Objective evaluation of stimulation artefact removal techniques in the context of neural spike sorting

Maarten Schelles[1], Jasper Wouters[2], Boateng Asamoah[3] Myles Mc Laughlin[3] and Alexander Bertrand[2,4]

Abstract—

Objective **We present a framework to objectively test and compare stimulation artefact removal techniques in the context of neural spike sorting.**

Approach **To this end, we used realistic hybrid ground-truth spiking data, with superimposed artefacts from *in vivo* recordings. We used the framework to evaluate and compare several techniques: blanking, template subtraction by averaging, linear regression, and a multi-channel Wiener filter (MWF).**

Main results **Our study demonstrates that blanking and template subtraction result in a poorer spike sorting performance than linear regression and MWF, while the latter two perform similarly. Finally, to validate the conclusions found from the hybrid evaluation framework, we also performed a qualitative analysis on *in vivo* recordings without artificial manipulations.**

Significance **Our framework allows direct quantification of the impact of the residual artefact on the spike sorting accuracy, thereby allowing for a more objective and more relevant comparison compared to indirect signal quality metrics that are estimated from the signal statistics. Furthermore, the availability of a ground truth in the form of single-unit spiking activity also facilitates a better estimation of such signal quality metrics.**

Key terms **Artefact removal; Neural spike sorting; Linear filter design**

I. Introduction

Simultaneous neural stimulation and recording are required to gather electrophysiological evidence on the effect of neural stimulation on cortical microcircuits [1], [2]. When the stimulation and recording take place in the same brain region, the recording will typically be obscured by large stimulation artefacts [1]. One example where such artefacts are a major nuisance is the study of physiological effects of direct cortical stimulation (DCS), in which stimulation electrodes are placed on the cortical surface to send a small current through the brain tissue [3]. DCS has numerous clinical applications, such as the treatment of neuropathic pain and Parkinson's disease, and the recovery from stroke [4]. Although behavioural evidence for the effect of DCS has been provided, the exact underlying physiological mechanisms remain unidentified [5]. Collecting such electrophysiological evidence is difficult, as it requires studying the single-unit activity, i.e., spiking patterns of individual neurons, and in particular the differences between inhibitory and excitatory neurons [6], [7], [8], [9].

In order to extract single-unit activity from neural recordings, each neural spike has to be assigned to a putative neuron. This process is generally referred to as 'spike sorting', for which several automatic algorithms and software packages have been developed [10], [11]. However, in the presence of a stimulation artefact, the latter heavily dominates the signal, in which case the spike sorting process fails to assign these waveforms to their putative neurons, thereby failing to identify the spike times of individual neurons [12], [13], [14]. Therefore, an artefact removal step is necessary prior to spike sorting.

Stimulation artefact removal techniques can be divided into three categories [12]. The first category are the artefact prevention techniques (e.g. by performing bipolar recordings [15]). These techniques mitigate the artefacts during the recording itself, but rarely succeed in fully preventing them [12]. The second category are techniques for front-end artefact immunity. This category is situated on the hardware side, and mostly contains techniques to ensure the artefacts are recorded undistorted and thus usable for later data processing. The third category are back-end signal processing techniques, which can be implemented either offline or online. In this paper, we focus on (the evaluation of) such back-end signal processing techniques for stimulation artefact removal.

Current state-of-the-art stimulation artefact removal techniques may largely succeed in removing artefacts and recognizing underlying neural data, but lack a suitable framework to quantify how much of the underlying neural data can exactly be retrieved, in particular in a context of spike sorting. This failure rises from either the absence of ground-truth data at all [16], or the use of incomplete ground-truth data, e.g. only ground-truth data for the stimulation artefact itself, based on simulated artefacts, but no ground-truth data for the neural information [13]. With these types of quantification, the effect of artefact removal on the spike sorting accuracy remains unclear and subjective.

In this paper, we propose a framework to objectively evaluate stimulation artefact removal techniques in a context of neural spike sorting. The strength of our framework lies in the use of *in vivo* recorded neural data and stimulation artefacts, which are used to create so-called hybrid ground truth data. The term *hybrid* here refers to the fact that the evaluation data are artificially constructed by manipulating real *in vivo* recordings such that the underlying ground truth is known (both for the underlying spiking data as well as the

[1] KU Leuven, Electrical Engineering Dept. (ESAT), MNS Micro- and Nanosystems, Leuven, Belgium

[2] KU Leuven, Electrical Engineering Dept. (ESAT), Stadius Center for Dynamical Systems, Signal Processing, and Data Analytics, Leuven, Belgium

[3] Exp ORL, Department of Neurosciences, The Leuven Brain Institute, KU Leuven, B-3000, Leuven, Belgium

[4] Leuven.AI - KU Leuven Institute for AI, B-3000, Leuven, Belgium

artefact component). The ground-truth spiking data allow an accurate and direct quantification of the effect of artefact removal on the spike sorting, and therefore an objective evaluation between different state-of-the-art artefact removal techniques in terms of the obtained spike sorting accuracy and for various signal-to-noise ratio (SNR) levels. Moreover, the artefact ground truth can be used to investigate and quantify the suppression of the artefact.

We investigated four different artefact removal techniques:

1) *Blanking* is a frequently used method, as it is easy to implement, also in hardware [13].
2) *Template subtraction*, for the same reasons as blanking [17].
3) *Linear regression* is currently seen as the state-of-the-art in artefact removal, both offline and online [14], [15].
4) Lastly, the *multi-channel Wiener filter* (MWF) was also evaluated, which is a rather recent, but very promising artefact removal method borrowed from the field of electroencephalography (EEG) signal processing [18], [19].

All investigated techniques are semi-supervised, meaning they incorporate extra information in addition to the recording. These typically perform better than fully unsupervised (blind) techniques, like independent component analysis (ICA) [18]. Here, the extra information that is assumed to be available are the timestamps that denote when each artefact starts and ends. Since the stimulation pulses are generated by the system, such information is typically readily available.

The outline of this paper is as follows. In section II, we discuss the origin of the ground-truth data, and how the objective evaluation framework is constructed. In section III, we briefly review the four artefact removal methods and their relevant properties. Section IV shows the quantitative results of our objective evaluation framework, based on hybrid ground-truth data, as well as a qualitative evaluation on our own *in vivo* recordings. Finally, in section V we briefly discuss the results and the advantages and limitations of our evaluation framework.

II. Hybrid-data benchmarking framework

The main contribution of this study is the use of hybrid ground-truth spiking data in order to facilitate an objective evaluation of artefact removal methods. To this end, we started from an actual neural recording without stimulation and performed an automatic (but manually curated) spike sorting. From these spike sorting results, we generated hybrid ground-truth data using the method in [20]. Lastly, we linearly superimposed actual artefacts measured during *in vivo* experiments on top of these hybrid neural recordings, in which the signal-to-noise ratio (SNR) could be varied. In the remainder of this section, we provide the details on how we obtained this hybrid benchmark data set. An overview of our evaluation procedure can be found in figure 1.

A. Neural recordings

To evaluate the impact of stimulation artefact (removal), we employ two different neural recordings, originating from [21]. Note that these recordings do not include stimulation, hence they are artefact-free. The signals were recorded *in vivo* with two completely different probes, with 32, respectively 128, channels. This demonstrates that the framework can be applied regardless of the probe or number of electrodes that are being used. The neural recordings are then processed by the SpyKING CIRCUS spike sorting package [22], in combination with the SpikeInterface framework [23], which results in information about the spike times of individual neurons, also referred to as single-unit spike trains, after manual curation [24]. The first recording is 9 minutes and 17 seconds long and contains 8 high-quality neuron clusters, for a total of 14329 spikes. The second recording is 5 minutes long and contains 7 high-quality neuron clusters, for a total of 4525 spikes.

B. Generation of hybrid ground truth data

As previously mentioned, the use of an accurate and realistic ground truth spiking data is the basis for a reliable and objective evaluation of different stimulation artefact removal techniques. These ground-truth data must thus resemble real neural data as much as possible. In figure 2, we explain the spike hybridization procedure that is used in this paper to make such ground-truth data based on *in vivo* recordings of spikes and artefacts. The basic idea behind spike hybridization is to first capture the spikes from a spike-sorted single-unit cluster and to remove them from the recording [20]. This spike train is then shifted in time and injected at another location on the probe. This location shift is crucial to get reliable ground-truth recordings. Indeed, spikes at the original location that would not have been identified during the initial sorting, could be identified during a second spike sorting process, but would then be classified as false positives, meaning they are not recognized as correct spikes, while they actually are. The relocation of the spike train, and thus of the neuron, resolves this issue as it prevents the attribution of potential leftover spikes to the cluster of interest, while preserving probe-dependent spatio-temporal characteristics of the neural spiking data. This relocation process also preserves spike amplitude variations within a single-unit cluster [20].

