the useful signal. Another difficulty is that unlike their intracellular counterparts, extracellular potentials do not have a universally stereotyped shapes and amplitudes. Depending on various parameters such as cell geometry, the distribution and density of individual ionic channels, and the position of a recording electrode with respect to electrically active membranes, extracellular potentials can have different shapes, phases and amplitudes [4]. All of these issues make the problem of spike detection challenging.

Because of its practical importance to experimental neuroscience, the detection of spikes in noisy extracellular observations is a classical problem. There exist dozens of algorithms that accomplish the task of spike detection and they can be classified as manual and automated, supervised and unsupervised. As discussed below, we are particularly interested in spike detection methods that are automated and unsupervised.
The most common manual detection tool is the window discriminator [5]. Extracellular signals that exceed a simple amplitude threshold and pass through a subsequent pair of user-specified time-voltage boxes are identified as spikes. Although practically successful, this method requires human supervision and its manual nature makes it especially time consuming when using multiple electrodes. Also, the statistical nature of this method is not well understood. Another widely used technique for spike detection is amplitude thresholding, where the threshold value can be set automatically (e.g., as a multiple of the estimated noise standard deviation), or manually. While this detection method is simple, its performance deteriorates rapidly under low signal-to-noise ratio (SNR) conditions. Other detection methods include power detection [6], [7], matched filtering [7], principal components [8], Haar transformation [9] and the discrete wavelet packet transform [10]. In the power detection method, also known as energy detection [11], the instantaneous power of the signal, calculated using a sliding window approach, is compared against a threshold derived from the mean and the standard deviation of the noise power. While this method usually offers some improvement over the amplitude threshold method, it performs almost as poorly in a low SNR environment. A matched filter (generalized matched filter) approach is known to maximize SNR when a signal is embedded in a white (colored) noise, respectively. However, because it is based on template matching, this method cannot be applied in an unsupervised fashion—supervision is required to construct the templates. The method of principal components requires the construction of the spike autocorrelation matrix, where multiple spikes from different cells are collected before the actual detection procedure. Generally, methods based on template matching or prior knowledge of "typical" spike shape are supervised detection algorithms, and not the subject of this paper. Yang and Shamma [9] proposed a spike detection method using the discrete Haar transformation, which is essentially a wavelet idea. However, our wavelet method does not require their white noise assumption and unnecessary inverse transformation from wavelet domain to time domain. Owies [10] proposed a multi-resolution version of the generalized likelihood ratio test (GLRT) for spike detection, a method somewhat similar in spirit to the one developed here. The major differences between the two methods will be pointed out later. It suffices to say that the multi-resolution version of GLRT developed in [10] cannot be implemented in an unsupervised manner.

Our interest in unsupervised methods arises from our current efforts to develop control algorithms that autonomously reposition extracellular recording electrodes so as to optimize and then maintain

1The ultimate purpose of window discriminators is to classify spikes originating from different cells. In order to be classified, spikes need to be detected first.
high signal recording quality in the face of inherent electrode and cell migrations in neural tissue [12]. To assess recording quality in the feedback loop, action potentials must be isolated on an ongoing basis. Because moving electrodes will experience displacements of hundreds of microns, the shape, phase, and amplitude of the detected spikes will vary significantly over the electrodes’ movement ranges. Consequently, template-based methods are inappropriate for our use. Moreover, within the range of the electrode’s movement, the recorded signal may experience extreme SNR conditions, ranging from low-noise–high-signal amplitude near the axon hillock, to high-noise–low-signal amplitude when no neurons are present nearby. Such a wide “dynamic range” requires an unsupervised spike detection algorithm with robust performance over a wide range of parameters, which, to our knowledge, none of the existing unsupervised algorithms is capable of. The algorithm presented in this paper provides this capability in an unsupervised fashion. Moreover, our performance evaluation (Section III) shows that this approach performs better than traditional manual methods, and therefore can be profitably applied in situations where manual methods are currently used. The paper is organized as follows. In Section II we use the theory of continuous wavelet transform combined with basic detection and estimation theory to develop a new neural transient detection algorithm. Section III evaluates performance of the algorithm with respect to other commonly used spike detection methods. The results were cross-validated using data synthesized from actual recording experiments. Concluding remarks are given in Section IV. Some of the mathematical derivations are given in the Appendix.

II. SPIKE DETECTION VIA WAVELETS

The core of our detection algorithm is presented in this section. The methodology consists of a combination of several techniques stemming from multi-resolution wavelet decomposition, statistics, detection theory and estimation theory. For the readers’ convenience we state the five major steps of the algorithm up-front. Each step will be explained in the subsequent sections.

1) Perform multi-scale decomposition of the signal using an appropriate wavelet basis.
2) Separate the signal and noise at each scale.
3) Based on results from steps 2 and 3, perform Bayesian hypothesis testing at different scales to assess the presence/absence of spikes.
4) Combine the decisions at different scales into a single detection decision.
5) Estimate the arrival times of individual spikes.

The chances of detecting a signal embedded in noise are improved when one can take advantage of prior information about the signal and the noise. The prior information could be acquired through
biophysical considerations or experimental trials. To keep the algorithm general and unsupervised, these
prior information have to be as vague as possible. This will be discussed in Section II-A and Section II-
B. Two assumptions about the noise are used in the derivations throughout the article, the background
noise is: i) stationary ii) Gaussian. Although these assumptions are not crucial for implementation, they
insure the mathematical tractability of the derivations. When they are violated, the performance of the
algorithm may be different from the results reported in Section III. The stationarity of neural noise cannot
be assumed in general. For time scales considered in this article we assume that the statistics of the noise
does not change appreciably. If the stationarity ever becomes a concern, the data can be broken into
shorter segments that can be analyzed separately. Also, wavelet bases of compact support are well suited
for representation of non-stationary signals. In Section III-A we shall give a justification of the Gaussian
assumption.

The problem of detecting transients in a collection of noisy observations has been studied for decades.
The presence/absence of a useful signal in a background noise is normally cast as a hypothesis testing
problem, where no signal is present under the null hypothesis [11]. If the signal to be detected is not
perfectly known, which is the case for unsupervised problems, usually no uniformly most powerful test
exists [13]. In these cases, the performance of a detector depends on the signal representation. A signal’s
representation can be model based or expansion based. When no appropriate model for the signal can
be found, one normally projects the signal onto a canonical set of basis functions, which gives rise to
a set of expansion coefficients. Depending on the signal representation, the detection problem can be
formulated in a variety of domains: the time domain, frequency domain, time-frequency domain, etc.

In this paper we use a time-scale expansion based signal representation. As we will show, there exist
wavelet basis functions that are well suited to spike detection. Moreover, there exist fast algorithms for
wavelet filtering [14], thereby allowing real-time implementation. A brief introduction to the continuous
wavelet transform is presented next. For more details on this subject, interested readers can consult [15]
or other texts. For a look at other approaches to signal detection using wavelet bases, see [16]–[18].
These techniques are based on dyadic wavelets (see Section II-B), and are mainly concerned with the
detection of a single transient within the observation sequence, without the estimation of parameters such
as transient arrival times.

A wavelet $\psi$ is a function of finite energy, i.e. $\psi \in L^2(\mathbb{R})$, and zero average:

$$\int_{\mathbb{R}} \psi(t) \, dt = 0 .$$
It is normalized $\| \psi \|_2 = 1$ and centered in the neighborhood of the origin. From this function, also called the *mother wavelet*, one can obtain a family of time-scale waveforms by translation and scaling

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi \left( \frac{t-b}{a} \right) \quad a, b \in \mathbb{R} ,$$

(1)

where $a > 0$ represents the scale and $b$ is the translation. The functions $\psi_{a,b}$ are called wavelets and they share the properties of the mother function (note that $\psi = \psi_{1,0}$, so that the mother function is a member of the family defined by Eq. (1)).

The *wavelet transform* of an arbitrary function $x \in L^2(\mathbb{R})$ is a projection of that function onto the wavelet basis

$$T x(a,b) = \int_{\mathbb{R}} x(t) \psi_{a,b}(t) \, dt .$$

(2)

For a fixed scale $a_0$ and translation $b_0$, the wavelet transform of the function $x$, denoted by $X(a_0,b_0)$, represents its resemblance index to the wavelet $\psi_{a_0,b_0}$. If the index is “large” the resemblance is strong and vice versa. This index is termed the wavelet coefficient. For wavelets of compact support [19], the domain of integration in Eq. (2) is confined to the support. Consequently, the wavelet coefficient only depends on the part of the signal within the support. If the scale of the wavelet function is relatively small, the wavelet of compact support provides a local analysis tool suitable for capturing transient signal phenomena and for coping with non-stationarities.

