A Supervised Multi-Sensor Matched Filter for the Detection of Extracellular Action Potentials

Agnieszka F. Szymanska1, Michael Doty1, Kathryn V. Scannell2, Zoran Nenadic1,3

Abstract—Multi-sensor extracellular recording takes advantage of several electrode channels to record from multiple neurons at the same time. However, the resulting low signal-to-noise ratio (SNR) combined with biological noise makes signal detection, the first step of any neurophysiological data analysis, difficult. A matched filter was therefore designed to better detect extracellular action potentials (EAPs) from multi-sensor extracellular recordings. The detector was tested on tetrode data from a locust antennal lobe and assessed against three trained analysts. 25 EAPs and noise samples were selected manually from the data and used for training. To reduce complexity, the filter assumed that the underlying noise in the data was spatially white. The detector performed with an average TP and FP rate of 84.62% and 16.63% respectively. This high level of performance indicates the algorithm is suitable for widespread use.

I. INTRODUCTION

Before any signal analysis can be performed, action potentials (APs), representative of neurophysiological activity, must be identified from the data. Detection is therefore a crucial component of any neurophysiological signal processing algorithm. Biological noise in extracellular neurophysiological recordings is composed of the activity of background neurons and ion-channel noise. It is therefore both correlated with, and statistically similar to the AP signal. This compounds the detection problem and makes most standard signal detection tools used in other fields unsuitable.

Template matching, or more generally the matched filter, have proven to effectively detect spikes from various types of neurophysiological data [1]–[5]. The general basis behind template matching is to generate a template representative of neural activity, and then compare the data to the template. This approach can be used with a single template for detection purposes or many different templates for purposes of simultaneous detection and classification [6].

Matched filtering can be either supervised or unsupervised. Traditional template approaches are supervised, using both spike and noise training measurements for various parameter estimation before detection. The template is then usually generated from extracellular action potentials (EAPs) selected by an analyst from the data [3], [6]. Un-supervised algorithms using wavelets as templates, and assuming no prior knowledge about the signal, are also becoming increasingly popular [4], [5], [7], [8].

Although widely applied to single-sensor extracellular data, there are few examples of matched filters applied to multi-sensor data. Gozani and Miller [6] developed a technique using multiple matched filters to simultaneously detect and classify multi-sensor EAPs. Their technique tries to both maximize EAP signal-to-noise ratio (SNR) and minimize the interference between EAPs. However, this method relies on the construction of many EAP templates, which requires some prior knowledge of the number of recorded neurons, and may be very time consuming, as a single tetrode in a neuron-dense region may record activity from up to 20 neurons [9]. Furthermore, minimizing the interference between EAPs may compromise their detectability.

The work presented here concentrates on the generation of a more reliable and simpler supervised matched filter for use on multi-sensor extracellular recordings. The filter takes into account the noise statistic of the data and is designed to use a single template to detect EAPs from multiple different neurons.

II. METHODS

A. Data Collection

Data used in this experiment is publicly available online [10] and the full data collection procedure is described by Pouzat et al. in [11]. Briefly, a planar silicon probe was placed below the surface (~50-100 µm) of an adult locust’s antennal lobe and used for recording. Recorded data was sampled at 15 kHz and bandpass filtered from 300 - 5,000 Hz. A total of 20 seconds of data was provided from four of the probe tip sensors. All subsequent data analysis was performed in MATLAB.

B. Ground-Truth Estimation

Half of the collected data was used for training (training data), and the remaining half was used for further analysis (test data). Three trained analysts independently tagged all of the spikes in the test data.

C. Detection Algorithm Design

In the case of extracellular signals, the two models being investigated represent noise only data, and data containing both EAPs and noise. More formally, EAP detection can be interpreted as a hypothesis testing problem, where under the null hypothesis, $H_0$, the signal contains noise only, and under the alternative hypothesis, $H_1$, the signal contains both an EAP and noise.
Given a $C$-sensor signal of length $N$, where $N$ is the number of samples spanned by an EAP, we can express the hypotheses mathematically as

$$H_0 : x = n$$

$$H_1 : x = s + n$$

where $x \in \mathbb{R}^{C \times N}$ is a multi-sensor signal, $s \in \mathbb{R}^{C \times N}$ is an EAP, and $n \in \mathbb{R}^{C \times N}$ is zero-mean sensor noise. The likelihood ratio test (LRT) for this problem can accordingly be expressed as

$$L(x) = \frac{p(x|H_1)}{p(x|H_0)} ,$$

accept $H_1$ if $L(x) > \gamma$

accept $H_0$ if $L(x) < \gamma$

where $x \in \mathbb{R}^{1 \times CN}$ is the row vector form of the signal matrix $x$, and $\gamma$ is the threshold.