We used the open-source software tool SHYBRID [20] to generate the hybrid ground-truth spiking data from the neural recording. The result was a new recording, with the spikes of the original recording, but on different locations on the probe and at different times. The full recordings are used, including all the original spike templates and their relative timings, to create a new, ground-truth, recording. This has the advantage that it is more realistic than ground-truth recordings that rely on artificial spike models or simulations [20]. The recording is accompanied by ground-truth labels, which indicate for each single-unit cluster when there is a spike. As we moved each spike, we could also control its amplitude, and therefore change the SNR of the recording. This is a feature we used to

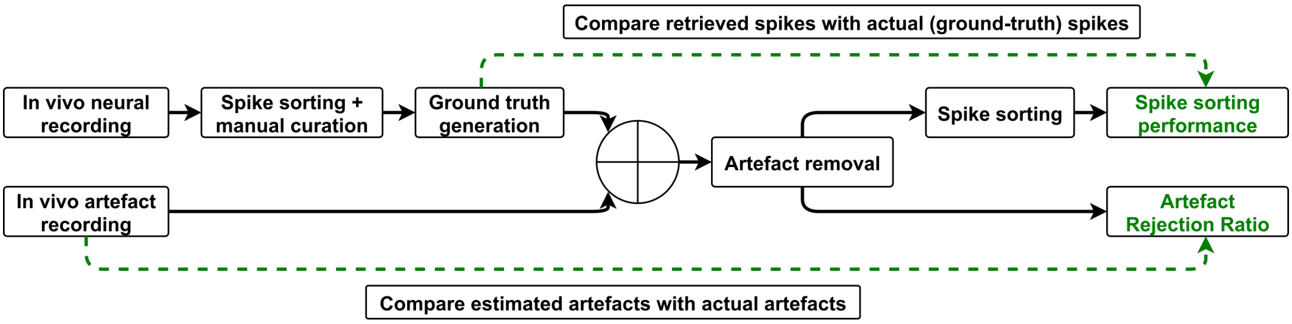

Compare retrieved spikes with actual (ground-truth) spikes
In vivo neural recording
Spike sorting + manual curation
Ground truth generation
Artefact removal
Spike sorting
Spike sorting performance
In vivo artefact recording
Artefact Rejection Ratio
Compare estimated artefacts with actual artefacts

Fig. 1: Flowchart of the used framework. The two green boxes at the end show the used performance metrics for a quantitative evaluation of the artefact removal methods.

assess the effect of artefact removal in low-SNR conditions (see Subsection IV-C).

As mentioned before, all 15 different neuron spike trains (each with a different spatio-temporal signature spike waveform) are included in the framework and the analysis of the different artefact removal techniques. The results are always averaged over all spike trains from a recording. Moreover, the SHYBRID software also includes realistic spike overlaps, leading to a realistic neural ground truth.

C. *Artefact model*

The next step is to superimpose stimulation artefacts onto the spiking data. In the 'hybrid' philosophy, we do not use a simulated artefact, but instead extract it from a real recording. All procedures were approved by the KU Leuven animal ethics committee for laboratory experiments under project number P201/2018. The stimulation pulses were recorded from an *in vivo* experiment with DCS, using a similar procedure as in [3]. Here, a 32-channel silicon probe from Cambridge NeuroTech (A554-37H6b-sharp) [25] was inserted into a burr-hole craniotomy which was drilled above the motor cortex of a rat. The reference for this recording probe was a metal wire placed on the brain, in the proximity of the recording probe. Two stimulation electrodes were placed on the cortex approximately 2 mm from the craniotomy; one in the anterior and the other in the posterior direction. The electrodes were attached to a current source stimulator (AM 2200 Analog) which received a voltage input from an NI-card (NI USB-6216) which received a waveform signal generated in a custom Matlab 2014a software. The signal consisted of biphasic pulses with phase widths of 200 μs. All techniques considered in this paper are linear techniques, which assume a linear superposition of the undistorted artefacts on top of the spikes. Therefore, it is important that the artefacts are not clipped against the maximum amplitude of the analog-to-digital converter during the recording.

In this work, we used a continuous stimulation, with a frequency of 800 Hz, which is a clinically relevant frequency in DCS [5]. There was a total artefact length, per pulse, of about 700 μs: 2 biphasic pulses with 200 μs per phase and a small artefact tail after the stimulation stopped, shown in figure 3. With a continuous stimulation of 800 Hz almost 60% of the total signal length was corrupted by artefacts. The artefact recordings were high-pass filtered with a cut-off frequency of 300 Hz (same filter as for the neural recording), and then added to the hybrid data from subsection II-B. As the artefacts were recorded with a different probe than the neural data (and thus of the hybrid ground-truth data), the following mapping was used. The top channel of the probe with which we recorded the artefacts corresponded to the top channel of the hybrid probe, and so on, going downwards on the probe. Channels on the hybrid probe whose location fell in between two electrodes on the artefact probe, were interpolated linearly based on the distance to these two electrodes. Finally, artefacts on two different electrodes at the same depth of the hybrid probe would have exactly the same artefacts according to this linear scaling, so a very small amount of Gaussian noise was introduced. This noise was in the order of magnitude of the variations between two original adjacent channels for the artefacts. This was done to ensure that all hybrid artefacts were a little bit different, as this could otherwise favour certain artefact removal methods.

The timestamps to denote the start and end of an artefact (the semi-supervised information for our techniques) were provided by the data acquisition system, coupled to the stimulator system. The observed amplitude of the artefacts is about one order of magnitude larger than the spikes (as later also shown in figure 6). If the timestamps would not be available in another experimental context, this property can be used to design a simple threshold operation to detect the onsets of the artefacts [26].

D. *Performance metrics*

Using the spiking ground-truth data, we can define three performance metrics to accurately capture the performance of a certain spike sorting process: the precision, the sensitivity, and the F1-score. The spike times of all spikes in a retrieved cluster, are compared to the true spike times of the corresponding neuron. The *precision* is defined as the ratio of correctly identified spikes (true positives) over all identified spikes within a particular cluster (true positives and false positives). The *sensitivity*, also called sensitivity, is defined as the ratio of the correctly identified spikes (true positives) over all ground-truth spikes of the corresponding neuron. Finally, the *F1-score* captures the two previous metrics into one: $F1 = 2 \times \frac{precision \times sensitivity}{precision + sensitivity}$. These three are able to capture sufficient information on the spike sorting performance, and give very detailed information on single

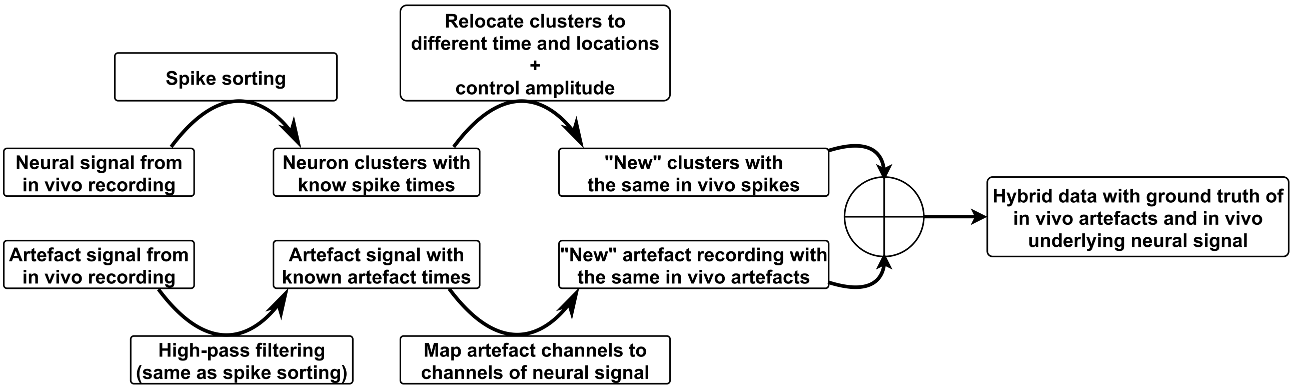

Spike sorting
Relocate clusters to different time and locations + control amplitude
Neural signal from in vivo recording
Neuron clusters with know spike times
"New" clusters with the same in vivo spikes
Hybrid data with ground truth of in vivo artefacts and in vivo underlying neural signal
Artefact signal from in vivo recording
Artefact signal with known artefact times
"New" artefact recording with the same in vivo artefacts
High-pass filtering (same as spike sorting)
Map artefact channels to channels of neural signal

Fig. 2: Detailed schematic description of the generation of hybrid ground-truth data.

clusters [27].

Apart from the spike sorting performance, we also used another performance metric solely based on the artefacts: the Artefact-to-Residue ratio (ARR), or artefact-rejection ratio. This measures how much the artefact is suppressed on each channel. The ARR at channel k is defined as [19]:

$$ARR_k = 10\log_{10}\frac{E\Big\{a_k^2[t]\Big\}\Big|_{t\in\mathcal{A}}}{E\Big\{(a_k[t]-\hat{a}_k[t])^2\Big\}\Big|_{t\in\mathcal{A}}} \tag{1}$$

where $\mathcal{A}$ is the set of time samples in which the artefact is present, $E\{.\}|_{t\in\mathcal{A}}$ denotes the expectation operator over the samples in A, $a_k[t]$ is the ground truth artefact signal, and $\hat{a}_k[t]$ is the estimated artefact signal that is subtracted by the artefact removal method.

The total ARR of the signal is the weighted average of the ARRs of the channels: $ARR = \sum_{k=1}^{K} p_k \times ARR_k$, with the normalized weights p_k defined as the proportion of the artefact power of that channel, over the total artefact power. The artefact power of a channel can be estimated by subtracting the signal power during a clean period from the signal power during an artefact.

$$p_k = \frac{E\Big\{x_k^2[t]\Big\}\Big|_{t\in\mathcal{A}} - E\Big\{x_k^2[t]\Big\}\Big|_{t\notin\mathcal{A}}}{\sum_{k=1}^{K}\Big(E\Big\{x_k^2[t]\Big\}\Big|_{t\in\mathcal{A}} - E\Big\{x_k^2[t]\Big\}\Big|_{t\notin\mathcal{A}}\Big)} \tag{2}$$

The whole framework, including the performance metrics and how they are calculated based on ground-truth data, is schematically summarized in figure 1.

III. Evaluated artefact removal methods

In this section, we briefly review the four stimulation artefact removal methods that are evaluated in this paper.

