Statistical modelling of spike libraries for simulation of extracellular recordings in the cerebellum

Thorbergsson, Palmi Thor; Garwicz, Martin; Schouenborg, Jens; Johansson, Anders J

Published in:
Annual International Conference of the IEEE Engineering in Medicine and Biology Society

DOI:
10.1109/IEMBS.2010.5627177

2010

32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

Statistical Modelling of Spike Libraries for Simulation of Extracellular Recordings in the Cerebellum

P. T. Thorbergsson, *Student Member, IEEE*, M. Garwicz,
J. Schouenborg, A. J Johansson, *Member, IEEE*

Abstract— **Brain machine interfaces with chronically implanted microelectrode arrays for signal acquisition require algorithms for successful detection and classification of neural spikes. During the design of such algorithms, signals with a priori known characteristics need to be present. A common way to establish such signals is to model the recording environment, simulate the recordings and store ground truth about spiking activity for later comparison. In this paper, we present a statistical method to expand the spike libraries that are used in a previously presented simulation tool for the purpose described above. The method has been implemented and shown to successfully provide quick access to a large assembly of synthetic extracellular spikes with realistic characteristics. Simulations of extracellular recordings using synthesized spikes have shown to possess characteristics similar to those of in-vivo recordings in the cat cerebellum.**

I. INTRODUCTION

Brain Machine Interfaces (BMIs) are an emerging field within neuroscience. BMIs allow uni-/bidirectional communication with the central nervous system (CNS), facilitating studies of neuronal mechanisms as well as extraction of control signals for operating prosthetic devices. One class of BMIs uses extracellular recordings in the cerebral cortex as their input signals. These recordings are done with chronically implanted electrode arrays connected to external devices for data acquisition and signal processing. A major problem in this type of BMIs is the amount of data obtained from the recordings. This makes it necessary to implement efficient algorithms for extraction of relevant information and thereby reduction of data to be stored or transmitted through the system.

The extracellular recordings consist of two major components; a low frequency local field potential, representing mainly synaptic activity, and high frequency "spiking activity", representing activity of single neurons [1]. Extraction of information from single-unit spiking activity depends on successful detection and classification of spikes. During development of algorithms for these tasks, signals with a priori known characteristics (spike times and classes) are needed. We have previously implemented and reported on a simulator that is based on statistical models for spike times and basic assumptions about the recording environment [2], [3]. The simulator assigns a spike waveform to every contributing neuron and assumes that the waveform does not change during the recording. The waveforms are randomly selected from an assembly of experimentally obtained spikes. Such an assembly is referred to as a spike library. Spike shapes depend on several factors, including type and geometry of the neuron and spatial relationship between the neuron and the recording electrode [4], [5]. Although the original spike library spans a wide range of waveform morphologies, its discrete and sparse nature poses obvious restrictions in this regard.

In this paper we report on a method to expand a spike library to an arbitrary size in order to cover a wider range of waveform morphologies. We find the basis waveforms (principal components) that describe the original spike library and estimate parameters in a statistical model describing their weights. We then use the model to generate new weights that, when applied to the principal components, result in new spike waveforms that follow the statistics of the original data. The method has been shown to be successful in synthesizing an arbitrary number of spike waveforms to use in the simulation of extracellular recordings for testing of spike detection and sorting algorithms.

II. BACKGROUND

A. The Original Spike Library

Spikes were detected in and extracted from several recordings in the cat cerebellum [6]. Spikes from each recording were sorted using *Chronux* [7], [8], ensemble averaging was used for noise reduction and average spike waveforms were stored. The original spike library consists of 85 spike waveforms. In the simulations presented in this paper, some of the spikes in the original library are considered to be outliers due to excessive deviations in location and shape of major waveform landmarks and are therefore discarded in the modelling procedure described here.

B. Principal Component Analysis

When performing principal component analysis (PCA) on an ensemble of spikes, we find an orthonormal basis to describe the spikes by applying singular value decomposition (SVD) on the original spike matrix with the mean waveform subtracted from each spike. The output of the analysis are the basis vectors (principal components), their relative contributions to the total variability in the dataset (eigenvalues of the covariance matrix of the data, "latent roots"), and component

This work was supported by a Linnaeus Grant from the Swedish Research Council, ID: 60012701 and a grant from the Knut and Alice Wallenberg Foundation, nr 2004.0119.

P. T. Thorbergsson and A. J Johansson are with the Neuronano Research Center and at Dept. of Electrical and Information Technology, both at Lund University, Lund, Sweden. M. Garwicz and J. Schouenborg are with the Neuronano Research Center and at the Dept. of Experimental Medical Science, both at Lund University, Lund, Sweden.

E-mail: palmi.thor.thorbergsson@eit.lth.se

weights for every spike in the dataset [9]. By using all the principal components, the dataset can be entirely described by

$$\mathbf{S} = \mathbf{PW} \tag{1}$$

where the i-th original spike is in column i of the matrix $\mathbf{S}$, principal component j is in column j of the matrix $\mathbf{P}$ and the weight of principal component j for spike i in column i and row j of the matrix $\mathbf{W}$.