A. Wavelet Functions for Spike Detection

It is important to choose a wavelet that is suitable for the signal of interest. Our choice is motivated by the shape of the extracellular potentials to be detected in the background noise. This shape can be explained with the help of biophysics. Despite the differences in extracellular and intracellular potentials pointed out in Section I, they are fundamentally related. The extracellular potential depends on the transmembrane current, which consists largely of a resistive component and a capacitive component. Since the capacitive component is proportional to the time derivative of the transmembrane potential, it is approximately bi-phasic\(^2\). The capacitive component may dominate the membrane current during an action potential, so the time course of the extracellular spike is typically bi-phasic \([4]\). Because a wavelet coefficient represents the measure of similarity between the signal and the basis function, it is reasonable to look for a wavelet that is spike-like. Accordingly, the neural signal would be represented

\(^2\)If the transmembrane action potential undergoes after-hyperpolarization, the capacitive current can be tri-phasic, with the third phase having significantly smaller amplitude than the rest of the waveform.
by a few coefficients. In approximation theory this is known as a sparse representation and the basis functions corresponding to these coefficients can be interpreted as intrinsic signal structures [15]. In the presence of noise, the sparse representation becomes an important condition for successful separation of signal and noise (see Section II-C).

Fig. 1 shows several different wavelet functions. The simplest wavelet is the Haar function. While it is compact, its discontinuity implies that it is not well localized in the frequency domain. The other wavelet functions shown are $db2^3$ and two wavelets from the family of biorthogonal wavelets: $bior1.3$ and $bior1.5$. Biorthogonal wavelets are constructed from splines and more about these wavelets can be found in [20]. We note that unlike $db2$, the biorthogonal wavelets are symmetric and that $db2$ wavelets do not have explicit analytic expression. Fig. 2 shows five spike templates recorded from the posterior parietal cortex of a Rhesus monkey (Macaca mulatta) using a single electrode (see Section III-A for recording details). Note that the two biorthogonal wavelets appear to match the intrinsic structure of the templates better than $db2$, as their bi-phasic shape is reminiscent of action potentials. It is therefore expected that the biorthogonal wavelets provide a sparser representation of neural signal than $db2$, which together with its near symmetric variate was used in [10].

B. Choice of Scale

The continuous wavelet transform defined by Eq. (2) operates on a continuous set of scales and translations. Hence, the basis functions $\psi_{a,b}$ are not orthogonal and the representation of the signal $x$ by its wavelet coefficients is redundant. One can choose dyadic scales and translations from a discrete set \{\(a = 2^j\); \(b = k2^j\); \(j, k \in \mathbb{Z}\)\}, so that the corresponding wavelets $\psi_{a,b}$ form an orthonormal basis of $L^2(\mathbb{R})$. Here, we will restrict the set of scales and translations in a different manner. Practically, all extracellularly recorded signals are sampled in discrete time. Thus we choose the set of basis function translations to be finite, where this set is determined by the sampling rate of the signal $f_s$ (KHz) and its duration $T$ (s), i.e. $b \in \mathcal{B}$, where

$$\mathcal{B} = \{0, 1, \cdots, k, \cdots, N - 1\},$$

This is a family of wavelets introduced by I. Daubechies, hence the name. The number, 2 in this case, represents the order of wavelet and is equal to the number of vanishing moments of $\psi$.3
and \( N = T f_s + 1 \) is the number of samples of the discrete signal (time series). Therefore, in the continuous wavelet transform\(^4\) the set of translations coincides with the discrete time vector.

Similarly, biophysical considerations of the duration of extracellular potentials can be used to restrict the relevant scales of the wavelet basis functions. Despite their variability in shape and amplitude, the vast majority of extracellular spikes are highly localized in time, with a characteristic duration. For example, the typical durations of action potentials in primate cortex lie in the range of 0.4 to 0.5 (ms) for signals recorded near the axon or initial segment, and 0.7 to 1.0 (ms) for spikes recorded in the vicinity of the soma-dendritic complex [4]. Assuming 20 samples per spike are sufficient for its accurate description, we arrive at Nyquist frequency of 20 to 40 (kHz). Indeed most neural data are acquired within this range of sampling rates. Similar results have been reported in [21], although depending on the species and the brain area, action potentials can last up to 3.0 (ms). Based on this biophysical knowledge of the duration of the transient to be detected, we can limit the set of scales for the analyzing wavelet functions. This practically serves to filter out a significant amount of noise and also appreciably reduce the real-time computational requirements. In summary, we use a limited set of scales

\[
\mathcal{A} = \{a_0, a_1, \ldots, a_j, \ldots, a_J\}
\]

where \( a_0 \) and \( a_J \) are determined from the sampling rate of the signal and the minimum and maximum durations of the spikes, denoted by \( W_{\text{min}} \) and \( W_{\text{max}} \), respectively, and these two parameters are chosen by the user. In contrast to the conventional dyadic scaling used in most wavelet applications, we choose the intermediate scales \( \{a_1, a_2, \ldots, a_{J-1}\} \) uniformly sampled between the two extrema \( a_0 \) and \( a_J \) with an arbitrary step chosen by the user. In the case of dyadic scaling, the scales of wavelet decomposition range from very fine to very coarse, where these limits are determined by the signal duration and sampling rate. The coefficients at very fine scale contain nothing but signal discontinuities and high frequency noise and are not relevant for detection of neural transients. Likewise, at coarser scales the information about relatively short transients is contaminated by the noise that is picked by the basis functions of excessively large support. This is another major difference between our approach and the one proposed by [10] that is based on dyadic scaling. Hereafter, the wavelet coefficient of discrete signal \( x \) at scale \( a_j \in \mathcal{A} \) and translation \( k \in \mathcal{B} \), is denoted by \( X(j, k) \), where \( X(j, k) = \langle x, \psi_{j,k} \rangle \), and \( \langle , \rangle \) stands for the inner product in \( \mathbb{R}^N \).

\(^4\)We keep the name continuous wavelet transform despite the discretization of signal, scales and translations. The name discrete wavelet transform is reserved for discrete signals with dyadic scales and translations.
C. The Statistics of Wavelet Coefficients

By applying the continuous wavelet transform, with possibly a restricted set of scales and translations, we obtain a multi-scale representation of the signal in terms of its wavelet coefficients. If the discrete observations $x$ contain useful signal $s$ and noise $w$, i.e.

$$x[n] = s[n] + w[n] \quad n \in B,$$

then the statistical properties of the wavelet coefficients are largely determined by those of the noise. In particular, if $w$ is a white Gaussian noise (WGN) with mean $\mu$ and variance $\sigma^2$, it follows from Eq. (3) and the properties of the wavelet transform that the mean and the variance of the coefficient $X(j, k)$ are

$$E\{X(j, k)\} = S(j, k) \quad \text{and} \quad \text{Var}\{X(j, k)\} = \sigma^2,$$

where $S(j, k) = \langle s, \psi_{j,k} \rangle$ is the wavelet coefficient corresponding to the true signal $s$. If the representation of the signal is sparse, only a few of the coefficients $S(j, k)$ are different from zero. In other words, the coefficients $X(j, k)$ corresponding to noise are zero-mean random fluctuations, and the coefficients $X(j, k)$ corresponding to signal plus noise are random variables with the mean appreciably different from zero.

For purposes of unsupervised signal detection, we must separate these coefficients by estimating the noise level $\sigma$ in each coefficient from the sampled data. To obtain these estimates, we borrow ideas from Donoho and Johnstone [22] who studied the problem of nonlinear estimation of signals from noisy data under the sparse representation. Their remarkably good wavelet denoising method is based on accepting only those wavelet coefficients that exceed a threshold, followed by the inverse wavelet transform. In our case, the threshold will become a part of the hypothesis testing procedure at the level of coefficients (see Section II-D), without the need for inverse transformation. They proposed a hard thresholding rule

$$\rho_T(X) = \begin{cases} X & \text{if } |X| > T \\ 0 & \text{if } |X| \leq T \end{cases}$$

which acts on the wavelet coefficients $X$ at different scales, and was shown to perform close, with respect to a certain metric, to an ideal estimator. For a near-optimal performance it is sufficient to choose the threshold $T = \sigma \sqrt{2 \log_{10} N}$, where, as before, $N$ is the number of samples of the analyzed time series. If the background noise is not white, the threshold becomes basis function dependent [15]. In this case, each wavelet coefficient generally has a threshold of its own.