A. Blanking

Blanking is the most straightforward artefact removal technique, as it doesn't assume an underlying model on the artefacts or the neural signal. It only requires the knowledge of the artefact timestamps, which are known in the experimental context of interest for this work (but can also be easily detected in another experimental context [26]).

For each artefact, a linear interpolation is performed between the last sample before, and the first sample after the artefact. By replacing the artefact samples instead of just cutting them out of the signal, the timing information is preserved, and so is the true time distance between consecutive spikes. This is important for the spike sorting, where the refractory period plays a role in determining whether a waveform is indeed a true spike.

B. Template subtraction (averaging)

This technique aims to estimate a template of the recorded artefact pulse waveform, by averaging the time-aligned artefact pulses across the recording [13], [17], [26], [28], [29]. In this template, the neural signal will be averaged out. Two different methods are considered to construct such an average template. In the first method, we average over all artefacts per channel, leading to 1 template per channel. In the second method, we average over all channels per artefact pulse, leading to 1 template per stimulation pulse (event). The first method thus assumes that the artefact waveform is approximately the same across the time axis (but can be different across channels), whereas the second assumes the artefact waveform is approximately equal across all channels (but can change over time).

C. Multi-channel filtering

Instead of simple spatial averaging, where the same weight is given to each channel, multi-channel filtering techniques aim to construct an optimal linear filter for the artefact removal, using signals from other channels. Two specific techniques are described and investigated: the linear regression filter which directly uses the other channels (but leaving out the adjacent ones to preserve neural information) and the MWF, which is the stochastic variant, where all channels are combined based on their underlying second-order statistics. We only give a high-level description here to highlight the non-standard features relevant for the interpretation of the results, and we refer to appendix A for a more detailed mathematical description of the method.

1) Linear regression: Linear regression assumes that the relationship between the stimulation artefacts at different channels can be described by linear time-invariant filters. Figure 3 clearly shows that artefacts among all channels are indeed very similar over the entire range of the probe, even on channels several 100 μm away. On the other hand, the spatial similarity of the neural signal is much smaller. The amplitude of spikes quickly decreases for channels further

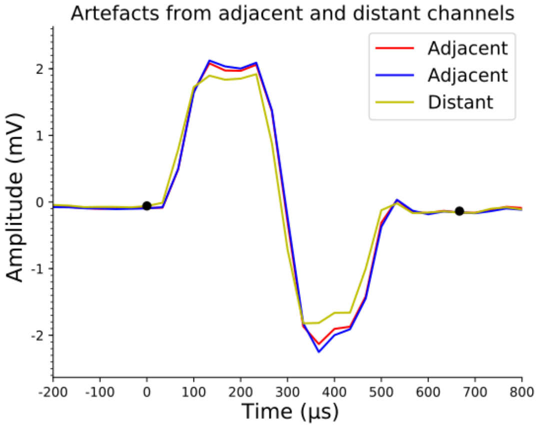


Fig. 3: Artefacts from the same stimulation pulse, but from different channels. The black dots mark the start and end of the artefact, as given by the acquisition system. Waveforms of two adjacent channels (channels 1 and 2, 25 μm apart) are very similar, and even show a great similarity to waveforms from distant channels (channel 32, 400 μm apart). The difference in amplitude of the artefacts is caused by a different depth of the channels on the probe.

away from the neuron. Thus, while adjacent channels will show similar artefacts and similar spike waveforms, channels about 100 μm further away, will still show similar artefacts, but no more similarity in spikes. This observation is the basis for the linear regression, where the artefact component of each channel is estimated from the recorded data at non-adjacent channels [14].

When constructing an optimal spatio-temporal filter $\boldsymbol{w}_k$ for channel k, we only take channels that are sufficiently far away from this channel of interest into account, in an attempt to preserve the neural information, an approach first explored on a depth probe with 1 column in [16]. To this end, we define the vector $\bar{\boldsymbol{x}}_{-k}[t]$ as the vector stacking the L most recent samples (up to the current sample time t) from all channels, except those that are located within a distance ϵ from channel k on the probe (including channel k itself).

The spatio-temporal filter $\boldsymbol{w}_k$ which estimates the artefact component $\hat{a}_k[t] = \boldsymbol{w}_k^\intercal \bar{\boldsymbol{x}}_{-k}[t]$ at channel k by filtering the data in $\bar{\boldsymbol{x}}_{-k}[t]$, is optimized to estimate the artefact signal in channel k in minimum mean squared error (MMSE-)sense, by using the far-away channels as reference signals, i.e.,

$$\hat{\boldsymbol{w}}_k = (\boldsymbol{R}_{x_{-k}x_{-k}} + \lambda \boldsymbol{I})^{-1} \boldsymbol{r}_{x_{-k}x_k}. \tag{3}$$

Here, $\boldsymbol{R}_{x_{-k}x_{-k}}$ and $\boldsymbol{r}_{x_{-k}x_k}$ respectively represent the spatio-temporal covariance matrix and cross-correlation vector, where a diagonal loading is added to the former in order to perform Ridge regression with L2-norm regularization. The estimated artefact signal $\hat{a}_k[t] = \hat{\boldsymbol{w}}_k^\intercal \bar{\boldsymbol{x}}_{-k}[t]$ can then be subtracted from the k-th channel $x_k[t]$. We refer to appendix A for further details.

Note that the distance ϵ to determine which channels to take into account should be non-zero. Indeed, in the trivial case of using the channel of interest to construct an optimal template for its own, the filter would only use that channel, as it would result in a perfect representation. If the distance is chosen to be almost zero, all other channels are taken into account. If the distance is chosen a bit larger than the inter-electrode distance, only non-adjacent channels are taken into account, etc. The larger this distance is, the less accurate the template estimate will become, yet the lower the risk of removing spikes in channel k due to leakage of correlated neural activity. The goal is to determine an optimal distance ϵ such that the template only contains the true artefact and no neural information. In our study, we have determined the optimal ϵ by using the F1-score as performance metric.

Important to note is that this method only applies linear regression between the different recorded channels. Another kind of linear regression could be to map the input stimulation block pulses onto the recorded artefacts. However, the mapping of the stimulation pulse through the stimulation electrode, brain tissue, and recording electrode, is highly non-linear, and therefore difficult to model with a linear filter [13], [30].

2) Multi-channel Wiener Filter: The MWF is an existing artefact removal method for EEG, where it has been successfully applied to remove eye blinking and muscle artefacts [18], [19]. To our best knowledge, this technique has not been used for stimulation artefact removal in the context of neural spike sorting. Let $\boldsymbol{x}[t] = \boldsymbol{a}[t] + \boldsymbol{n}[t]$ denote the recorded multi-channel signal containing and artefact component $\boldsymbol{a}[t]$ and a neural component $\boldsymbol{n}[t]$. The MWF exploits the differences in the spatio-temporal statistics of the unwanted signal component (the artefact) and the wanted signal component (the neural signal) to suppress this unwanted component as much as possible within the recorded signal $\boldsymbol{x}[t]$ [19]. Similar to the previous method, the MWF takes the recorded multi-channel probe signal as an input, and produces at its output an estimate of the artefact component at each channel, which can then be subtracted from the signal. However, the MWF uses all channels, including the target channel and its neighbourhood. Here, we again give a high-level description and refer to appendix A for the mathematical details. We use the MWF framework as proposed in [19].

The MWF is represented by a matrix W of which the k-th column contains the spatio-temporal filter $\boldsymbol{w}_k$ that estimates the artefact signal $a_k[t]$ at channel k. As opposed to the previous method, the MWF computes this matrix in a single shot, i.e., all filters for all channels are computed at once. The optimization criterion for this spatio-temporal linear filter $\boldsymbol{W}$ is again the linear minimum mean squared error (MMSE). The MWF can be computed as:

$$\hat{\boldsymbol{W}} = \boldsymbol{R}_{xx}^{-1} \boldsymbol{R}_{aa} \tag{4}$$

where $\boldsymbol{R}_{xx}$ and $\boldsymbol{R}_{aa}$ represent the spatio-temporal covariance matrices for the signals $\boldsymbol{x}[t]$ and $\boldsymbol{a}[t]$, respectively. Note that the signal $\boldsymbol{a}[t]$ is not directly observable, such that $\boldsymbol{R}_{aa}$ cannot be directly estimated through temporal averaging of the observations. However, when the stimulation is not active, $\boldsymbol{x}[t] = \boldsymbol{n}[t]$, which allows to estimate $\boldsymbol{R}_{nn}$ (i.e. the covariance matrix of $\boldsymbol{n}[t]$) from the recorded data. To retrieve a reliable estimate for $\boldsymbol{R}_{aa}$, and to prevent it to become ill-conditioned or even indefinite, we compute a

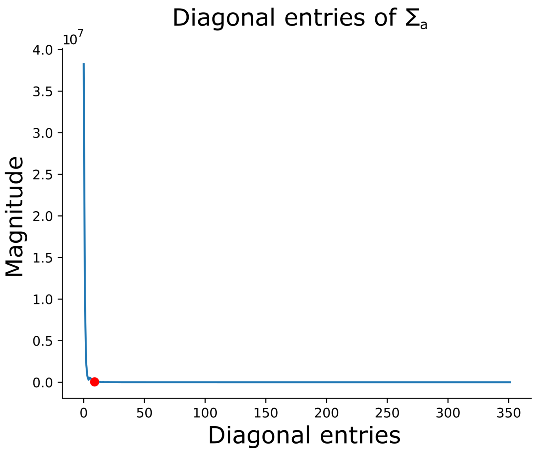


Fig. 4: Magnitude of the diagonal entries of $\mathbf{\Sigma}_a$ of the artefact model for the 32-channel probe. Out of 352, only 9 (see red dot) are kept to build the artefact signal with.

generalized eigenvalue decomposition (GEVD) of the matrix pencil $(\boldsymbol{R}_{xx}, \boldsymbol{R}_{nn})$, with as end result [19]:

$$\begin{aligned} \boldsymbol{R}_{aa} &= \boldsymbol{V}^{-\mathsf{T}} \mathbf{\Sigma}_a \boldsymbol{V}^{-1} \\ with \quad \mathbf{\Sigma}_a &= \mathbf{\Sigma}_x - \mathbf{\Sigma}_n \end{aligned} \tag{5}$$

with $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_n$ both diagonal matrices and $\boldsymbol{V}$ containing the corresponding generalized eigenvectors in its columns. The strength of the GEVD now lies in the manipulation of $\mathbf{\Sigma}_a$, for which two alternatives have been described in [19]:

1) All negative entries of $\mathbf{\Sigma}_a$ are set to zero, leading to a covariance matrix that is certainly (semi)-positive definite.
2) Only the first Q (largest) diagonal entries of $\mathbf{\Sigma}_a$ are retained. This explicitly exploits the low-rank structure of the artefacts.