In the case of Gaussian noise the LRT takes on the form of a generalized matched filter (GMF) [12]

$$S(\bar{x}) = \bar{x} \Sigma^{-1} \bar{x}^T ,$$

accept $H_1$ if $S(\bar{x}) > \gamma'$

accept $H_0$ if $S(\bar{x}) < \gamma'$

(1)

where $\bar{x} \in \mathbb{R}^{1 \times CN}$ is the row vector form of the EAP, $\Sigma \in \mathbb{R}^{CN \times CN}$ is the spatio-temporal noise covariance matrix calculated from the row vector form of the sensor noise $n$, and $\gamma'$ is the updated threshold subsuming $\gamma$ and other data-independent constants. It is important to note that the test statistic, $S(\bar{x})$, is linearly dependent on the signal $x$. Even if the noise is not Gaussian, $S(\bar{x})$ still has the highest achievable SNR of all other linear combinations of the data. In this case, however, the GMF could underperform the more general LRT.

D. Parameter Estimation

Twenty five EAPs, 2 ms each ($N = 30, \ C = 4$), were manually selected from the training data, aligned to their peak values, and averaged to generate the matched template, $s$. Similarly, 25 noise samples, roughly 30 ms each, were manually selected from the training data, concatenated into a single four-sensor time series, $n_T$, and used to generate the noise covariance matrix, $\Sigma$.

To simplify the noise covariance calculation we can represent $\Sigma$ as

$$\Sigma = \begin{bmatrix}
\Sigma_{1,1} & \Sigma_{1,2} & \cdots & \Sigma_{1,C} \\
\Sigma_{2,1} & \Sigma_{2,2} & \cdots & \Sigma_{2,C} \\
\vdots & \vdots & \ddots & \vdots \\
\Sigma_{C,1} & \Sigma_{C,2} & \cdots & \Sigma_{C,C}
\end{bmatrix} \quad (2)$$

where the submatrix $\Sigma_{i,i} \in \mathbb{R}^{N \times N}$ is the temporal covariance matrix of noise at sensor $i$, and $\Sigma_{i,j} \in \mathbb{R}^{N \times N}$ ($j \neq i$) is the temporal cross-covariance matrix of noise at sensors $i$ and $j$. Estimating the full covariance matrices may require a prohibitively large noise sample, however, the calculation becomes more tractable if we assume the noise is uncorrelated in space. This forces all $\Sigma_{i,j}$ ($j \neq i$) to 0. Although neurophysiological noise exhibits spatial correlations [13], our experience shows that the spatially white noise assumption outperforms fully colored noise in detection.

The training noise $n_T$ was subdivided into 375 noise windows ($N = 30$). Auto-covariance sequences, $r_i(k)$, were then calculated at lags $k \in [-29, 29]$ for each window and each sensor $i \in [1, 4]$. The sequences were averaged across all 375 noise windows, and used to generate each $\Sigma_i$.

E. Threshold Setting

Threshold values were determined using the median standard deviation, $\sigma_M$, of the filtered noise training data, $S(n_T)$,

$$\sigma_M = M \left\{ | S(1) \rangle - M, \ldots, | S(T) - M \rangle \right\} \frac{1}{\sqrt{2Erff^{-1}(M)}} \quad (3)$$

where $S(n_T)$ was calculated using (1), $M$ is the median operator, $M$ is the median of the filtered noise training data $S(n_T)$, and $T$ is the number of samples spanning $n_T$ [8]. The threshold is then set as a multiple, $a$, of median standard deviations above $M$

$$\gamma' = M + a\sigma_M$$

$a$ will be referred to as the threshold multiple for the remainder of this article.

The median was used for this calculation because it is less affected by spikes that may be accidentally present in the training noise than a mean. $M$ and $\sigma_M$ are therefore robust estimates of the true noise mean and standard deviation.

F. Performance Analysis

Detected EAPs, given 40 incrementally increasing thresholds, $a = [1, 1.5, \ldots, 20.5]$, were compared against those tagged by each trained analyst, with the analyst acting as the ground-truth. True positive (TP), and false positive (FP) rates for each threshold and analyst were then calculated as

$$\text{TP Rate} = \frac{TP}{Spikes_A} \quad (5)$$

$$\text{FP Rate} = \frac{FP}{Spikes_D} \quad (6)$$

where $Spikes_A$ is the total number of spikes tagged by the analyst, and $Spikes_D$ is the total number of spikes detected by the detector, and $TP$ and $FP$ are the total numbers of true positives, and false positives, respectively. The false negative (FN) rate can be calculated as $(1 - TP)$, however as it gives no new information about the detector performance, it is not reported here. The TP and FP rates at each threshold were then used to generate ROC curves for each analyst.

III. RESULTS

The detection algorithm performed well compared with all three analysts. The optimal threshold for each analyst (Table I) was determined by minimizing the distance between their respective ROC curve (Fig. 1) and theoretically perfect performance (100% TP, 0% FP). At the optimal threshold, the detector performed with a TP rate of 90.79% and a FP rate of 20.66% compared with Analyst 1, a TP rate of 80.00% and a FP rate of 10.31% compared with Analyst 2, and a
TP rate of 83.06% and a FP rate of 18.91% compared with Analyst 3 (Table I). The detector’s average TP and FP rates were 84.62% and 16.63%, respectively.