From the translation invariance of the continuous wavelet transform and the underlying noise, it follows that all of the coefficients at scale $a_j$ have the same threshold, i.e. $T_j = \sigma_j \sqrt{2 \log_{10} N}$, where $\sigma_j^2$ is the
variance of the noise coefficients $W(j,k)$ at scale $a_j$. Since the coefficients $W(j,k)$ are not known, this variance must be estimated from the coefficients $X(j,k)$. Due to a possible presence of useful signal $s$ in the observation $x$, such an estimate will be biased. That is, the values of $X(j,k)$ significantly different from $W(j,k)$ will contain useful signal. However, such coefficients will be outliers because few coefficients contain signal due to our sparse representation. The sample variance of $X_j = \{X(j,k) : k \in B\}$ will be affected by those outliers, in particular it will over-estimate $\sigma^2_j$. From robust statistics [23], we know that the median of a random variable is less sensitive to outliers than its variance. For a Gaussian random variable (see Appendix I) it can be shown that the median of its absolute deviation effectively estimates the standard deviation, i.e.

$$\hat{\sigma}_j = M \{ |X_j - \bar{X}_j| \}/0.6745 . \quad (5)$$

Eq. (5) was derived under the assumption that the $X(j,k)$ are independent Gaussian random variables, which may not hold in general. It has been shown, however, that for a rich class of so-called $1/f$ processes, the wavelet decomposition at a fixed scale nearly whitens (decorrelates) coefficients at that scale [24]. In other words, a fixed scale wavelet decomposition of a colored random process tends to approximate the Karhunen-Loève decomposition, a canonical representation in which all expansion coefficients are uncorrelated, of the projection of such process to the wavelet basis. Under the Gaussian assumption, this implies that the sequence $X_j$ is independent. Fig. 3 shows a colored noise and its continuous wavelet representation for several different scales, together with the corresponding auto-covariance sequences (ACVS). It is apparent that the ACVS of the wavelet coefficients at different scales resemble the Dirac function, indicating that the coefficients are uncorrelated. Also note that the white noise approximation is less valid at larger scales, due to a significant amount of overlap in the basis functions. This can be circumvented by sub-sampling the sequence of wavelet coefficients at scale $a_j$, say $X_j = \{X(j,k) : k = 0, a_j, 2a_j, \cdots \}$ which makes the resulting sequence nearly white and the estimate given by Eq. (5) approximately valid. The estimate $\hat{\sigma}_j$ is then used in the detection algorithm that will be presented next.

D. The Detection Algorithm at a Single Scale

We now formulate the detection problem in the time-scale domain. The problem of detecting spikes in a noisy signal can be seen as a binary hypothesis testing problem, where under the null hypothesis $\mathcal{H}_0$ the signal is not present, and under the alternative $\mathcal{H}_1$ both signal and noise are present. More formally

$$\mathcal{H}_0 : x[n] = w[n] \quad n \in B$$

$$\mathcal{H}_1 : x[n] = s[n] + w[n] \quad n \in B ,$$

September 11, 2003 DRAFT
where $x[n]$ represents a noisy observation (evidence) at a discrete time $n$, $s$ is the transient (spike) to be detected and $w$ is the background noise. Because of the transient nature of $s$, the hypothesis $H_1$, if true, will be so only for an interval of time, or equivalently for a subset of the discrete time $\mathcal{B}$. Moreover, multiple transients could be present and these represent the main differences between the problems of the classical signal detection and the detection of transients.

We formulate the first step of our detection problem as a sequential binary hypothesis test at each scale $a_j \in \mathcal{A}$. Section II-E discusses how to combine the coefficient level decisions into an overall decision. Appendix II derives the following hypothesis testing rule for each wavelet coefficient $X(j,k)$:

$$|X(j,k)| \begin{cases} H_1 & \frac{\hat{\mu}_j}{2} + \frac{\hat{\sigma}_j^2}{\hat{\mu}_j} \log_e \gamma_j \leq \Theta_j, \\ H_0 & \forall k \in \mathcal{B}, \end{cases}$$  \hspace{1cm} (6)$$

where $\hat{\sigma}_j$ is determined from Eq. (5), $\hat{\mu}_j$ is the sample mean of the absolute value of the wavelet coefficients under the hypothesis $H_1$, $\gamma_j$ is a parameter that depends upon the acceptable costs of false alarms and omissions, denoted by $\lambda_{FA}$ and $\lambda_{OM}$, and the prior probabilities of the two hypotheses, denoted by $P(H_0)$ and $P(H_1)$ (see below and Appendix II for more details). The parameter $\Theta_j$ can be viewed as an acceptance threshold for the hypothesis $H_1$ at scale $a_j$.

Note that $\hat{\mu}_j$ cannot be estimated from $\mathcal{X}_j$, as many of the coefficients from $\mathcal{X}_j$ will contain noise only. However, we can get a rough estimate as to which coefficients contain noise only, by applying the thresholding defined by Eq. (4), where the value of the threshold is determined by $T_j = \hat{\sigma}_j \sqrt{2 \log_e N}$. Such a procedure splits $\mathcal{X}_j$ into two disjoint subsets: a noise subset $\mathcal{X}^n_j$ and a signal subset $\mathcal{X}_j^s$, where

$$\mathcal{X}_j^n \triangleq \{X(j,k) \in \mathcal{X}_j : \rho_{T_j}(X(j,k)) = 0\} \hspace{1cm} \mathcal{X}_j^s \triangleq \mathcal{X}_j \setminus \mathcal{X}_j^n.$$

Thus, $\hat{\mu}_j$ is found as a sample mean of $|\mathcal{X}_j^n|$. If $\mathcal{X}_j^n = \{\emptyset\}$, we define $\hat{\mu}_j = 0$.

In order to evaluate $\gamma_j$ we must specify the costs $\lambda_{OM}$ and $\lambda_{FA}$. The ratio $P(H_0)/P(H_1)$ of two prior\(^5\) probabilities is determined as $\|\mathcal{X}_j^n\|/\|\mathcal{X}_j^s\|$, where $\|\cdot\|$ stands for the size of a set. Also note that $\gamma_j$ depends only on the ratio of $\lambda_{FA}$ and $\lambda_{OM}$ (see Appendix II), and we can constrain the two costs by $\lambda_{FA} + \lambda_{OM} = 1$. In that case $\gamma_j$ can be conveniently reparametrized, i.e.

$$\log_e \gamma_j = \log_e \frac{\lambda_{FA}}{1 - \frac{\lambda_{FA}}{P(H_1)}} \triangleq L L_M + \log_e \frac{P(H_0)}{P(H_1)},$$  \hspace{1cm} (7)$$

where $L$ is in effect a new parametrization of $\lambda_{FA}/\lambda_{OM}$ and $L_M = 36.7368$ is a conveniently chosen scaling factor. The choice of $L_M$ is not unique and the value above corresponds to the natural logarithm of

\(^5\)Since these probabilities are estimated from the data, they are not priors in the strict Bayesian probability sense.
the maximum ratio of $\lambda_{FA}$ and $\lambda_{OM}$ that does not cause the arithmetic overflow under a double-precision floating point representation. The convenience of this reparametrization is that a relatively narrow range of $L$ covers a relatively wide range of false alarm-to-omission cost ratios. For example, $L \in [-0.188, 0.188]$ corresponds to $\lambda_{FA}/\lambda_{OM} \in [0.001, 1000]$, with $L = 0$ equivalent to $\lambda_{FA} = \lambda_{OM}$. Thus, for all practical applications it is sufficient to chose $L \in [-0.2, 0.2]$.

Once $L$ is selected (which is equivalent to the choice of $\lambda_{FA}/\lambda_{OM}$), the acceptance threshold $\Theta_j$ is determined from Eq. (6) and a decision is made at each scale $a_j$. Section II-E describes how the decisions at different scales are combined to produce a single decision. Before we proceed, let us closely examine an unlikely, but possible, scenario that could happen in the outcome of the thresholding outlined above.

The estimate of the acceptance threshold $\Theta_j$ for the hypothesis $H_1$ is data dependent, and largely depends on how the set $X_j$ is split into two subsets $X_j^p$ and $X_j^n$. The question we want to answer is what happens in the degenerate case when $X_j^p = \{ \emptyset \}$? Such a case happens when no wavelet coefficients exceed the threshold $T_j$, and is conceivable if the hypothesis $H_0$ is true $\forall n \in B$, i.e. observations contain noise only. Fig. 4 shows two such observation sequences containing colored and white Gaussian noise. If $X_j^p = \{ \emptyset \}$, it follows from Eq. (6) and (7) that $\Theta_j = \infty$ regardless of $L$, and we expect our detection algorithm to accept $H_0$ sequentially $\forall k \in B$, thereby not reporting any detected transients. Other detection methods, such as amplitude thresholding, are not capable of recognizing if observations contain any useful signal or noise only.

To illustrate this point, Table I shows the number of events detected from the two observation sequences given by Fig. 4 using two different detection methods: our wavelet method, which will be completed in Section II-E, and amplitude thresholding. The detection based on our wavelet method is parameter independent ($\Theta_j = \infty$) and produces no detected transients. The detection based on amplitude thresholding is performed for several different values of the threshold determined by the number of standard deviations of noise. It is clear that amplitude thresholding suffers from a number of false positives, even for conservative estimates of the threshold based on 3.75 or 4 standard deviations of the noise. This feature of wavelet detection algorithm may be useful if one is to avoid false positives, and this mode of operation of the algorithm is termed “conservative”.