The largest entries of $\mathbf{\Sigma}_a$ represent the largest part of the artefact power, where the artefact power is defined as the trace of the matrix $\mathbf{\Sigma}_a$, i.e., the cumulative sum of its diagonal entries. The parameter to be optimized is then the power fraction, being the share of all included diagonal entries divided by the total artefact power. In our research, we tested both the normal MWF (without setting any of the diagonal entries of $\mathbf{\Sigma}_a$ to zero, thus a power fraction of 1), and the GEVD-based MWF, where we only preserve a subset of Q diagonal entries of $\mathbf{\Sigma}_a$, thus a power fraction smaller than 1. In this case, only the Q largest entries are used, to enforce the artefact model to be low-rank. In subsection IV-A, we explain how the value of Q was selected in our experiments.

IV. Results

A. Spike sorting performance

Our hybrid ground-truth framework allows us to accurately compare the performance of different artefact removal methods. We will use the previously explained metrics: precision, sensitivity, F1-score, and ARR. For the ease of comparison, the metrics for each removal method will be averaged over all available neurons (clusters), since the performance was found to be largely the same across different neurons for a given method.

TABLE I: Comparison of the different techniques for artefact removal. The spike sorting performance (F1-score, precision, sensitivity) is averaged over all clusters. To evaluate the robustness, the standard error on the mean across all single-unit clusters from that recording is shown between parentheses. As precision and sensitivity can be interchangeable (depending on the threshold during spike sorting), the optimal point is chosen as the point where the (unrounded) F1-score was the highest.

32-channel probe

	F1-score	**Precision**	**Sensitivity**	**ARR (dB)**
Clean recording	1.00 (<0.01)	1.00 (<0.01)	1.00 (<0.01)	/
Blanking	0.77 (0.02)	0.89 (0.03)	0.69 (0.02)	35.01
Channelwise	<0.01 (<0.01)	<0.01 (<0.01)	0.04 (0.01)	9.82
Eventwise	<0.01 (<0.01)	<0.01 (<0.01)	0.32 (0.07)	20.68
Spatial lin. reg.	0.98 (<0.01)	0.98 (<0.01)	0.98 (<0.01)	**36.40**
Lin. reg.	0.98 (<0.01)	0.98 (<0.01)	0.98 (<0.01)	35.12
MWF	0.81 (0.10)	0.83 (0.10)	0.82 (0.08)	31.36
Spatial GEVD-MWF	0.98 (0.01)	0.98 (<0.01)	0.97 (0.02)	26.28
GEVD-MWF	**0.99** (<0.01)	**0.99** (<0.01)	**0.98** (0.01)	34.36

128-channel probe

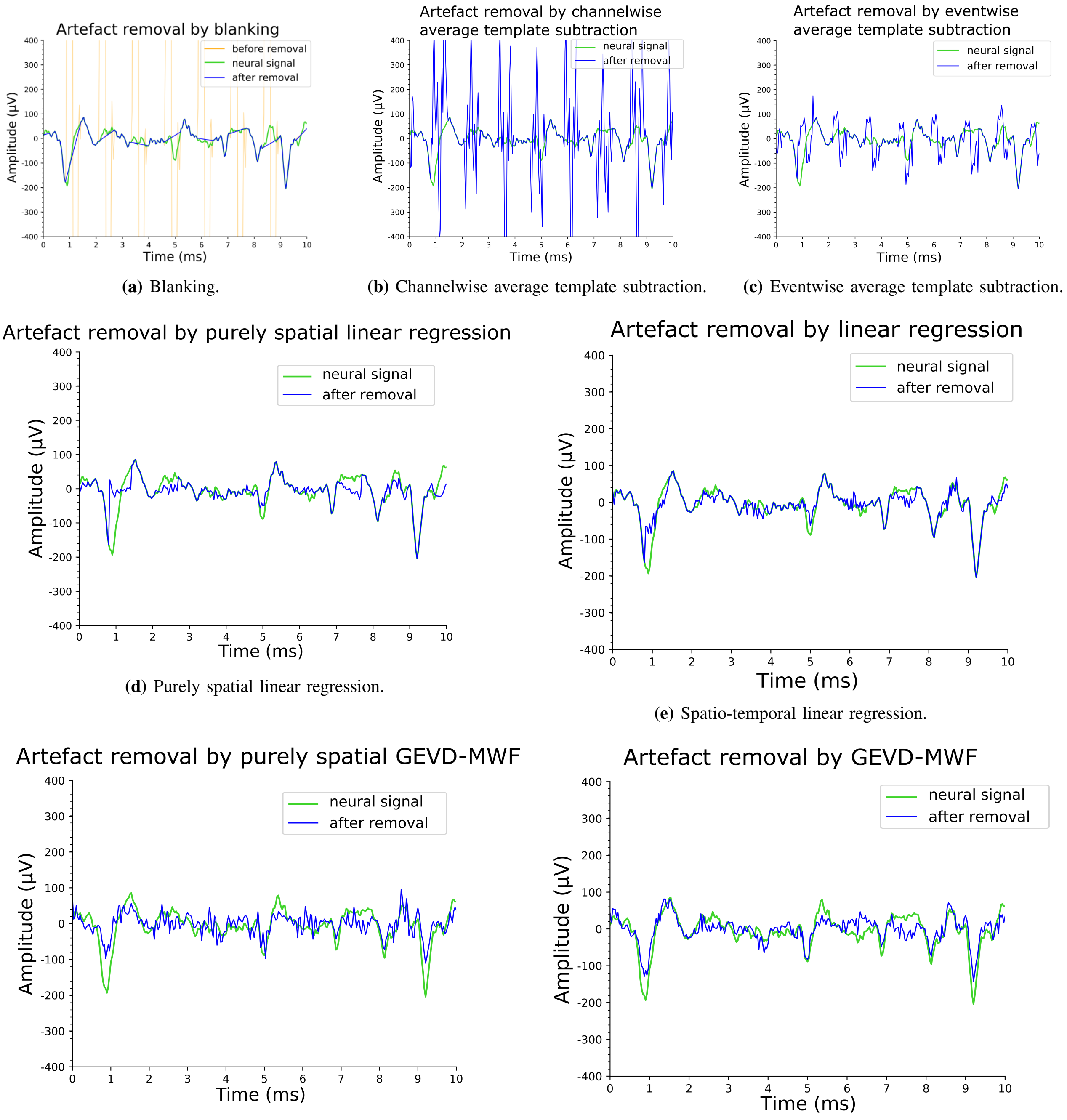
	F1-score	**Precision**	**sensitivity**	**ARR (dB)**
Clean recording	1.00 (<0.01)	1.00 (<0.01)	1.00 (<0.01)	/
Blanking	0.71 (0.05)	0.93 (0.02)	0.59 (0.06)	42.18
Channelwise	0.01 (<0.01)	0.01 (<0.01)	0.04 (<0.01)	17.39
Eventwise	0.32 (0.13)	0.31 (0.14)	0.38 (0.13)	25.00
Spatial lin. reg.	0.97 (<0.01)	0.97 (<0.01)	0.97 (<0.01)	39.68
Lin. reg.	**0.98** (<0.01)	**0.98** (<0.01)	**0.97** (<0.01)	39.59
MWF	0.96 (<0.01)	0.97 (<0.01)	0.95 (0.01)	40.17
Spatial GEVD-MWF	0.97 (<0.01)	0.97 (<0.01)	0.96 (<0.01)	**42.37**
GEVD-MWF	**0.98** (<0.01)	**0.98** (0.01)	0.96 (0.01)	33.89

Table I shows results for the different artefact removal techniques. All parameters were determined in function of the optimal F1-score. The ϵ for the linear regression was tuned in steps of 10 μm and the distance giving the highest F1-score, was determined to be 30 μm in this case . The optimal regularization constant λ was found to be around 0.001 times the maximum entry (in absolute value) of the covariance matrix. The number of time lags L for the spatio-temporal linear regression filter was found to be 7, with only a very small dependency on the number of time lags. For the MWF, the number of time lags was 10. The fraction of the cumulative sum of the diagonal entries in $\mathbf{\Sigma}_a$ was set to 99% for the 32-channel probe, and 99.9% for the 128-channel probe, leading to values Q = 9 and Q = 11, respectively. However, this fraction is not dependent on the number of channels, but rather on the amplitude of the artefacts, which was higher for the 128-channel probe.

As can be seen from table I, for the 32-channel probe, there is a large difference in using the normal MWF, and using the GEVD-MWF. This difference is explained when looking at the diagonal elements of $\mathbf{\Sigma}_a$, as calculated in (25). Indeed, figure 4 shows that only Q = 9 values are actually sufficient to retain 99% of the trace of the matrix. In the 128-channel probe, the difference in spike sorting performance between MWF and GEVD-MWF is less pronounced.

