# Title

Sorting Overlapped Spikes using Superimposed Template Matching

By Srdjan Sumarac

# Significant Contributions

Designed and evaluated an algorithm to improve classification of non-isolated neural spikes.

# Poster

To add later

# Executive summary

A problem that is still present in modern day spike sorting is the presence of overlapping spikes. Resolving these waveforms at a higher accuracy would significantly improve the decoding of neural populations in the brain. In this paper, an add-on module approach that is capable of classifying and resolving overlapping spikes automatically at a higher accuracy is proposed. After constructing template waveforms from initial clustering of well-isolated spikes, a superimposed template generated from pair-wise combinations of those templates at all possible time shifts from each other is then implemented. Subsequently, overlapping waveforms were identified using a weighted absolute difference template matching method, and each spike waveform is assigned to their respective overlapping template label, and then automatically translated to the equivalent single-unit combination. This method was examined using simulated data from the SYNTH MONOTRODE dataset and the performance was evaluated using the Spike Forester accuracy metrics such as precision, recall and overall accuracy. It was found that the add-on module was able to resolve overlapping waveform at an average of 54% across the whole study, with various signal-to-noise ratios, an improvement of 4% from standard methods without the module. TALK ABOUT PRECISION AND RECALL TOO… As a result, this module is easily added as an extra step to improve standard low-dimensional clustering techniques such as k-means, dbscan, etc, in order to boost performance.

# Introduction

Neurons are the fundamental units of the brain that transmit and receive information through action potentials. These brief voltage spikes can be recorded with devices such as microelectrode arrays, silicone electrodes and tetrodes (REFERENCES) the enable the detection of large number of action potentials simultaneously. Being able to sort these spikes into their corresponding neuron is crucial for (ADD response variability, correlation, latencies, and information rates, REFERENCE). However, spike sorting has been a challenge for many years, and there is an increasing need to develop reliable spike sorting algorithms that are able to accurately distinguish neurons from background electrical noise and other spikes. It is common for neurons to fire simultaneously and alter the detected voltage waveforms. This results in spikes that significantly deviate in shape and amplitude from single-unit waveforms. These overlapped waveforms pose a significant problem for fully automated spike sorting algorithms due to the limiting constraints of available methods. Typical methods that have been proposed in literature to deal with overlapping spikes in extracellular recordings involve creating combinations of previously identified single-unit waveforms. These template waveforms are then used to measure the similarity between the detected spike and the template combination. There are many proposed methods that are used to find the best match, such as (amplitude values, Fourier coefficients, ML techniques such as SVM, decoupled parameter estimation algorithm, subtraction of templates from overlaps, Neural Networks). In each of these methods, a certain threshold needs to be determined to find the best matching template. These methods also tend to be slow, since comparisons of all potential overlapping waveforms are needed. (TALK ABOUT DIFFERENT METHODS). Methods that do not involve template matching focus on extracting robust features other than amplitude to cluster overlapping waveforms with single-unit waveforms. Due to the advancements in high density multielectrode arrays with much smaller spacing between electrodes have taken advantage of the fact that the activity of one neuron appears on several different electrodes. In order to improve some of the above methods, Superimposed Template Matching (STM) was proposed as an automated method to resolve overlapping waveforms. Single-unit templates are generated using any popular spike sorting technique, such as K-means, and then using the superposition of pair-wise combinations of single unit templates at all possible time shifts from each other, it is able to sort overlapping spikes. This method has been designed to be added as an in-between step for any standard spike sorting methods, with the advantage that it can improve overlapped spike accuracy. The method is fully automated and parameter less and unsupervised and can successfully improve the performance of spike sorting in simulated data.

# Literature Review

## Review of literature surrounding the research and how my work is positioned within the field of research

# Overview

The traditional approach to spike sorting is illustrated in Figure X.

A screenshot of a cell phone

Description automatically generated

Raw data is initially bandpass filtered and spike detection is usually performed using an amplitude threshold. The spikes are then extracted from the filtered signal and relevant features of the spike shapes are extracted, which consequently become the input to a clustering algorithm that usually performs classification of spikes into their single unit equivalent. The proposed STM model is an add-on step to any traditional spike sorting algorithm after initial classification. It can be broken down into the following main steps

1. Construction of single-unit templates
2. Generation of overlapping spike templates
3. Matching an identified overlapping waveform to the closest template match based on weighted absolute difference.