Although the conservative mode is desirable regarding the probability of false alarm, it may be inadequate for some applications. Namely, if the observations contain a very weak transient signal (low SNR), the coefficients of useful signal and noise are rather similar, and the algorithm may refuse to detect any transients. The reason for this is that none of the coefficients will survive the hard thresholding given by (4). Consequently the set $X_j^p$ will be empty and the acceptance threshold of $H_1$ will never be attained.
TABLE I
THE NUMBER OF DETECTED TRANSIENTS FOR DIFFERENT CHOICE OF DETECTION PARAMETER USING TWO DIFFERENT
DETECTION METHODS: WAVELET (CONSERVATIVE AND LIBERAL) AND AMPLITUDE THRESHOLDING.

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter</th>
<th>Colored</th>
<th>White</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wavelet (conservative)</td>
<td>any</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Amplitude</td>
<td>3.00</td>
<td>11</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>3.25</td>
<td>8</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>3.50</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>3.75</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4.00</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Wavelet (liberal)</td>
<td>$\epsilon \in [0.0, 0.2]$</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Alternatively, we can slightly modify the algorithm so that it works in a “liberal” acceptance mode. In this case, we perturb the parameters comprising the acceptance threshold $\Theta_j$ that depend on $X_j^w$, namely the prior probability of $\mathcal{H}_1$ and $\hat{\mu}_j$, and assign them some non-zero values. Since $P(\mathcal{H}_1) = 0$ under $X_j^w = \{\emptyset\}$, we replace this probability by the smallest non-zero probability of $\mathcal{H}_1$, which is $P(\mathcal{H}_1) = 1/N$. This is equivalent to the assumption that there is a single wavelet coefficient at scale $a_j$ that exceeds the threshold $T_j$. Under this assumption we know that the sample mean of $X_j^w$ is at least $T_j$, therefore we choose $\hat{\mu}_j = T_j$. These will render the acceptance threshold $\Theta_j$ finite, and hence the mode of operation is termed “liberal”. Therefore, if $X_j^w = \{\emptyset\}$, the parameters in the liberal mode can be viewed as a local perturbation of the parameters in the conservative mode. If $X_j^w \neq \{\emptyset\}$, the two modes are equivalent. The bottom part of Table I shows that the performance of the algorithm is not compromised using liberal mode for a wide range of values of parameter $L$, and we will be using this mode of operation through the rest of the article.

E. The Overall Detection Algorithm

This section describes our overall spike detection and estimation algorithm. The methodology consists of two major steps:

i) Combine the decisions at individual scales into a single decision

ii) Estimate the spike arrival times.

**Combining Decisions at Individual Scales.** Because they are highly localized in time, the samples corresponding to neural transients occupy contiguous subsets of the discrete time vector $B$. This property
of transients is often referred to as a temporal contiguity. Temporal contiguity translates into the contiguity of the coefficients in the wavelet domain [25], i.e. the wavelet coefficients corresponding to the same transient tend to be neighbors in both time and scale. Since we use the continuous wavelet transform with the basis functions of compact support roughly matched to the scale of neural transients, the temporal contiguity in the wavelet domain is inherently preserved. The scale contiguity follows from a broad frequency spread of a time-limited signal, namely if a scale is thought of as an approximation of the frequency, a time-limited transient will be spread across many scales. The presence of noise, however, may obscure the picture at the scales that are not relevant. The scale contiguity can also be viewed in the present context as a cross-correlation (redundancy) of the wavelet coefficients (decisions) at different scales, respectively. Before we turn to the problem of redundancy, let us introduce a few notations.

Let \( B^H_{\alpha_j} \) be a subset of the translation set \( B \) that corresponds to the acceptance of \( H_1 \) at scale \( \alpha_j \), i.e. \( B^H_{\alpha_j} \triangleq \{ k \in B : |X(\alpha_j, k)| > \Theta \} \). If non-empty, the sets \( B^H_{\alpha_j} \) comprise a number of contiguous regions, where a contiguous region at scale \( \alpha_j \) is defined as the subset of the translation set \( B \) over which the hypothesis \( H_1 \) is accepted in succession at the scale \( \alpha_j \). These are illustrated by Fig. 5, which shows a segment of a simulated spike train, the wavelet coefficients that support the acceptance of \( H_1 \), and the corresponding translation indices defining \( B^H_{\alpha_j} \) together with its contiguous constituents. The scales are chosen assuming \( W_{\min} = 0.5 \) (ms) and \( W_{\max} = 0.8 \) (ms), so that the support of the basis functions at scales \( \alpha_0, \alpha_1, \alpha_2, \) and \( \alpha_3 \) are approximately 0.5, 0.6, 0.7 and 0.8 (ms), respectively. Since the duration of the noise-free transients is 0.6 (ms), we expect the transients to be best detected at scale \( \alpha_1 \). Note that at each scale \( \alpha_0, \alpha_1, \) and \( \alpha_2 \) there are three similar contiguous regions, and that the information provided by these scales is highly redundant. Also note that the first transient is hardly detected at scale \( \alpha_3 \) and that the second one is omitted, which is not surprising given that the most relevant scale for the transient considered is \( \alpha_1 \), and that the degradation in performance increases as scales deviate from \( \alpha_1 \).

The redundancy is lost by defining the combined set of acceptance of \( H_1 \), denoted by \( B^{H_1} \), as

\[
B^{H_1} = \bigcup_{\alpha_j \in \mathcal{A}} B^{H_1}_{\alpha_j}. \tag{8}
\]

By applying Eq. (8) the contiguities across different scales are combined. We note that such a procedure only makes sense in the case of the continuous wavelet transform because the basis functions at different scales are defined over the same set of translations \( B \). Therefore, if \( B^{H_1} \neq \{\emptyset\} \) it comprises a number of contiguous regions, where a contiguous region is defined as the subset of the translation set \( B \) over which the hypothesis \( H_1 \) is accepted in succession at any of the analyzing scales. The combined set of acceptance of \( H_1 \) containing three contiguous regions is shown at the bottom of Fig. 5.

September 11, 2003 DRAFT
Estimation of Spike Arrival Times. We now turn to the issue of estimating the spike arrival times from the wavelet coefficients supporting the acceptance of $\mathcal{H}_1$. One could estimate the spike arrival times in a two step process: (1) reconstruct a signal fragment corresponding to the wavelets whose coefficients pass the threshold test; and (2) estimate the arrival time as the peak of the reconstructed signal fragment. Here we take a different approach that doesn’t require signal reconstruction, but instead arrives at an estimate directly from the coefficients. This method more directly ends up with essentially the same result, while also implementing a few additional processing steps. Joel, I have some doubts about what was written in red (above). What you proposed makes perfect sense when scales and translations are chosen from dyadic set, in which case the basis functions are orthonormal. In that case the reconstruction simply becomes the weighted sum of the basis functions. If the basis is not orthogonal, e.g. wavelet packet, it is not clear how to do the reconstruction, as we are facing a non-uniqueness issue. Different regularization techniques are proposed e.g maximum entropy, which provides a best tree. I am not sure if regularization algorithms exist for the continuous wavelet transform which is hyper-redundant, and I have never seen any document that deals with reconstruction in the continuous wavelet transform. We could probably argue that the inverse transformation is unnecessary even when it is possible (e.g dyadic scales and translation), and that there is nothing wrong with doing the estimation in the wavelet domain.

In a noise-free environment, the wavelet function that provides the maximum correlation with the transient to be detected, corresponds to a wavelet coefficient of maximum magnitude. The time associated with the translation index of the wavelet function with maximal coefficient can be taken as a good approximation to the occurrence time of the underlying transient. Because we choose wavelet supports the set of translations $\mathcal{B}$ with time resolution down to the smallest sampling period, this approximation is essentially as good as the sampling period. Tracking of modulus maxima of the wavelet coefficients across scales has been proposed for the detection of signal singularities [26]. In a noisy environment, there is naturally a jitter associated with the location of this maximal coefficient. This jitter can be reduced by averaging the locations of the maxima across different scales. This is the heuristic behind our approach.