B. Visual assessment

Because of the hybrid framework, we have access to the ground-truth neural signal. We can thus make a visual comparison between the ground-truth signal, and the recovered neural signal after artefact removal. Figure 5 gives a visual assessment of how the different techniques remove



(a) Blanking.

(b) Channelwise average template subtraction.

(c) Eventwise average template subtraction.

(d) Purely spatial linear regression.

(e) Spatio-temporal linear regression.

(f) Purely spatial GEVD-MWF.

(g) Spatio-temporal GEVD-MWF.

Fig. 5: Visual assessment of the different artefact removal techniques, shown over the same period of time for the same recording. The blue signal shows the ground-truth neural data without artefact, the green signal is the recording with artefacts, and the red signal is the recording after artefact removal.

the artefacts and handle spikes. The signals after blanking (figure 5a) and both types of averaging (figures 5b and 5c) clearly lead to a poorer reconstruction of the underlying signal, relative to the multi-channel filtering techniques.

TABLE II: Comparison of the linear regression and a GEVD-MWF for a neural recording with low PSNR (15 dB). In the first three columns, the spike sorting performance (F1-score, precision, sensitivity), Artefact-to-Residue Ratio, and PSNR after artefact removal are averaged over all clusters. The two last columns show the PSNR of the cluster with the lowest, respectively highest, PSNR after artefact removal. The spike sorting performance of the neural signal (without artefacts) is also shown as a reference.

	F1-score	**Prec.**	**Sensitivity**	**ARR**	**Avg. PSNR**	**Min. PSNR**	**Max. PSNR**
Clean recording	0.90	0.94	0.89	/	15.00 dB	15.00 dB	15.00 dB
Lin. reg.	0.58	0.69	0.58	40.45	11.96 dB	10.57 dB	13.35 dB
GEVD-MWF	0.56	0.64	0.59	37.42	11.38 dB	7.75 dB	14.16 dB

When comparing the recovered signal after a purely spatial filter (figures 5d and 5f) with the ones after a spatio-temporal filter (figures 5e and 5g), it can be seen that the largest part of the filtering exists out of spatial filtering, while the temporal filtering only has a small extra effect. This is also in line with the spike sorting performance from table I. In general, when looking at the results from the linear regression filter and the GEVD-MWF (figures 5e and 5g), it is clear that both methods lead to a good reconstruction of the underlying neural signal, thereby still allowing for a reliable spike sorting performance. Nonetheless, we observed that each technique irrevocably leads to a decrease in SNR, compared to the ground-truth signal.

C. *Low-SNR regime*

In the results shown in table I, all spike clusters had a peak-signal-to-noise-ratio (PSNR $= 10\log_{10}\frac{P_{peak}}{P_{noise}}$ [20]) going from 18 dB up to over 30 dB, before adding the stimulation artefacts. Spikes with such a high PSNR, will still keep a sufficiently high PSNR for a reliable spike sorting, even after artefact removal. In reality, there will also be clusters with a lower PSNR. To simulate this situation, we scaled all spikes down to a PSNR of 15 dB (for every cluster) within the SHYBRID software, and performed the same steps as described earlier. Table II shows the spike sorting performance results for the linear regression and the GEVD-MWF.

In the low-SNR regime, the ARR is still high. However, for clusters with a lower PSNR, the inherent decrease in PSNR has an impact on the spike sorting performance. We can also look at the PSNR of the clusters after artefact removal, instead of the pure performance metrics. The average PSNR over all clusters is very similar for linear regression and MWF. However, there is a small inter-cluster difference in PSNR for the linear regression. For the MWF on the other hand, there is a much larger variation between different clusters.

D. *In vivo validation*

To validate the framework used in this paper, and therefore the obtained results, we also performed artefact removal on unmanipulated *in vivo* data. The data are recorded with a protocol similar to [3], with one very long burst of stimulation pulses. Figure 6 shows the signal before and after artefact removal. These signals closely resemble the signals simulated by our hybrid framework. The left plot shows the output of the linear regression and the GEVD-MWF method in overlay. It can be observed that spikes are retrieved, even in the part that is heavily affected by a stimulation artefact. A closer inspection on the zoom box (middle figure) illustrates that the spike which is buried under an artefact pulse is indeed recovered by both methods, while the artefact pulse is mitigated. To compare, the result of the blanking and template matching method is shown in the rightmost plot. Here, a clear artefact residual is visible for the template matching method, which could be wrongly detected as a spike and which dominates the actual spike that coincides with the first artefact pulse. The blanking method removes the spike due to the linear interpolation.

V. Discussion

A. *Study design choices and limitations*

When determining the use and focus of our framework, we made a number of choices that influenced the choice of the evaluated techniques. For example, when electrodes are sparsely divided throughout the brain, neural information is not shared by adjacent electrodes, and other techniques should be used to exploit the shared behaviour of artefacts over electrodes [31]. However, due to advancements in neuroscience and microfabrication technologies, there is a trend towards large-scale, high-density electrode arrays [10], [32]. Therefore, in our study, we have chosen to focus on the context of spike sorting and artefact removal through multi-channel filters for such arrays, where neural information is shared among several electrodes.

Furthermore, we have chosen to compare a subset of commonly used linear techniques. Non-linear techniques for artefact removal have also been investigated in the past, such as artefact estimation based on nonlinear adaptive models with self-oscillations [33]. Furthermore, there have been even studies that used information from the artefact removal in the subsequent spike sorting process [34]. While we have only focused on commonly used linear techniques, in further research, the developed framework with ground-truth data could also be used to test such non-linear artefact removal techniques. It must be noted, however, that the framework in its current form, is only representative in the context of spike sorting. Indeed, the spike sorting is an inherent part to our framework, and different performances might be expected in another context, such as artefact removal for EEG or local field potentials [19], [29], [35]. This also means that our framework requires stimulation frequencies that are not too high (<2kHz). For a higher stimulation frequency (or a higher pulse width with the same frequency), it will be impossible to observe clean neural spikes in between the stimulation pulses, and thus to train the filters such as MWF with this neural information. On the other hand, we don't expect a lower limit on the stimulation frequencies for the framework to remain valid. The reason we have chosen this range of frequencies is therefore purely practical: a lower frequency would show more undistorted neural signal, leading to better a-priori spike sorting performance, such that the performance difference between techniques would be more difficult to see.

Lastly, it must be noted that recent research has also explored cancelling or mitigating the artefacts by dedicated circuit or experiment design [15], [32], or sometimes even succeeds in completely circumventing the necessary stimulation artefact removal. This can be done by using different physiological ways to record and stimulate the brain. Recent examples are calcium imaging with microstimulation [36], electrophysiology together with microLED optoelectrodes [37], or the combination of calcium imaging, optogenetic

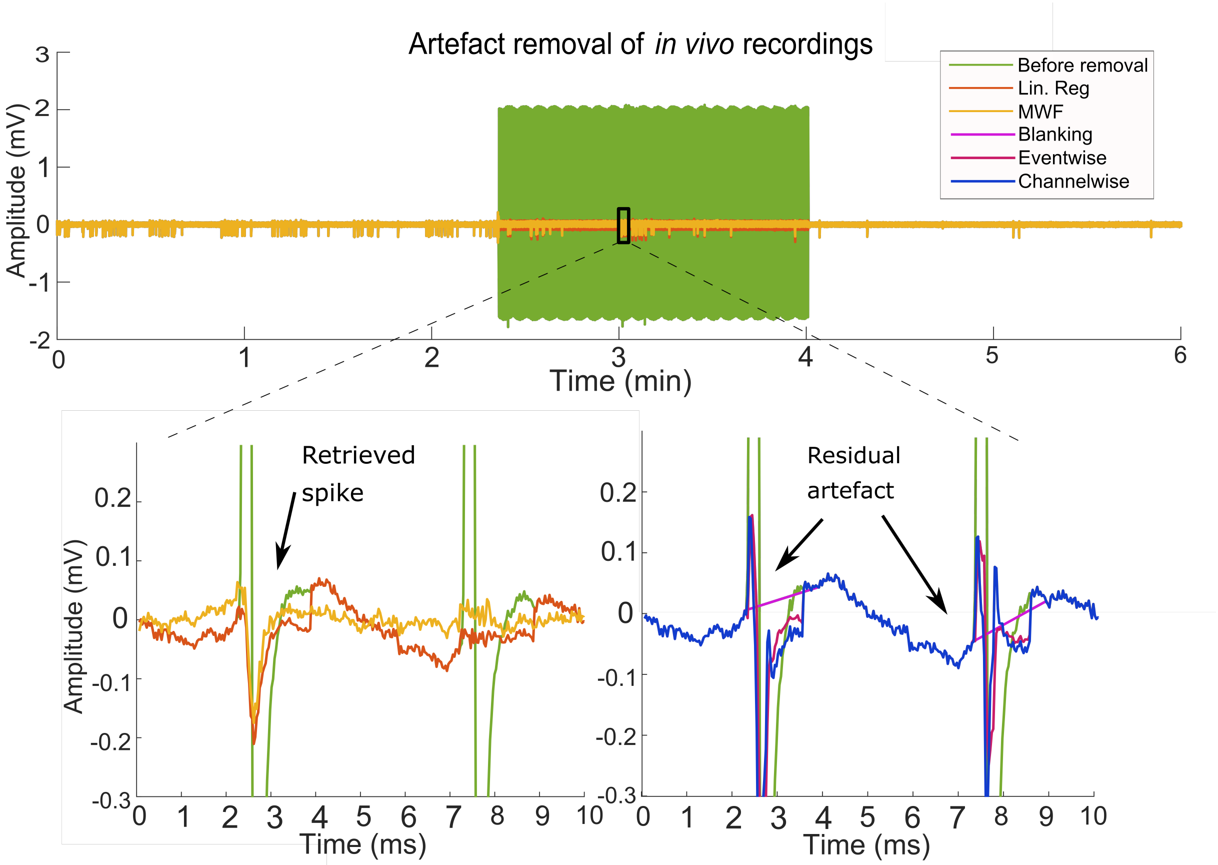


Fig. 6: Top: *In vivo* recording before and after artefact removal, with linear regression and a GEVD-MWF. After artefact removal, spikes can be seen and retrieved throughout the whole recording, also when stimulation takes place. Bottom: a zoom of the black window indicated in the left figure. For comparison, the results with blanking and template subtraction are also shown on the right.

stimulation and electrophysiology [38]. Nonetheless, electrical recording and stimulation remain the most widely used combined strategy to interfere with brain circuits, especially with techniques aiming for clinical translation [1].