These methods will be described in the following sections in detail.

# Detailed Discussions

## Spike Detection

Initially, the raw data is bandpass filtered from 300 to 3000 Hz and used as an input the spike sorting algorithm. This allows for much easier spike visualisation and detection in the presence of noisy background activity. Spike peaks were then detected using an automatic amplitude threshold that was set as a multiple of an estimate of the median absolute deviation (MAD) of the filtered signal. This robust method proposed by (REF Quiroga) assumes noise is gaussian in nature and since the spike components do not affect the median deviation, i.e., median(|X|) = median(|N|). Even if the noise distribution deviates from Gaussian, studies have shown that MAD yields better results than using standard deviation estimates (REF). Therefore, the estimate used to compute the threshold is



Four times the standard deviation of the background noise is commonly used to set the threshold (REF), and this value was also used in this study. The value of this multiple was carefully considered. If the multiple was set too small, noise fluctuation would lead to more false positive events whereas if the multiple was set too large, low-amplitude spikes will be missed. As a result, the peaks and time stamps of any spike that exceeded this threshold was stored as a vector. A picture containing antenna, object

Description automatically generated

After peak detection, the next step in the algorithm was to extract all data points around each peak for a specified window size. The window size was initially set at 3 ms and later cut to 1.5 ms. The sampling rate set for the datasets was 24 kHz, so this would be corresponding to a spike waveform with 36 samples in total. Since each extracted waveform was aligned at the peak, further alignment algorithms were not required. Spike extraction could have further been improved using energy operators, continuous wavelet transform, cubic spline simplification or probability theories (REF). However, since this study focuses on overlapped waveform classification, these further improvements were omitted for simplicity.

A picture containing drawing

Description automatically generated

## Initial Clustering

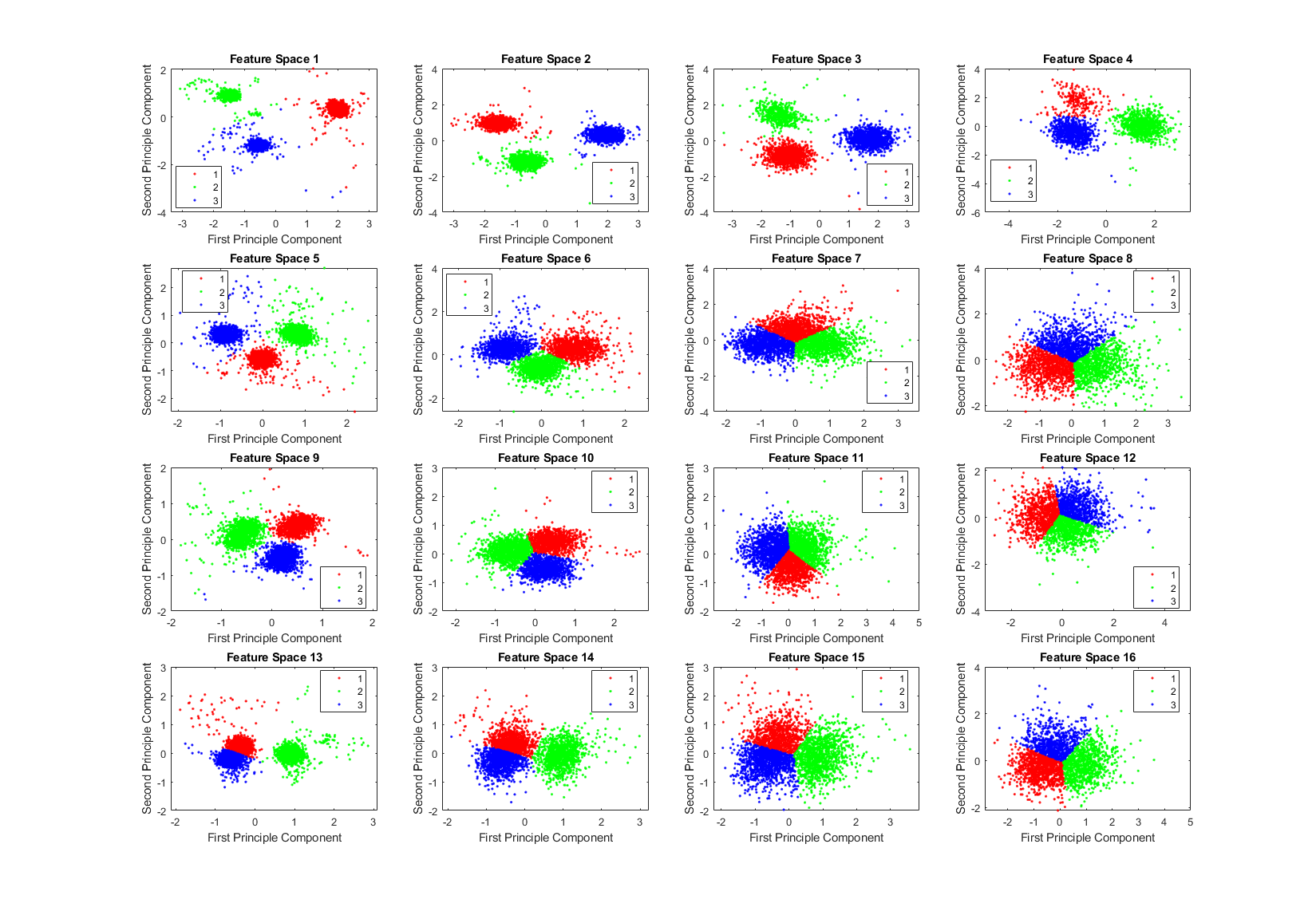
An effective way to extracting features from spike waveforms is known as principle component analysis (PCA). The logic behind PCA is that it gives an ordered set of orthogonal basis vectors that capture the directions in the data of largest variation. In order to represent a single spike, the principle components are scaled and added together and the scale factor for each component is known as the score. Since majority of the variance can be explained by the ﬁrst two principle components, only the first two principle components were used.