To estimate the spike arrival times, we start by organizing the acceptance set of $\mathcal{H}_1$ into its contiguous constituents, i.e.

$$\mathcal{B}^{\mathcal{H}_1} = \bigcup_{i=1}^{N_c} \mathcal{C}_i^{\mathcal{H}_1}, \quad (9)$$

where $\mathcal{C}_i^{\mathcal{H}_1}$ are the contiguous regions of $\mathcal{B}^{\mathcal{H}_1}$ and $N_c$ is the number of contiguous regions. Let $T_i^j$ be the estimated location of the $i$th transient at scale $a_j$, i.e. it is the translation index of the maximum
magnitude coefficient in the \(i\)th contiguous region at scale \(a_j\):

\[
T_i^j \triangleq \arg \max_{k \in C_i^j} \{|X(j, k)| : |X(j, k)| > \Theta_j\} \quad \forall i = 1, 2, \cdots, Nc \quad \forall a_j \in A.
\] (10)

Note that \(T_i^j\) may not exist for all scale-contiguous region pairs. For example, \(T_1^3\) and \(T_3^3\) are not defined in Fig. 5, as no coefficient at scale \(a_3\) supports the acceptance of \(H_1\) over the sets \(C_1^H_1\) and \(C_3^H_1\). The arrival time candidate of the \(i\)th transient is then found by averaging \(T_i^j\) over those scales that support the acceptance of \(H_1\) on the \(i\)th contiguous region:

\[
T_i = \frac{1}{\|A_i^{H_1}\|} \sum_{a_j \in A_i^{H_1}} T_i^j \quad \forall i = 1, 2, \cdots, Nc,
\] (11)

where \(A_i^{H_1} \triangleq \{a_j \in A : |X(j, k)| > \Theta_j, k \in C_i^{H_1}\}\), and, as before, \(\|\cdot\|\) is the size of a set. In the example given by Fig. 5, \(A_1^{H_1} = \{a_0, a_1, a_2\}\), \(A_2^{H_1} = \{a_0, a_1, a_2, a_3\}\), and \(A_3^{H_1} = \{a_0, a_1, a_2\}\). That is, the arrival time estimate is derived from the arrival time estimates at each different scale. Before the candidate \(T_i\) is declared as the arrival time of an individual transient, an additional processing step may be necessary.

First note that as a consequence of the multiphasic shape of wavelet functions (see Fig. 1), the coefficients corresponding to a single transient will be both positive and negative over a subset of the translation set \(B\). This sign change occurs through a smooth transition due to a fine resolution of the translation set \(B\). After the threshold operation (6) is applied, only the most positive/negative coefficients will survive, resulting in the distribution of the wavelet coefficients in two (or more) contiguous regions. An example of this phenomenon is the first transient from Fig. 5 which has two corresponding contiguous regions. However, the disjoint contiguous regions that arise from a single transient will be very close in time, and the spurious transients associated with this phenomena can be reliably eliminated by merging the candidates that are sufficiently close in time, followed by a re-estimation of the arrival time based on the average of the merged intervals. This type of post processing is performed sequentially from the onset of the signal for two candidates at a time. The intermediate results are updated and the process is repeated until no further merging is possible. The arrival time of the first transient from Fig. 5 is estimated in this fashion, where the candidates \(T_1\) and \(T_2\) are combined to produce a single arrival time. The only remaining unknown is how close two candidates need to be in order to be combined. The answer to this question depends on the largest scale \(a_J\) used in the analysis, as the distance between two arrival times belonging to the same transient, tend to increase with the scale (see Fig. 5). Recall that the coarsest scale \(a_J\) is determined from the maximal duration of the transient (see Section II-B), therefore
two transients will be merged if their distance is less than $W_{\text{max}}$, where $W_{\text{max}}$ is the maximum duration of the transient to be detected.

This method, therefore, successfully eliminates the spurious candidates that belong to a single spike. The obvious drawback, however, is that we are unable to resolve two genuine transients that are closer than $W_{\text{max}}$. For a well isolated single unit recording, this is not a serious constraint, as two spikes cannot be generated arbitrarily close in time due to the refractory period, which can last up to a few milliseconds [27]. However, when recording from multiple neurons, the spikes that originate from different units can partially overlap. Thus, the post-processing of the spike candidates should be viewed as a trade-off between eliminating spurious transients and inability to resolve partially overlapped spikes.

Once the occurrence times of the spike events are estimated, the spikes are extracted by copying a fixed number of samples before and after the time of occurrence. This results in a vector representation of individual events, convenient for further analysis. We proceed by testing the success of the proposed algorithm in a realistically modeled detection task. Throughout the rest of the article, we will refer to this algorithm as the wavelet detection method (WDM).

III. Simulation Results

Analytical assessment of the performance of the detector (6) requires a full knowledge of the statistics of the wavelet coefficients under both hypotheses. In the absence of this knowledge, one can resort to Monte Carlo simulations in order to evaluate the detector performance. To ensure consistent results, the number of Monte Carlo trials needs to be sufficiently high and individual trials have to be independent. It is all but impossible to properly test the detection performance on real extracellular signals, as we do not have an independent information about the number of spike transients and their exact arrival times in each trial. The only reliable way to obtain this information is to perform simultaneous intracellular recording [6], [28], which is virtually impossible in behaving animals. Hence, to rigorously test the performance of this method, we use simulations that synthesize spike trains from actual recorded data.

A. Modeling Neural Data

One way to model neural data is to extract spikes from actual recordings, arrange them randomly in time with an arbitrary firing rate and corrupt them by a suitably modeled “neural noise”, where the noise level is determined by the value of SNR chosen for study. Fig. 6 shows a high SNR extracellular data set collected from the posterior parietal cortex of a Rhesus monkey. A single platinum-iridium microelectrode (Frederic Haer Company, Bowdoinham, ME) with the nominal impedance of 2 (MΩ) at 1 (kHz) was used...
for the recording. The data was acquired through a recording system (Plexon Inc, Dallas, TX) with a pre-amplifier and a band-pass filter (band 154 (Hz)-13 (kHz)). The signals were amplified and digitized (12 bit A/D converter, digitization rate 40 (kHz)) by a data acquisition card (PCI-MIO 16E-4 with LabView (National Instruments, Austin, TX)). A good separation of signal and noise in this data enables us to detect spikes reliably. Transients are detected from the data using the WDM with \( L = 1.0 \). This value of the parameter imposes an extremely high penalty for false alarms \( \lambda_{FA}/\lambda_{OM} = \exp(37.7368) \), and ensures a low probability of Type I error. The detected spikes \( S = \{ S_1, S_2, \cdots, S_N \} \) were then aligned using a maximum correlation method. Briefly, the spikes are normalized \( \|S_n\|_2 = 1 \), and the first detected spike is fixed. The second spike is then shifted locally in time until its correlation with the first one is maximal. The third spike is now shifted back and forth until the sum of its correlations with the first two spikes is maximal and so on. A similar alignment procedure has been proposed by [8]. This procedure effectively re-estimates the time of occurrence of each detected spike. Some \( N_s = 285 \) spikes were detected and aligned in this manner. The spikes were scaled back to their original size and their peak-to-peak amplitudes were calculated. From the scatter plot of the peak-to-peak amplitudes, five different clusters were visually identified. The spikes were clustered using the k-means method [29] and the cluster averages were calculated, resulting in five templates which are shown in Fig. 2. We use multiple spike template shapes in simulations to ensure that the algorithm is not overly sensitive to a particular waveform shape. The same template construction procedure was repeated using amplitude thresholding detection with the threshold set so that 285 spikes are detected, and the resulting templates did not show significant deviations from the ones shown in Fig. 2. Therefore, the use of WDM in the construction of templates does not impose any bias in the further analysis.

These templates were used to generate spike trains with known properties as follows. Spike onset times were generated by a homogeneous Poisson process with a certain firing rate (FR) and a refractory period of 2 (ms) enforced. The simulation was terminated once the number of arrival times reached a pre-specified number \( N_a \). The five templates were then randomly drawn with equal relative frequencies and were successively centered at the arrival times, barring boundary conditions. Fig. 6 also shows a monkey extracellular data record (recorded as above) that did not contain any visible spikes and was therefore used as a template for neural noise\(^6\). The subsequent panels show the periodogram (power spectrum)

\(^6\)The boundary between signal and noise in the context of extracellular recording is not clearly defined due to the fact that a large component of noise represents the activity of distant neurons. A data record such as the one shown on Fig. 6 is therefore treated as noise, even though it is likely to contain low amplitude spikes.
and the histogram of this neural noise signal. Note the characteristic $1/f$ trend in the spectrum and bell-shaped noise histogram indicating the Gaussian trend in the data. Gaussian distribution of background neural noise has been reported by [30], [31]. Each spike within the spike train was normalized so that $\|S_i\|_\infty = 1$, and a randomly selected sub-segment of the neural noise was scaled according to the SNR desired for each test. That is, the standard deviation of the neural noise was defined according to

$$\sigma \triangleq \frac{\|S_i\|_\infty}{\text{SNR}} = \frac{1}{\text{SNR}}$$

and added to the generated spike train. This procedure was repeated many times for different FR and SNR values. For each choice of FR and SNR, many trials were performed. Also note that random arrangements of the spike templates effectively remove any systematic bias.