All techniques discussed in this paper are linear techniques and thus assume a linear addition of the artefacts and the spikes. The same holds for the proposed hybrid framework, i.e., it assumes that the artefact waveform is linearly superimposed to the neural signal. In practice, this means that the amplitude of the artefacts may not go up to the non-linear saturation regime of the ADC. If that happens, non-linear artefact removal techniques should be used, in combination with an evaluation framework that also models such nonlinearities in the creation of hybrid ground-truth data. Such non-linear cases are beyond the scope of this study, and are also preferably avoided in practice. Indeed, with increasing artefact amplitude, the ADC will eventually saturate, thereby preventing retrieval of the underlying neural information.

B. *Discussion of results*

Based on our proposed framework, we have performed a benchmark study comparing several electrical stimulation artefact removal methods in the context of neural spike sorting. By using hybrid spiking data with a superimposed artefact from an *in vivo* recording, we have access to both the ground-truth spiking data, as well as the ground-truth artefact. This allowed us to perform an objective evaluation, both in terms of spike sorting performance and artefact rejection, while still working with highly realistic and representative data (note that both the spiking activity and artefacts were generated from actual *in vivo* recordings). To the best of our knowledge, this is the first study that introduces such a hybrid framework for objective benchmarking of different state-of-the-art stimulation artefact removal techniques for spike sorting. Previous studies typically estimated their artefact removal and spike sorting performance either directly from the *in vivo* recordings without a ground truth [16] or only used ground-truth data for the artefact itself, typically based on simulated instead of recorded artefact data [13]. Because of a lack of ground-truth data for the neural spiking data, it is impossible to objectively assess the effect of artefact removal on spike sorting performance.

Blanking leads to reasonable results, even though all information during an artefact is thrown away, which accounts for over half of the total signal. However, spikes that are partly cut off, are only partly thrown away. In combination with the fact that linear interpolation is used, the remaining waveform sometimes still looks enough like a spike to be recognized by the spike sorting algorithm. Template subtraction based on both types of simple averaging leads to poor results. Although popular because easy to implement in hardware, it is clear that a template through averaging results in a very poor reconstruction of the artefact. In particular when the template is computed by averaging multiple instances of the artefact pulse across time, subtraction of the template leads to very large residual artefacts (figure 5b). Purely spatial averaging (also called Common Average Referencing) leads to better results, but the residual artefact is still too large (figure 5c), leading to many spike sorting errors. Although the artefacts on all channels are similar, their large amplitude makes that even small relative differences result in relatively large residuals compared to the neural spiking activity [14]. Spatial averaging over the adjacent channels (also called a Laplacian spatial filter, and commonly used for EEG recordings [39])) could lead to a better artefact template, but neural information would also partly be lost, due to the spatial correlation across nearby channels [14]. More spiking data that is correlated to the spiking activity of the target channel, would leak into the template (as this is also partly seen on the adjacent channels). This problem is solved by constructing an optimized filter based on linear regression, where the weight of the included channels is automatically chosen, but the nearby channels are left out. The filter constructs an instantaneous estimate of the artefact, which can lead to different estimates (i.e. filter outputs) for each pulse, while automatically weighting the channels to (1) have a good fit to the artefact, and (2) capture as little correlated spiking data as possible. This leads to a decent spike sorting performance after artefact removal. Furthermore, the GEVD-MWF, another technique to construct such an optimal spatio-temporal filter, obtains similar results. For both of these optimal spatio-temporal filters, the largest part of the artefact reconstruction is performed by means of spatial filtering, while the temporal filtering has a small extra effect.

Furthermore, the *in vivo* validation offers anecdotal evidence which is in accordance with these findings that both linear regression and GEVD-MWF succeed in removing the artefact and reveal underlying spikes, while blanking and template subtraction either removes spikes, or leave a strong residual artefact. In the case of linear regression and GEVD-MWF, spikes are visible throughout the whole recording. We also spike sorted the *in vivo* data. Both the linear regression and the GEVD-MWF resulted in similar sortings and allowed to find spikes both during and outside stimulation belonging to the same neuron.

The proposed benchmarking framework allows an accurate comparison in performance between different techniques, and therefore also allows a direct comparison between the linear regression and GEVD-MWF as artefact removal techniques. Mathematically, they are very similar: both are optimal in a linear MMSE-sense, and use a weighted version of the other channels. The most important difference lies in the channels used for the computation of the filter and the type of regularization employed. In the linear regression, the most close-by channels are removed, while further away channels are used as filter inputs. After tuning the radius ϵ, it was observed that only the direct neighbours of the channel were removed, while electrodes that are only two steps away were preserved. The fact that only a few channels were discarded is probably because the linear regression aims to construct a signal that is as similar as possible to the channel of interest, and nearby channels are more similar than distant channels. This implies a trade-off: channels close-by are a better reference for the artefact, but also have a higher correlation in the neural data, thereby risking to remove some of the spiking information as well. The MWF, on the other hand, doesn't have this trade-off. All channels are used, and the largest power is automatically given to the channels that contain the least neural signal power, as they can be used to create an accurate artefact signal.

This also implies a difference in the computational complexity of the filter. The MWF uses the same covariance matrix ($\boldsymbol{R}_{xx}$) to compute the filter for all channels, such that it only has to be inverted once. In the case of the linear regression, each filter (one for every channel) uses a different subset of the channels, and therefore has to compute a different matrix inverse for each channel. This causes the linear regression filter to be more computationally intensive, especially for probes with high-channel counts. Recently, there is a move towards the use of high-channel count probes in electrophysiology and spike sorting. In this context, the much lower computational cost of the MWF is a strong advantage over the computationally intensive state-of-the-art linear regression filter, given that both methods achieve a similar spike sorting performance.

Using a GEVD to improve a MWF, has previously been proposed in a context of EEG artefact removal [19]. However, there it was mainly used to correct for an oversubtraction by removing negative entries in $\boldsymbol{\Sigma}_a$, as explicitly setting a low-rank did not influence the results in artefact removal for EEG [19]. Our study has shown that for artefact removal in the context of spike sorting, explicitly choosing a low rank using the GEVD-MWF instead of the full-rank MWF can make an important difference. This is due to the underlying nature of the stimulation artefacts, which are highly structured across space and time, thereby resulting in a covariance matrix with a low-rank structure. Using a GEVD-MWF then ensures that the artefact model will always be low-rank, also in cases when a normal MWF (without GEVD) would not automatically lead to a low-rank model. Imposing such low-rank structure also has an implicit regularization effect as it effectively reduces the number of degrees of freedom in the model.

VI. Conclusion

Stimulation artefact removal methods are necessary as a pre-processing step to retrieve neural information when recording and stimulation simultaneously take place in the same brain region. We developed a hybrid framework, using ground-truth data, to compare and evaluate different artefact removal methods. This hybrid framework allows us to acquire accurate, ground-truth based results on the spike sorting performance after artefact removal, allowing for an objective comparison of different methods. It further allows to test methods for recordings with an artificially chosen SNR, to get useful insight in how methods will behave in recordings of varying SNR values, and how the SNR changes after artefact removal. We have compared several existing stimulus artefact removal methods with each other, including a recent method (MWF) from the EEG literature that was never tested before for stimulation artefact removal in neural probe data. The GEVD-MWF and linear regression methods performed similarly and they both outperformed the blanking and template subtraction methods. However, in the context of high-channel count probes, MWF has an advantage over linear regression on the computational complexity.

A similar hybrid benchmark framework can be used in future research, e.g., when developing new artefact removal methods, to objectively evaluate and compare them with the state-of-the-art methods.

VII. Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

VIII. Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 802895), from FWO (Research Foundation Flanders) for project G.0D75.16N (Distributed signal processing algorithms for spike sorting in next-generation high-density neuroprobes), and from the Flemish Government (AI Research Program). Boateng Asamoah is SB PhD fellow at FWO.