For both identifying the single unit templates and overlapping waveforms, the K-means algorithm was used, a clustering algorithm that attempts to split a given dataset into a fixed number of clusters. K-means is easily used to cluster high dimensional data collected from single and multichannel recording electrodes. K-means only needs the number of clusters as its parameter, and for the dataset used in this study, the number of clusters were set to three. As the focus of this paper is on improving overlapped spike classification accuracy, the choice is clustering method is not critical. It should be noted other clustering techniques such as density based clustering and fuzzy clustering were also tested during evaluation, and both methods saw performance improvements with STM.

Figure X shows how the clustered spikes from the low dimensional feature space match to those spikes at high dimensional feature space. As we can see, clusters that occur close to the centroid represent well isolated spikes. On the contrary, spikes that have a large deviation from the centroids represent spikes that deviate significantly in shape. Many of these large deviated spikes represent overlapping spike waveforms, in which traditional clustering methods fail to classify at a high accuracy. Figure X shows that as the noise increases, the clusters move closer together, and the ability to classify accurately becomes significantly more difficult.

A close up of a map

Description automatically generated



## Single Unit Template Construction

The first step that was considered in the design of STM was to identify the single unit templates that are required to build the overlapping waveforms. To find these single units, the spikes that were initially clustered using K-means and dense clusters were carefully inspected. Using the output labels from K-means classification, the median waveform for each cluster waveform was computed and used as a single unit template. Any spikes that deviated significantly from the cluster centroid were therefore removed as outliers, and any clusters with very small amplitudes were also not considered as they most likely were the result of multi-unit activity or noise. Figure X shows three clusters where each cluster has a corresponding single unit template waveform.

A close up of a map

Description automatically generated

## Generation of Overlapped Spike Templates

The overlapped templates were then generated using the single unit templates described in the section above. To do this, for each single unit template, a second unit template was shifted relative to the first and linearly superposed together at each sample in the window size. Since each single template was initially extracted at 3ms (72 samples), each superposition iteration started at the 18th sample and ended at the 54th sample, with the rest of values set at zero. The superposed template was then cut back to 36 samples to eliminate the superimposed zeros. This is done so that it was possible to directly match a 36-sample overlapped waveform with a 36-sample overlapped template. As a result, these new waveforms that have been generated can be referred to as synthetic waveforms. As an example, three synthetic waveforms that have been used throughout the paper were extracted from the superposition of two templates at one phase shift are shown in Figure X.