Note that the optimal thresholds associated with the best performance for each analyst differed. This is a result of the discrepancy in spike tagging between analysts, which can be quantified in terms of signal-to-noise ratio (SNR). The SNR for each spike, \( u \), tagged by the analysts was calculated as

\[
\text{SNR}(u) = \max_i \left\{ \frac{\sigma^2_i(u)}{\sigma^2_i(n)} \right\}
\]

(7)

where \( \sigma^2_i(u) \) is the variance of spike \( u \) at sensor \( i \), and \( \sigma^2_i(n) \) is the variance of the training noise \( n \) at sensor \( i \). The maximum SNR across sensors was taken as \( u \)’s SNR.

Analyst 1 was the most selective with a median SNR of 28, followed by Analyst 2 with a median SNR of 20, and Analyst 3 was the most liberal, with a median SNR of 7. An example of the types of spikes tagged by each analyst, as well as the detected spikes at \( a = 5.5 \) is shown in Fig. 2.

### TABLE I

<table>
<thead>
<tr>
<th>Analyst</th>
<th>Optimal Performance</th>
<th>Threshold Multiple ( 'a' )</th>
<th>Median Analyst SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90.79</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>80.00</td>
<td>5.5</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>83.06</td>
<td>3.5</td>
<td>7</td>
</tr>
</tbody>
</table>

### IV. DISCUSSION

#### A. Analyst Reliability

The data used in this study was collected in vivo, therefore the ground-truth in terms of spike times was not available and was instead estimated as the spikes selected by each analyst. However, agreement between analysts was as low as 39%. The variability in the analysts’ tagged spikes therefore implies that human analysts may not be a reliable source of ground-truth estimation.

Besides not being consistent with each other, the analysts are also not internally consistent in spike selection. Analysts are likely to arbitrarily increase or decrease their internal thresholding criteria over time, even while tagging spikes in a single data set. Similarly, spikes with a relatively low amplitude but surrounded by noise are very likely to be selected by the analyst. However, similar spikes surrounded by other APs are less likely to be selected.

This bias artificially decreases the detector’s TP rate. In general human analysts are not a reliable or consistent source of ground-truth for the detection of neurophysiological activity. This conclusion was also drawn by Harris et al. [14] where 9 analysts sorting extracellular tetrode data had FN and FP rates as high as 30%.

---

**Fig. 1.** ROC Curves for all three analysts. TP and FP pairs were collected for 40 incrementally increasing detection thresholds with \( a = [1.0, 1.5, \ldots, 20.5] \), for each analyst. The black arrows point to the detectors optimal performance for each analyst, with the appropriate parameters listed below.
B. Detector Performance

The presented algorithm’s performance is under estimated in this study. As discussed previously, this is partially due to the analysts’ unreliability. To try to better assess the detector’s performance, the FPs in the optimal cases for Analyst 1 were further inspected. The detector performed optimally when compared against Analyst 1 at $a = 7$. In this case the detector reported a 20.66% FP rate (Table I). Because Analyst 1 is the most selective, we can compare these FPs to the spikes tagged by both Analysts 2 and 3. The comparison results show that 99% of these FPs were tagged by both Analysts 2 and 3. In other words, a majority of the analysts thought that 99% of the spikes identified as FPs in this case were actually TPs. This implies that the FP rate should be closer to 0%. Similar results were derived when analyzing the FPs identified against Analyst 2.

Given these results, we believe that the detector would perform much better if compared against the actual ground-truth as opposed to trained analysts. This can be accomplished by patch-clamping several neurons and simultaneously recording extracellularly with a tetrode as described in [14]. The patch-clamp recordings would be temporally matched against tetrode recordings to determine precise AP times. These could then be used as the ground-truth for detection results. Any spikes recorded from non-patched neurons would not be included in such an analysis.

C. Threshold Setting

When applying this matched filter, the user can adjust the threshold multiple $a$ to achieve their desired sensitivity. However, this threshold may vary drastically from user to user and an automated threshold may therefore be more appropriate. Assuming APs are outliers within a Gaussian noise distribution, we can define an outlier threshold as

$$
\gamma' = \sigma_M \sqrt{2 \ln l}
$$

where $l$ is the number of time samples spanning the full data series, and $\sigma_M$ is derived in (3) [8]. This threshold is essentially an upper bound on the noise distribution, therefore anything above it should belong to a different distribution, i.e. the signal. For the data used here this gives $\gamma' = 4.88$ and is roughly equivalent to $a = 6.0$, falling somewhere between the sensitivity levels of Analysts 1 and 2.

V. CONCLUSIONS

The multi-sensor matched filter developed in this study was assessed against three trained analysts and performed with average TP and FP rates of 84.62% and 16.63%, respectively. The detector’s performance presents it as a great candidate for wide-spread use as one of the only multi-sensor detectors of its kind. Furthermore, the algorithm’s performance is likely under estimated as a vast majority of the spikes identified as FPs for Analysts 1 and 2 were actually identified as TPs by Analyst 3. Because the analysts were proven to be inconsistent and unreliable, we believe that the detector would perform even better if compared against the ground-truth instead of trained analysts.

REFERENCES