Appendix

A. Mathematical description of spatio-temporal filtering

1) Linear regression: Let $x_k[t]$ denote the signal on channel k at sample time t. In order to construct a temporal filter, the time stacked vector $\bar{\boldsymbol{x}}_k[t]$ is defined as follows:

$$\bar{\boldsymbol{x}}_k[t] = \Big[x_k[t] \;\; x_k[t-1] \;\; ... \;\; x_k[t-L+1]\Big]^\intercal. \tag{6}$$

Here, L is the number of time lags of the filter that will be estimated. To further expand this to a spatio-temporal filter, all K channels are stacked together:

$$\bar{\boldsymbol{x}}[t] = \Big[\bar{\boldsymbol{x}}_1[t]^\intercal \;\; \bar{\boldsymbol{x}}_2[t]^\intercal \;\; ... \;\; \bar{\boldsymbol{x}}_K[t]^\intercal\Big]^\intercal. \tag{7}$$

Finally, when constructing an optimal filter for channel k, we only take channels that are sufficiently far away from this channel of interest into account, in an attempt to preserve the neural information, an approach first explored on a depth probe with 1 column in [16]. To this end, we define the vector $\bar{\boldsymbol{x}}_{-k}[t]$, which is defined as in (7), but all channels that are located within a distance ϵ from channel k on the probe (including channel k itself) are removed. The goal is to design a spatio-temporal filter $\boldsymbol{w}_k$ which estimates the artefact component $\hat{a}_k[t]$ at channel k by filtering the data in $\bar{\boldsymbol{x}}_{-k}[t]$.

$$\hat{a}_k[t] = \boldsymbol{w}_k^\intercal \bar{\boldsymbol{x}}_{-k}[t]. \tag{8}$$

Finally, the clean data $y_k[t]$ are defined as:

$$\begin{aligned} y_k[t] &= x_k[t] - \hat{a}_k[t] & \forall\ t \in \mathcal{A} \\ y_k[t] &= x_k[t] & \forall\ t \notin \mathcal{A} \end{aligned} \tag{9}$$

The filter $\boldsymbol{w}_k$ is optimized to estimate the artefact signal in channel k in minimum mean squared error (MMSE-)sense, by using the far-away channels as reference signals, i.e.,

$$\hat{\boldsymbol{w}_k} = \min_{\boldsymbol{w}_k} \;\; E\Big\{\Big(a_k[t] - \boldsymbol{w}_k^\intercal \bar{\boldsymbol{x}}_{-k}[t]\Big)^2\Big\}\Big|_{t\in\mathcal{A}} \tag{10}$$

where $E\{.\}$ denotes the expectation operator. Note that a_k is not known in practice. However, assuming the channels in $\bar{\boldsymbol{x}}_{-k}[t]$ are sufficiently far away from channel k, the neural (spiking) data will be uncorrelated between $x_k[t]$ and $\bar{\boldsymbol{x}}_{-k}[t]$, such that $a_k[t]$ can be replaced with $x_k[t]$:

$$\hat{\boldsymbol{w}_k} = \min_{\boldsymbol{w}_k} \;\; E\Big\{\Big(x_k[t] - \boldsymbol{w}_k^\intercal \bar{\boldsymbol{x}}_{-k}[t]\Big)^2\Big\}\Big|_{t\in\mathcal{A}}. \tag{11}$$

Solving (11) after adding an L2-norm regularization term with weight λ results in the Ridge regression solution:

$$\hat{\boldsymbol{w}_k} = (\boldsymbol{R}_{x_{-k}x_{-k}} + \lambda \boldsymbol{I})^{-1} \boldsymbol{r}_{x_{-k}x_k}. \tag{12}$$

In (12), $\boldsymbol{R}_{x_{-k}x_{-k}}$ and $\boldsymbol{r}_{x_{-k}x_k}$ respectively represent the spatio-temporal covariance matrix and vector:

$$\boldsymbol{R}_{x_{-k}x_{-k}} = E\Big\{\bar{\boldsymbol{x}}_{-k}[t]\bar{\boldsymbol{x}}_{-k}^\intercal[t]\Big\}\Big|_{t\in\mathcal{A}} \tag{13}$$

$$\boldsymbol{r}_{x_{-k}x_k} = E\Big\{\bar{\boldsymbol{x}}_{-k}[t]x_k^\intercal[t]\Big\}\Big|_{t\in\mathcal{A}} \tag{14}$$

2) Multi-channel Wiener Filter: The MWF exploits the differences in the spatio-temporal statistics of the unwanted signal component (the artefact $\boldsymbol{a}[t]$) and the wanted signal component (the neural signal $\boldsymbol{n}[t]$) to suppress this unwanted component as much as possible [18]. The recording $\boldsymbol{x}[t] = [x_1[t]...x_K[t]]^\intercal$ can thus be written as:

$$\boldsymbol{x}[t] = \boldsymbol{n}[t] + \boldsymbol{a}[t]. \tag{15}$$

The MWF takes the recorded multi-channel probe signal as an input, and produces at its output an estimate of the artefact component at each channel, which can then be subtracted from the signal:

$$\boldsymbol{y}[t] = \hat{\boldsymbol{n}}[t] = \boldsymbol{x}[t] - \boldsymbol{W}^\intercal \bar{\boldsymbol{x}}[t]. \tag{16}$$

Here, $\bar{\boldsymbol{x}}[t]$ is defined the same as in (7). The optimization criterion for this spatio-temporal linear filter $\boldsymbol{W}$ is again the linear minimum mean squared error (MMSE) with as cost function:

$$\hat{\boldsymbol{W}} = \min_{\boldsymbol{W}} \;\; E\Big\{\Big\|\boldsymbol{a}[t] - \boldsymbol{W}^\intercal \bar{\boldsymbol{x}}[t]\Big\|^2\Big\}\Big|_{t\in\mathcal{A}}. \tag{17}$$

Note the differences with the linear regression: the MWF takes all channels into account, and aims to estimate the artefact component of all channels at once. The optimal estimate for the filter $\boldsymbol{W}$ (a matrix containing the filters for all channels) then becomes:

$$\hat{\boldsymbol{W}} = \boldsymbol{R}_{xx}^{-1} \boldsymbol{R}_{aa}. \tag{18}$$

Note that we have not introduced a regularization constant as in (12), as we will later introduce a subspace-based regularization (see (22)-(25)). Again, $\boldsymbol{R}$ represents the spatio-temporal covariance matrix respectively for the signal, the artefact component and the neural data:

$$\boldsymbol{R}_{xx} = E\{\bar{\boldsymbol{x}}[t]\bar{\boldsymbol{x}}[t]^{\intercal}\}|_{t\in\mathcal{A}} \tag{19}$$

$$\boldsymbol{R}_{aa} = E\{\bar{\boldsymbol{a}}[t]\bar{\boldsymbol{a}}[t]^{\intercal}\}|_{t\in\mathcal{A}} \tag{20}$$

$$\boldsymbol{R}_{nn} = E\{\bar{\boldsymbol{n}}[t]\bar{\boldsymbol{n}}[t]^{\intercal}\}|_{t\notin\mathcal{A}} \tag{21}$$

$\boldsymbol{R}_{xx}$ can be directly estimated from the data, i.e. over all samples in $\mathcal{A}$. On the other hand, the covariance matrix of the artefact component $\boldsymbol{R}_{aa}$ is of course not known in practice (we can not use the ground-truth artefact in the design of the filter). However, the covariance matrix of the neural data $\boldsymbol{R}_{nn}$ can also be directly estimated from the data, by selecting parts of the recording where the stimulation is not active, i.e. all samples $\notin \mathcal{A}$. Note again that in most settings, including ours, the timestamps of the stimulation are known.

Since the artefact signal and the neural signal are uncorrelated, we have that $\boldsymbol{R}_{xx} = \boldsymbol{R}_{nn} + \boldsymbol{R}_{aa}$, which would in principle allow to estimate $\boldsymbol{R}_{aa}$ as $\boldsymbol{R}_{aa} = \boldsymbol{R}_{xx} - \boldsymbol{R}_{nn}$. However, this might lead to an ill-conditioned matrix and/or it might result in an indefinite matrix (note that covariance matrices are by definition positive (semi-)definite). To retrieve a reliable estimate for $\boldsymbol{R}_{aa}$, a generalized eigenvalue decomposition (GEVD) from $\boldsymbol{R}_{xx}$ and $\boldsymbol{R}_{nn}$ is constructed [19], resulting in a joint diagonalization of $\boldsymbol{R}_{xx}$ and $\boldsymbol{R}_{nn}$:

$$\boldsymbol{V}^{\intercal}\boldsymbol{R}_{xx}\boldsymbol{V} = \boldsymbol{\Sigma}_x \tag{22}$$

$$\boldsymbol{V}^{\intercal}\boldsymbol{R}_{nn}\boldsymbol{V} = \boldsymbol{\Sigma}_n \tag{23}$$

where $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_n$ are both diagonal matrices. Equations (22)-(23) result in a generalized eigenvalue problem:

$$\boldsymbol{R}_{xx}\boldsymbol{V} = \boldsymbol{R}_{nn}\boldsymbol{V}\boldsymbol{\Sigma} \tag{24}$$

where the matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_n^{-1}\boldsymbol{\Sigma}_x$ contains the generalized eigenvalues, sorted in descending order, and $\boldsymbol{V}$ contains the corresponding generalized eigenvectors in its columns. Using the fact that $\boldsymbol{R}_{xx} = \boldsymbol{R}_{nn} + \boldsymbol{R}_{aa}$, and combining (22) and (23), the following estimate of $\boldsymbol{R}_{aa}$ is found [19]:

$$\boldsymbol{R}_{aa} = \boldsymbol{V}^{-\intercal}\boldsymbol{\Sigma}_a\boldsymbol{V}^{-1} \tag{25}$$

$$with \quad \boldsymbol{\Sigma}_a = \boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_n.$$

References

[1] G. Hong and C. M. Lieber, "Novel electrode technologies for neural recordings," *Nature Reviews Neuroscience*, vol. 20, no. 6, pp. 330–345, 2019.

[2] E. J. Tehovnik, A. S. Tolias, F. Sultan, W. M. Slocum, and N. K. Logothetis, "Direct and indirect activation of cortical neurons by electrical microstimulation," *Journal of Neurophysiology*, vol. 96, no. 2, pp. 512–521, 2006.