### B. A Note on SNR

In communication theory, SNR at time $t$ is defined as the absolute value of the signal amplitude at time $t$ divided by the standard deviation of the noise [32]. Such a definition of SNR is time dependent and is not particularly useful in neurophysiology, where SNR can be viewed as a single number that characterizes the noisiness of a spike train. Colloquially, SNR can be defined as the ratio of the maximum amplitude of the spike and the “amplitude” of the background noise, hence the definition (12). If a spike train contains multiunit activity, different spikes are likely to have different SNRs, thus we need to apply averaging in order to obtain a single SNR. The “average” SNR may be biased depending on the relative frequency of spikes of different amplitudes, and we avoid this problem by normalizing the spike templates. Another commonly used definition of SNR involves the root-mean-square values of the spike and noise [7], and can be recovered from Eq. (12) under the Euclidean norm and a scale factor. This definition of SNR gives lower values than the one given by Eq. (12), under identical noise variance. Finally, the definition of SNR may involve the power of signal and noise as opposed to their amplitudes. This discrepancy can be reconciled by expressing the SNR in (dB) $(20 \log_{10}(\text{SNR})$ for amplitude and $10 \log_{10}(\text{SNR})$ for power). Because of this variability in the SNR definition, it is often difficult to objectively assess the level of noise in the data. To alleviate this difficulty and give some feeling for the noise level, Fig. 7 provides a snapshot of the simulated data at two different SNRs.

### C. Performance Evaluation

The performance of the WDM was tested at low, medium and high firing rates of 10, 30, and 100 (Hz). Because the spike arrival times follow the statistics of a Poisson process, the length of individual
trials is variable. Setting the total number of spikes per trial to $N_a = \text{FR}$, renders the average duration of trials to be approximately one second. Relatively high noise levels $\text{SNR} = \{3.5, 3.6, \ldots , 4\}$ were applied in the analysis.

The performance of the WDM was compared against other methods including power detection method (PDM), single amplitude thresholding method (SATM) (see Introduction) and double amplitude thresholding method (DATM). The performance of these methods was assessed using the receiver operating characteristic (ROC). The ROC curve compares the probability of (correct) detection ($P_D$) versus the probability of false alarm ($P_{FA}$). The ROC curve for each combination of FR and SNR was obtained by averaging the performance over 300 trials. Appendix III describes the details of our averaging methodology.

A recent detection method using nonlinear energy operator (NEO) [33] reports a successful detection under nearly 0 (dB), where, the authors defined SNR as a peak-to-peak amplitude of the smallest spike template divided by the noise standard deviation. However, without specifying the amplitudes and relative frequencies of other templates, followed by averaging, as discussed in Section III-B, SNR can be manipulated to any desired value. In other words, such a definition of SNR does not provide an objective measure of the noisiness of the data, as larger amplitude spike templates will have larger SNRs. Furthermore, the authors generated neural noise using autoregressive moving average (ARMA) model, although it has been shown that ARMA models are inadequate for capturing long-term correlation structure typically associated with $1/f$ processes [24], [34]. We tested the performance of the NEO detector under a properly defined SNR and realistic neural noise model and found its performance at best comparable, and often falling short of other conventional methods. Therefore, the detection using NEO will not be a subject of further investigation.

The results of our simulation tests are shown in Fig. 8, which describes the performance of the method for different combinations of firing rate and SNR. Each ROC curve is obtained by a systematic variation of the threshold parameter; parameter $L$ in the WDM and parameters $T_s$, $T_d$, and $T_p$, which represent a multiple of the estimated noise standard deviation, in SATM, DATM and PDM, respectively. Clearly, the ROC corresponding to WDM lies above the ROC of other detection methods. In other words for a fixed probability of false alarm $P_{FA}$, the probability of detection $P_D$ of the WDM is consistently bigger than those of the other methods. Conversely, at the same level of $P_D$, the WDM has consistently lower false alarm rate than the other methods. The improvement in performance is more apparent under low SNR and low FR conditions, which are commonly encountered in everyday recording. Also note that our results are consistent with the results of [7] which showed that the PDM outperforms SATM. Likewise, SATM
slightly outperforms DATM. This is not surprising given the near symmetric shape of the templates (see Fig. 2), making the double amplitude thresholding redundant, thereby increasing the probability of false alarm. In general, the two phases of a spike waveform may be substantially different in amplitude, and if one is to perform an unsupervised detection, double thresholding is certainly a better solution.

We also note that unlike the classical ROC which corresponds to the performance of binary hypothesis test (yes/no decision) and should lie above the chance (line $P_D = P_{FA}$), the ROCs shown in Fig. 8 have somewhat different character. First, they correspond to the result of sequential hypothesis testing. Second and more importantly, the signals to be detected are transient and their arrival times need to be estimated. Thus, the ROCs from Fig. 8 represent combined results of detection and estimation. Since the estimation problem is much harder and more sensitive to the presence of noise than detection [11], a chance detector would have performance with $P_{FA} \gg P_D$. Therefore, it is possible that the performance of detector and estimator lie below the line $P_D = P_{FA}$, although this only tend to happen for SATM and DATM and only under relatively low $P_D$ conditions.

**Parameter Settings for Unsupervised Applications.** For unsupervised detection one does not have the luxury of varying the detection threshold, hence a single threshold that “works” across a wide range of SNRs and FRs is sought. We investigate how choosing a single parameter affects the performance of different methods, and find that the WDM offers additional advantage over the other detection methods, as shown in Fig. 9. For approximately equal levels of $P_D$, panel (A) vs. panels (B),(C) and (D), we see that the dispersion of $P_{FA}$ of the WDM is much smaller than that of the other methods. Not only does the WDM provide a smaller $P_{FA}$, but also it has a more consistent behavior over a range of FRs. In addition, for a fixed FR, $P_{FA}$ remains fairly constant across different SNRs, as can be seen as almost vertical performance curves of panel (A). On the contrary, the performance curves of the other methods tend to be slanted (panels (B), (C) and (D)), and this becomes more apparent at low FR. We conclude this analysis by noting that the value of parameter $L = 0$ (panel (A)) offers a reasonable compromise between $P_{FA}$ and $P_D$ while maintaining a consistent performance over a wide range of FRs and SNRs. Therefore, for unsupervised spike detection, this might be a good choice of the parameter.

**Timing Jitter.** As outlined above, spike detection methods can be viewed as a combination of decision theory and parameter estimation, where estimated parameters are the arrival times of individual spikes. The presence of noise causes jitter in the estimated occurrence times (see Fig. 5). Two important performance parameters of any estimator are its bias and consistency. Ideally, an estimator should be unbiased with consistency proportional to the number of observations. Table II shows estimated jitter and its standard deviation averaged over different SNR-FR scenarios. The results correspond to WDM with $L = 0$ and
SATM with $T_s = 3.6$, and represent averages over 300 trials for each SNR-FR combination. The negative value of jitter indicates that the estimated spike times are lagging the true spike times. On average, WMD has a bias that is nearly 4 times smaller than that of SATM. On the other hand, the standard deviation of the jitter is smaller in the case of SATM, and this can be viewed as a bias-variance trade-off. Despite their presence, bias and inconsistency do not pose a serious threat in the process of spike detection for several reasons. First, their values are rather small compared to the duration of spike transients and they are comparable to the sampling period of the data 0.025 ms (40 kHz). Second, they can be eliminated through the process of spike alignment (see Section III-A), which effectively re-estimates the occurrence times of individual events.

**Characteristics of WDM False Positives.** Finally, we performed an analysis of the falsely detected spikes under the WDM ($L = 0$) and SATM ($T_s = 3.6$) to assess the nature of the failure modes of the two approaches. Our conjecture is that the false positives of WDM will resemble true spikes more than the false positives of SATM. Fig. 10 summarizes the results for SNR=3.5 and FR=10 (Hz), although similar behavior was observed over all SNR-FR combinations. The false positives over 300 trials for the two methods were extracted. Falsely identified spikes were then represented in a 2-D feature space by means of principal components. Although projected to the same subspace, the false positives corresponding to WDM (panel (A)) and SATM (panel (B)) are shown on separate plots for clarity. The false positives corresponding to WDM tend to cluster in 4 distinct groups, where those of SATM fall into a single broad cluster. The features of the five spike templates used for data modeling are also shown in these plots. Panels (C) and (D) show the cluster average waveforms, so that each spike within the cluster can be treated as a noisy realization of the corresponding waveform. The majority of the falsely detected spikes by the WDM are bi-phasic (blue and cyan waveforms) with a small number of tri-phasic spikes (red and green waveforms), where most of the false positives detected by the SATM are mono-phasic and do not have the realistic spike shape. Furthermore, the spike templates fall into one of the clusters of WDM, and this is not surprising given the shapes of the templates (Fig. 2) and the shape of blue waveform (panel (C)). This happened consistently for all SNR-FR combinations. Therefore, the false positives of WDM are well structured and could represent the activity of distant neurons, where the false positives of SATM mostly represent random fluctuations of measured voltage. Hence, one can generally make a good use of false positives of WDM where false positives of SATM have to be discarded as noise. For real applications, this has to be done through the process of spike clustering [30], [35], [36].
TABLE II

BIAS AND STANDARD DEVIATION OF ESTIMATED SPIKE ARRIVAL TIMES FOR WDM AND SATM AVERAGED OVER DIFFERENT SNR AND FR SCENARIOS.