A close up of a map

Description automatically generated

As we can see in Figure X, two different single unit templates were linearly added together, with each waveform being phase shifted at every sample. It was assumed that it was unlikely that more than two neurons would fire simultaneously, and as a result the error associated with not sorting these overlaps would be small. Thus, in the current implementation of STM only considered pair-wise combinations of template waveforms. Although it may be possible to extend STM to more than two combinations, this becomes much more complicated due to the significant increase in the number of combinations, and the increase in likelihood that different combinations can create similar overlapping waveforms.

## Template Matching

A close up of a map

Description automatically generated

The next phase of the template matching procedure was the comparison of the overlapping waveforms with all the synthetic waveforms generated from combining the single unit templates. In order to perform this, we developed a method based on clustering. We grouped all the recorded spike waveforms that did not belong to the template clusters with the synthetic waveforms, and then re-clustered this new set of waveforms using KlustaKwik. We assumed that the overlapping waveforms and their best matching synthetic waveforms would be clustered together, and this method would be more accurate for template matching, as we would be able to use more than one feature to compare the overlapping waveforms with available templates, compared to solely using amplitude. Also this allowed us to avoid using iterative subtraction of the templates from the overlapping waveforms, which Lewicki (1998) pointed out might lead to matching errors due to the noise generated by the inaccuracies of the templates. The results of the second clustering for the sample electrode and tetrode datasets shown earlier (Figure 2) are shown in Figures 4A,B, respectively. Waveforms from six of the clusters that included both synthetic and overlapping waveforms for the tetrode data are also shown (Figure 4C).

Features we used for re-clustering of the sample electrode data were area, trigger value, Fast Fourier Transform (FFT), and the first principal component of the waveform. For the sample tetrode data, we used area, trigger value, FFT, the first principal component of the waveform, spike width, and peak value. It should be mentioned again that we empirically found that these features gave us better clustering results, but these features may vary from dataset to dataset.

In order to identify the best matching synthetic waveform for an overlapping waveform, we computed the Pearson correlation coefficient for all possible pairs of synthetic and overlapping waveforms inside each of the clusters that included both synthetic and overlapping waveforms. The synthetic waveform with the highest correlation coefficient was selected as the best match and used in the next stage to determine how the overlapping waveform would be assigned to one or both of the single unit template clusters that formed that synthetic waveform. We also attempted using other similarity measures like sum-squared error but found that the correlation coefficient returned the best overall results. In Figure 5A, an overlapping waveform (green), from the sample tetrode recording shown in Figures 2, 4, is shown along with the synthetic waveform (black) with the highest correlation coefficient. Other synthetic waveforms in the same cluster are shown in blue for reference. It is important to note that if we had used only the correlation coefficient to find the best match when comparing an overlapping waveform with all possible synthetic waveforms, this could have incorrectly grouped waveforms with similar time courses but different amplitudes. However, since the correlation coefficients were computed only on waveforms that were clustered together using both amplitude and shape features, this significantly reduced the possibility of errors using the correlation coefficient.

# Evaluation

The inter-spike interval (ISI) histogram was inspected to isolate spikes with large numbers of refractory violations where the ISI is smaller than a specified refractory period of 1ms. These spikes are then considered as a combination of multiple waveforms from different neurons and are thus considered overlapping spikes. Figure X shows the ISI for the combined datasets, and the fitted distribution. From this we can deduce that ISI is exponentially distributed with a mean of 1/19.0585 = 0.525, and so Pr(ISI < 1ms) = 0.0511 = 5.11%

A screenshot of a social media post

Description automatically generated

# Results

# Future Work

Try to implement overlapping techniques to multi-channel recordings due to the improvement of dense electrode arrays.

# Conclusion

# References

Review of methods

<https://www-tandfonline-com.ezproxy.lib.monash.edu.au/doi/pdf/10.1088/0954-898X_9_4_001?needAccess=true>

Easy data

<https://www-sciencedirect-com.ezproxy.lib.monash.edu.au/science/article/pii/S0165027009004506>

# Appendices