[3] B. Asamoah, A. Khatoun, and M. Mc Laughlin, "tACS motor system effects can be caused by transcutaneous stimulation of peripheral nerves," *Nature Communications*, vol. 10, no. 1, p. 266, 2019.

[4] J. A. Brown and J. G. Pilitsis, "Motor cortex stimulation," *Pain Medicine*, vol. 7, no. SUPPL. 1, pp. S140–S145, 2006.

[5] A. Khatoun, B. Asamoah, and M. Mc Laughlin, "Simultaneously excitatory and inhibitory effects of transcranial alternating current stimulation revealed using selective pulse-train stimulation in the rat motor cortex," *Journal of Neuroscience*, vol. 37, no. 39, pp. 9389–9402, 2017.

[6] X. Jiang, S. Shen, C. R. Cadwell, P. Berens, F. Sinz, A. S. Ecker, S. Patel, and A. S. Tolias, "Principles of connectivity among morphologically defined cell types in adult neocortex," *Science*, vol. 350, no. 6264, p. aac9462, 2015.

[7] R. Tremblay, S. Lee, and B. Rudy, "GABAergic Interneurons in the Neocortex: From Cellular Properties to Circuits," *Neuron*, vol. 91, pp. 260–292, jul 2016.

[8] R. Gao, E. J. Peterson, and B. Voytek, "Inferring synaptic excitation/inhibition balance from field potentials," *NeuroImage*, vol. 158, pp. 70–78, sep 2017.

[9] M. Mahmud and S. Vassanelli, "Differential modulation of excitatory and inhibitory neurons during periodic stimulation," *Frontiers in Neuroscience*, vol. 10, feb 2016.

[10] C. Rossant, S. N. Kadir, D. F. Goodman, J. Schulman, M. L. Hunter, A. B. Saleem, A. Grosmark, M. Belluscio, G. H. Denfield, A. S. Ecker, A. S. Tolias, S. Solomon, G. Buzski, M. Carandini, and K. D. Harris, "Spike sorting for large, dense electrode arrays," *Nature Neuroscience*, vol. 19, pp. 634–641, mar 2016.

[11] S. Gibson, J. W. Judy, and D. Marković, "Spike sorting: The first step in decoding the brain," *IEEE Signal Processing Magazine*, vol. 29, no. 1, pp. 124–143, 2012.

[12] A. Zhou, B. C. Johnson, and R. Muller, "Toward true closed-loop neuromodulation: artifact-free recording during stimulation," *Current Opinion in Neurobiology*, vol. 50, pp. 119–127, 2018.

[13] Y. Erez, H. Tischler, A. Moran, and I. Bar-Gad, "Generalized framework for stimulus artifact removal," *Journal of Neuroscience Methods*, vol. 191, pp. 45–59, aug 2010.

[14] D. Young, F. Willett, W. D. Memberg, B. Murphy, B. Walter, J. Sweet, J. Miller, L. R. Hochberg, R. F. Kirsch, and A. B. Ajiboye, "Signal processing methods for reducing artifacts in microelectrode brain recordings caused by functional electrical stimulation HHS Public Access," *Journal of Neural Engineering*, vol. 15, no. 2, p. 26014, 2018.

[15] A. E. Mendrela, J. Cho, J. A. Fredenburg, V. Nagaraj, T. I. Netoff, M. P. Flynn, and E. Yoon, "A Bidirectional Neural Interface Circuit With Active Stimulation Artifact Cancellation and Cross-Channel Common-Mode Noise Suppression," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 4, pp. 955–965, 2016.

[16] D. J. O'shea and K. V. Shenoy, "ERAASR: An algorithm for removing electrical stimulation artifacts from multielectrode array recordings," *Journal of Neural Engineering*, vol. 15, no. 2, p. 26020, 2018.

[17] T. Hashimoto, C. M. Elder, and J. L. Vitek, "A template subtraction method for stimulus artifact removal in high-frequency deep brain stimulation," *Journal of Neuroscience Methods*, vol. 113, pp. 181–186, jan 2002.

[18] A. Borowicz, "Using a multichannel Wiener filter to remove eye-blink artifacts from EEG data," *Biomedical Signal Processing and Control*, vol. 45, pp. 246–255, aug 2018.

[19] B. Somers, T. Francart, and A. Bertrand, "A generic EEG artifact removal algorithm based on the multi-channel Wiener filter," *Journal of Neural Engineering*, vol. 15, no. 3, p. 036007, 2018.

[20] J. Wouters, F. Kloosterman, and A. Bertrand, "SHYBRID: A graphical tool for generating hybrid ground-truth spiking data for evaluating spike sorting performance," *Neuroinformatics*, 2020.

[21] K. Lab, "Neural recording from a 32-channel probe." URL: http://www.kampff-lab.org/validating-electrodes, last checked on 23/04/2020.

[22] P. Yger, G. L. Spampinato, E. Esposito, B. Lefebvre, S. Deny, C. Gardella, M. Stimberg, F. Jetter, G. Zeck, S. Picaud, J. Duebel, and O. Marre, "A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo," *eLife*, vol. 7, pp. 1–23, 2018.

[23] A. Buccino, C. Hurwitz, S. Garcia, J. Magland, J. Siegle, R. Hurwitz, and M. Hennig, "Spikeinterface, a unified framework for spike sorting," *eLife*, 2020.

[24] C. Rossant, "Phy - an open-source python library providing a graphical user interface for visualization and manual curation of large-scale electrophysiological data." URL: https://phy.readthedocs.io/en/latest/, last checked on 02/05/2020.

[25] Cambridge NeuroTech, "Cambridge NeuroTech Product Catalog," Tech. Rep. April, 2020.

[26] E. B. Montgomery, J. T. Gale, and H. Huang, "Methods for isolating extracellular action potentials and removing stimulus artifacts from microelectrode recordings of neurons requiring minimal operator in-

tervention," *Journal of Neuroscience Methods*, vol. 144, pp. 107–125, may 2005.

[27] J. Wouters, F. Kloosterman, and A. Bertrand, "Towards online spike sorting for high-density neural probes using discriminative template matching with suppression of interfering spikes," *Journal of Neural Engineering*, vol. 15, p. 056005, jul 2018.

[28] Xing Qian ; Yue Chen ; Yuan Feng ; Bozhi Ma ; Hongwei Hao ; Luming Li, "A Method for Removal of Deep Brain Stimulation Artifact From Local Field Potentials An identification method of mechanical properties of materials based on the full-field measurement method based on the fringe pattern View project," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 12, pp. 2217–2226, 2017.

[29] L. Sun and H. Hinrichs, "Moving average template subtraction to remove stimulation artefacts in EEGs and LFPs recorded during deep brain stimulation," *Journal of Neuroscience Methods*, vol. 266, pp. 126–136, jun 2016.

[30] S. Basir-Kazeruni, S. Vlaski, H. Salami, A. H. Sayed, and D. Markovic, "A blind Adaptive Stimulation Artifact Rejection (ASAR) engine for closed-loop implantable neuromodulation systems," in *International IEEE/EMBS Conference on Neural Engineering, NER*, pp. 186–189, IEEE Computer Society, aug 2017.

[31] K. J. Paralikar, C. R. Rao, and R. S. Clement, "New approaches to eliminating common-noise artifacts in recordings from intracortical microelectrode arrays: Inter-electrode correlation and virtual referencing," *Journal of Neuroscience Methods*, vol. 181, no. 1, pp. 27–35, 2009.

[32] H. Cao, T. Coleman, T. K. Hsiai, and A. Khademhosseini, *Interfacing bioelectronics and biomedical sensing*. 2020.

[33] T. I. Aksenova, D. V. Nowicki, and A. L. Benabid, "Filtering out deep brain stimulation artifacts using a nonlinear oscillatory model.," *Neural computation*, vol. 21, no. 9, pp. 2648–2666, 2009.

[34] G. E. Mena, L. E. Grosberg, S. Madugula, P. Hottowy, A. Litke, J. Cunningham, E. J. Chichilnisky, and L. Paninski, *Electrical stimulus artifact cancellation and neural spike detection on large multi-electrode arrays*, vol. 13. 2017.

[35] L. F. Heffer and J. B. Fallon, "A novel stimulus artifact removal technique for high-rate electrical stimulation," *Journal of Neuroscience Methods*, vol. 170, no. 2, pp. 277–284, 2008.

[36] D. W. Park, J. P. Ness, S. K. Brodnick, C. Esquibel, J. Novello, F. Atry, D. H. Baek, H. Kim, J. Bong, K. I. Swanson, A. J. Suminski, K. J. Otto, R. Pashaie, J. C. Williams, and Z. Ma, "Electrical Neural Stimulation and Simultaneous in Vivo Monitoring with Transparent Graphene Electrode Arrays Implanted in GCaMP6f Mice," *ACS Nano*, vol. 12, no. 1, pp. 148–157, 2018.

[37] K. Kim, M. Vöröslakos, J. P. Seymour, K. D. Wise, G. Buzsáki, and E. Yoon, "Artifact-free and high-temporal-resolution in vivo opto-electrophysiology with microLED optoelectrodes," *Nature Communications*, vol. 11, no. 1, pp. 1–12, 2020.

[38] M. Thunemann, Y. Lu, X. Liu, M. Desjardins, M. Vandenberghe, S. Sadegh, P. A. Saisan, Q. Cheng, K. L. Weldy, H. Lyu, S. Djurovic, O. A. Andreassen, A. M. Dale, A. Devor, and D. Kuzum, "Deep 2-photon imaging and artifact-free optogenetics through transparent graphene microelectrode arrays," *Nature Communications*, pp. 1–12, 2018.

[39] L. Guo, *Neural Interface Engineering*. Springer, 2020.