<table>
<thead>
<tr>
<th></th>
<th>WDM</th>
<th>SATM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jitter (ms)</td>
<td>Std (ms)</td>
<td>Jitter (ms)</td>
</tr>
<tr>
<td>Average</td>
<td>-0.0396</td>
<td>0.0633</td>
</tr>
</tbody>
</table>

IV. CONCLUSIONS

Despite the large number of existing algorithms for detection of extracellular potentials in noisy observations, robust, fully automated detection algorithms have been scarce. We presented a novel detection scheme and compared its performance to many commonly used spike detection methods. The detection is cast in the standard hypothesis testing framework and since the signal to be detected is unknown, the detector performance is representation dependent. Spike transients have historically been detected by a simple amplitude thresholding, where the threshold level is chosen with respect to the (estimated) standard deviation of the noise. We have shown how this can yield erroneous results, especially if no signal is present. Spike waveforms are not just samples whose average amplitude exceeds some baseline level. They also have a characteristic shape and duration. Using wavelets we are taking advantage of this additional information that is ignored by amplitude or power thresholding methods.

The algorithm is completed by combining detection, which arises from sequential hypothesis testing, and parameter estimation, where the occurrence times of individual spike transients are the parameters of interest. Additional post processing of estimated arrival times may be necessary. This may result in inability to resolve the transients that are within a millisecond of each other. However, the same problem is inherent to any other detection method, and there have been some attempts to provide solution [7], [35], although in the context of spike classification rather than spike detection.

The proposed algorithm has been extensively tested over a wide range of conditions via Monte Carlo simulations of spike trains that were synthesized from actual data. We found that the WDM consistently outperforms other common detection methods, and the differences have been quantified. The basic findings are: i) the wavelet method provides a significant improvement under extremely low SNRs and low FRs, a situation commonly found in actual experiments; ii) for a single choice of detection parameter, the wavelet detection method offers more consistent performance under different SNR-FR scenarios; iii) therefore, it
is possible to come up with a single parameter that performs well for a wide range of SNRs and FRs, which is very useful for unsupervised on-line applications, where no prior knowledge about the data can be utilized; iv) the jitter in estimated spike arrival times and its variance are comparable for WDM and commonly used amplitude thresholding; v) falsely detected spikes of WDM are likely to be caused by the activity of distant neurons and can be utilized as neural data, as opposed to false positives of amplitude thresholding methods, which typically represent random fluctuations (noise).

We conclude the article by noting that the wavelet method is also suitable for off-line analysis, where by trading off the cost of omission and false alarm, the user can modify the sensitivity of the method, similar to changing the threshold level in amplitude thresholding method.

APPENDIX I

MEDIAN ABSOLUTE DEVIATION

Let $X$ be a Gaussian random variable with mean $\mu$ and variance $\sigma^2$, denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$. Let us define random variables $Y \triangleq X - \mu$ and $Z \triangleq |Y|$. Clearly $Y \sim \mathcal{N}(0, \sigma^2)$ and the probability density function (pdf) of $Z$ can be written as

$$f_Z(z) = \frac{dP(Z \leq z)}{dz}, \quad (13)$$

where $P(Z \leq z)$ is the cumulative distribution function of $Z$. It follows from the definition of $Z$ that

$$P(Z \leq z) = \begin{cases} P(-z \leq Y \leq z) & z \geq 0 \\ 0 & z < 0 \end{cases}. \quad (14)$$

Combining Eqs. (13) and (14) yields

$$f_Z(z) = \begin{cases} \frac{dz}{dz} \int_{-z}^{z} f_Y(y) \, dy = 2 f_Y(z) & z \geq 0 \\ 0 & z < 0 \end{cases}, \quad (15)$$

where the result follows after differentiating the integral from above with respect to the upper and lower bounds. The median of $Z$, denoted by $M\{Z\}$, can be defined as the middle point of the pdf $f_Z(z)$, i.e.

$$\int_{-\infty}^{M\{Z\}} f_Z(z) \, dz \triangleq \frac{1}{2}. \quad (16)$$

By substituting Eq. (15) into (16) and by noting that $f_Y(z) = (2\pi\sigma^2)^{-1/2} \exp(-z^2/2\sigma^2)$ one obtains

$$M\{Z\} = \sigma \sqrt{2} \text{erf}^{-1}\left(\frac{1}{2}\right), \quad (17)$$

where erf stands for the error function defined by $\text{erf}(x) = 2(\pi)^{-1/2} \int_{0}^{x} \exp(-t^2) \, dt$, and erf$^{-1}$ represents its inverse. After numerical evaluation it follows that $M\{Z\} \approx 0.6745 \sigma$. The significance of this result
is that one can use the median of a sample for estimation of its standard deviation. Since the median is much less susceptible to the presence of outliers, it provides a more robust estimate of the standard deviation. Let $\mathcal{X}_j$ be a sequence of $N$ independent identically distributed Gaussian random variables with mean $\mu_j$ and variance $\sigma_j^2$. From the analysis above, it follows that $\hat{\sigma}_j = M \left( |\mathcal{X}_j - \hat{\mathcal{X}}_j| \right) / 0.6745$, where $M$ is the sample median of $|\mathcal{X}_j - \hat{\mathcal{X}}_j|$ and $\hat{\mathcal{X}}_j \triangleq 1/N \sum_k X(j,k)$ is the sample mean of $\mathcal{X}_j$.

**APPENDIX II**

**SEQUENTIAL BINARY HYPOTHESIS TESTING**

Recall that we formulate our detection problem as a binary hypothesis testing problem, where under the null hypothesis $\mathcal{H}_0$ the signal is not present, while under hypothesis $\mathcal{H}_1$ the signal is present:

$$
\mathcal{H}_0 : \quad x[n] = w[n] \quad n = 0,1,\ldots,N-1 \\
\mathcal{H}_1 : \quad x[n] = s[n] + w[n] \quad n = 0,1,\ldots,N-1 ,
$$

where $x[n]$ represents a noisy observation at time $n$, $s$ is the signal to be detected and $w$ is noise. By the linearity of wavelet transform at scale $a_j \in \mathcal{A}$, these two hypotheses have the following form:

$$
\mathcal{H}_0 : \quad X(j,k) = W(j,k) \quad k \in \mathcal{B} \\
\mathcal{H}_1 : \quad X(j,k) = S(j,k) + W(j,k) \quad k \in \mathcal{B} .
$$

As in any hypothesis testing problem, the goal is to determine whether the evidence supports the rejection of $\mathcal{H}_0$. This decision should be made optimally with respect to a suitably chosen objective function. Let $\mathcal{R}(\mathcal{H}_0|X)$ ($\mathcal{R}(\mathcal{H}_1|X)$) be a conditional risk associated with accepting (rejecting) the hypothesis $\mathcal{H}_0$ given the evidence $X$, respectively. These risks can be expressed as:

$$
\mathcal{R}(\mathcal{H}_0|X) = \lambda_{00} P(\mathcal{H}_0|X) + \lambda_{01} P(\mathcal{H}_1|X) \\
\mathcal{R}(\mathcal{H}_1|X) = \lambda_{10} P(\mathcal{H}_0|X) + \lambda_{11} P(\mathcal{H}_1|X) ,
$$

where $\lambda_{ij} \geq 0$ is the cost of accepting $\mathcal{H}_i$ given that the true state of nature is $\mathcal{H}_j$ and $P(\mathcal{H}_i|X)$ is the probability of accepting $\mathcal{H}_i$ given the evidence. It is customary not to penalize for correct decisions, therefore $\lambda_{00} = \lambda_{11} = 0$. Furthermore, $\lambda_{10}$ represents the cost of rejecting $\mathcal{H}_0$ when it is true (Type I error or false alarm) and $\lambda_{01}$ is the cost of accepting $\mathcal{H}_0$ when it is not true (Type II error or omission error). Hence, we adopt the following notation: $\mathcal{R}(\mathcal{H}_0|X) = \lambda_{OM} P(\mathcal{H}_1|X)$ and $\mathcal{R}(\mathcal{H}_1|X) = \lambda_{FA} P(\mathcal{H}_0|X)$, where $\mathcal{R}(\mathcal{H}_0|X)$ and $\mathcal{R}(\mathcal{H}_1|X)$ are the conditional risks associated with false alarms and omissions. The
overall Bayes risk is then defined as:

$$\mathcal{R} = \int_{\Omega} \left[ \mathcal{R}(\mathcal{H}_0|X) + \mathcal{R}(\mathcal{H}_1|X) \right] p(X) \, dX,$$

(18)

where $p(X)$ is the pdf of $X$. The goal of Bayes decision theory is to partition the space $\Omega$ into acceptance regions for the two hypotheses while minimizing the cost (18). Not surprisingly, the optimal decision rule is to accept the hypothesis with a smaller conditional risk [37], i.e.

$$\mathcal{R}(\mathcal{H}_0|X) \geq \mathcal{R}(\mathcal{H}_1|X).$$

(19)

That is, Eq. (19) is read as: “accept $\mathcal{H}_1$ if $\mathcal{R}(\mathcal{H}_0|X) > \mathcal{R}(\mathcal{H}_1|X)$, and vice versa”. After invoking the Bayes rule $P(\mathcal{H}_i|X) = p(X|\mathcal{H}_i) \, P(\mathcal{H}_i)/p(X)$, the decision rule (19) becomes

$$\frac{p(X|\mathcal{H}_1)}{p(X|\mathcal{H}_0)} \overset{\mathcal{H}_1}{\underset{\mathcal{H}_0}{\gtrless}} \frac{\lambda_{FA} \, P(\mathcal{H}_0)}{\lambda_{OM} \, P(\mathcal{H}_1)} \triangleq \gamma,$$

(20)

where $p(X|\mathcal{H}_i)$ is the likelihood of $X$ given $\mathcal{H}_i$ and $P(\mathcal{H}_i)$ is the prior probability of the hypothesis $\mathcal{H}_i$. Note that $\gamma$ represents the acceptance threshold for $\mathcal{H}_1$ and in the typical case that $\lambda_{FA} > \lambda_{OM}$ and $P(\mathcal{H}_0) \gg P(\mathcal{H}_1)$, we have that $\gamma \gg 1$. Under the assumption that the noise is Gaussian we have $p(X|\mathcal{H}_0) \sim \mathcal{N}(0, \sigma^2)$ and $p(X|\mathcal{H}_1) \sim \mathcal{N}(\mu, \sigma^2)$, where for $\mu > 0$ for positive coefficients, and $\mu < 0$ for negative coefficients. To ensure compact notation, it is preferred to use $|X|$ instead of $X$. Then, by the Gaussian assumption one can equivalently rewrite Eq. (20) as:

$$|X| \overset{\mathcal{H}_1}{\underset{\mathcal{H}_0}{\gtrless}} \frac{\mu}{2} + \frac{\sigma^2}{\mu} \log_e \gamma,$$

(21)

where $\mu$ is the mean value of $|X|$ under the hypothesis $\mathcal{H}_1$ and $\sigma$ is the standard deviation of $X$. The parameters $\mu$ and $\sigma$ are not known and have to be estimated from the data. The same stands for the prior probabilities $P(\mathcal{H}_0)$ and $P(\mathcal{H}_1)$. In this case, the test (20) becomes essentially GLRT and its derivative (21) represents a sufficient statistic, and the only difference between GLRT and Bayes decision is in the way the acceptance threshold $\gamma$ is chosen. The condition (21) is checked sequentially ($\forall k \in B$) at each scale $a_j \in A$. 
APPENDIX III
AVERAGING OVER TRIALS

For ith trial $P_D$ and $P_{FA}$ are estimated as

$$\hat{P}_D^{(i)} = \frac{N_{cd}^{(i)}}{N_a^{(i)}} \quad \text{and} \quad \hat{P}_{FA}^{(i)} = \frac{N_{fa}^{(i)}}{N_d^{(i)}},$$

where $N_{cd}^{(i)}$ is the number of correctly detected spikes per trial, $N_a^{(i)}$ is the total number of spikes per trial, $N_{fa}^{(i)}$ is the number of falsely detected spikes per trial, and $N_d^{(i)}$ is the total number of detected spikes per trial. A correct detection is declared if the detected arrival time is within 0.5 ms of the true arrival time. The overall estimate is obtained as an ensemble average, i.e. $\hat{P}_D = \left\langle \hat{P}_D^{(i)} \right\rangle$ and $\hat{P}_{FA} = \left\langle \hat{P}_{FA}^{(i)} \right\rangle$.

Since $N_a^{(i)}$ is constant across trials, the first average simply becomes

$$\hat{P}_D = \frac{1}{M} \sum_{i=1}^{M} \hat{P}_D^{(i)},$$

where $M$ is the total number of trials ($M = 300$ in our case). However, the number of detected spikes $N_d^{(i)}$ is not constant over trials, therefore the trials cannot be given uniform weights in the estimate of $\hat{P}_{FA}$. Since $\hat{P}_{FA}^{(i)}$ is conditioned upon $N_d^{(i)}$, we write based on the law of total probability

$$\hat{P}_{FA} = \sum_{k=0}^{\infty} \hat{P}_{FA|k} P(N_d = k),$$

where $\hat{P}_{FA|k}$ is the estimate of $P_{FA}$ given that the number of detected spikes is $k$, and $P(N_d = k)$ is estimated as the number of trials with $k$ detected spikes divided by $M$. We also note that

$$\hat{P}_{FA|k} = \begin{cases} \frac{N_{FA|k}}{k} & k \in \mathbb{N} \\ 0 & k = 0 \end{cases},$$

where $\hat{N}_{FA|k}$ is the average number of false alarms given that the number of detected spikes is $k$.

REFERENCES


**FIGURE CAPTIONS**

**Fig. 1:** Wavelet functions of four different families: (A) Haar, (B) db2, (C) bior1.3, and (D) bior1.5.

**Fig. 2:** Five different spikes identified in a single electrode recording from the cortex of a Rhesus monkey. The spikes were detected, aligned and clustered (see Section III-A). The waveforms corresponding to the same class were then averaged for smoothing purposes and the five templates were obtained.

**Fig. 3:** (A) (Left) Colored noise obtained as a steady state response of a first order linear system (time constant 1.3 ms, sampling rate 20 kHz) driven by a WGN of zero mean and unit variance [38]. (Right) The normalized ACVS of the process showing correlation up to several milliseconds, consistent with the time constant of the system. (B), (C), (D) and (E) (Left) Continuous wavelet transform of the signal at scales 2, 4, 6 and 8, respectively. (Right) The ACVS of the corresponding multiscale representations of the signal indicating significant correlation in the vicinity of zero lag only (white noise).

**Fig. 4:** (A) Colored noise generated in the same way as in Fig. 3. (B) White noise. The horizontal lines mark the 3.75 standard deviation bounds.

**Fig. 5:** A segment of a spike train showing two transients. The spikes (red) with known occurrence times are modeled as two semi-cycles of a sine wave with different amplitudes. For simplicity, the signal is corrupted by a zero-mean WGN and the resulting signal is shown in blue. The subsequent plots show the wavelet coefficients at scales $\{a_0, a_1, a_2, a_3\}$ that support the acceptance of $H_1$ (blue) and the corresponding contiguous regions that define sets $B^{H_1}_j$ (red), as well as the set $B^{H_1}$, obtained by Eq. (8). The tick marks represent the true arrival times of spikes (blue) and their estimates (red).

**Fig. 6:** (A) An extracellularly recorded signal with high SNR. (B) A signal containing no distinguishable spikes - neural noise. (C) Power spectrum of the neural noise. (D) Histogram of the neural noise together with a Gaussian pdf with the same mean and variance as those of the noise samples, indicating that the

7The nomenclature for the scales of continuous wavelet transform used throughout this article is consistent with that of MATLAB® Wavelet Toolbox.
Gaussian assumption is valid.

**Fig. 7:** (A) Spike train at SNR=3.5 (left) and the blow-up of the shaded region (right). The tick marks indicate the locations of the spikes. Note how certain noise samples have amplitudes comparable to the amplitudes of spikes. (B) Equivalent plots for SNR=4.0.

**Fig. 8:** ROC for different FR and SNR scenarios. The horizontal axis of each panel represents the probability of false alarm and the vertical axis represents the probability of detection. Each column represents ROC at fixed mean firing rate: low 10 (Hz), medium 30 (Hz) and high 100 (Hz). Each row represents the performance at fixed SNR: 3.5, 3.6, 4 (bottom right corner of each panel). The detection methods are: WDM (blue), DATM (red), SATM (green), and PDM (magenta). Filled circles correspond to detection with $L = 0$ for WDM and $T_a = 3.6$ for SATM.

**Fig. 9:** Comparison of performances of WDM, SATM, DATM and PDM with the detection parameters set at $L = 0$, $T_s = 3.6$, $T_d = 3.75$ and $T_P = 3$, respectively (see Section III-C for details). (A) Performance of WDM. The arrows show the direction of increase of FR and SNR. The dotted lines represent interpolated performance at fixed SNR and for FR between 10 Hz and 100 Hz. (B) Performance of SATM. (C) Performance of DATM. (D) Performance of PDM.

**Fig. 10:** (A) Features (first two principal components) of false positives of WDM clustered in 4 distinct groups (blue, red, green and cyan). Black circles indicate the features of the five templates. (B) Equivalent plot for SATM, where data tend to fall into a single cluster (blue), with one outlier (red). (C) The waveforms (cluster averages) corresponding to clusters from (A), where blue and cyan clusters account for over 90% of the spikes. (D) Equivalent plot for cluster from (B) and the outlier.

**Figures**
Fig. 5.

Fig. 6.

Fig. 7.
Fig. 8.
Fig. 9.

Fig. 10.