We estimate the number of principal components needed to describe the information contained in the data by looking at the relative contributions of the principal components to the variance in the data. The remaining components are assumed to describe background noise and are discarded. By setting a threshold for what percentage of variance should be considered to contain information, we can automatically find the number of principal components needed do describe the data. Since the number of principal components needed is usually smaller than the number of samples in each spike, this allows us to reduce the dimension of the problem from the original number of samples/spike to the number of principal compoents used to describe the data. This is a commonly used approach in spike sorting, where principal component weights are used as spike features. An approximation of the spike matrix in the first N principal components is

$$\hat{\mathbf{S}} = \mathbf{P}_N \mathbf{W}_N \tag{2}$$

where $\mathbf{P}_N$ contains the first N columns of $\mathbf{P}$ and $\mathbf{W}_N$ contains the first N rows of $\mathbf{W}$ in Equation (1).

C. Weight Distributions

The statistics of the spike waveforms can be examined by looking at the distributions of the weights of the first N principal components across the entire original dataset. The distributions of individual component weights can be visualized in a histogram over the rows of $\mathbf{W}_N$ in Equation (2). However, it is assumed that certain combinations of principal components are less likely than others. This motivates us to look at the joint distributions of all component weights and assume that the component weight distribution is described by a Gaussian mixture model in N dimensions and with K mixture components. I.e. the columns of $\mathbf{W}_N$ in Equation (2) are assumed to be stochastic variables coming from an $N-$dimensional $K-$modal Gaussian distribution.

A key assumption of this paper is that the original spike library is a sample drawn from a large population of spikes. This sample can be used to derive information about the statistical properties of the underlying population. By estimating model parameters, we get an idea of what the rest of the spikes in the population might look like and by generating principal component weights according to this model, we can synthesize an arbitrary number of spikes with similar characteristics as the original spikes, and with shapes within the spectrum of "possible" shapes.

III. METHODS

A. Derivation and Utilization of Model Parameters

Principal component analysis is performed on the original spike library to obtain principal components, component weights and latent roots. The cumulative sum of latent roots is plotted and a variance threshold of 99% is applied to select the number of principal components to use, N. The component weights are fitted to an N-dimensional K-modal Gaussian mixture model using the function *gmdistribution.fit* in *MATLAB®* . Since the number of modes, K, is unknown, the parameter estimation is carried out for one to six modes ($K \in [1, 6]$) and the model with the lowest Bayesian information criterion (BIC) is selected. The BIC is used as it favors models with low complexity. The estimated model is used to generate a matrix of random principal component weights, $\tilde{\mathbf{W}}$ and the new spikes are constructed by

$$\tilde{\mathbf{S}} = \mathbf{P}_N \tilde{\mathbf{W}}. \tag{3}$$

The entire procedure is illustrated in Figure (1).

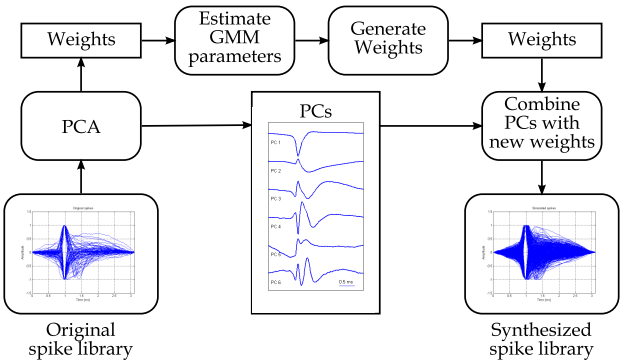

Fig. 1. Principal component analysis (PCA) is performed on the original spike library and the parameters of a Gaussian mixture model describing the resulting weight distribution are estimated. The model is used to generate new weights which are applied to the first N principal components, resulting in a synthesized spike with similar characteristics as the original spikes.

B. Evaluation of Synthesized Spike Libraries

To evaluate the overall quality of the modelling, we carry out several comparisons between the original and synthesized spike libraries. The library features of interest are distribution of spike durations, distribution of Euclidean interspike distances and sample intensity. The features are examined in histograms across the spike libraries. Usability in simulation of extracellular recordings is evaluated by running simulations in EAPSim [2], [3] with a real and synthesized spike library and comparing the power spectral densities of the simulated recordings. General appearance of spikes is evaluated in a double blind test on neuroscientists with long experience in working with spike data.

1) Feature Comparison: Original vs. Synthesized Spike Libraries: We define spike duration as the time period during which the absolute amplitude of the largest phase of the spike is above half its peak value [2]. Spike duration is calculated for all spikes in the real and synthesized spike library and the distributions are compared in histograms.

Euclidean interspike distance between spikes s_i and s_j is defined as

$$d_{i,j} = \sqrt{\sum_{m=1}^{M} (s_i(m) - s_j(m))^2}. \quad (4)$$

where m is the sample index. For each spike library, the distance is calculated for every spike pair in that library.

We define sample intensity as the histogram across every row of the spike matrices $\mathbf{S}$ and $\tilde{\mathbf{S}}$. Sample intensity provides a qualitative measure of the range of spike morphologies spanned by a spike library. A similar measure has previously been used in [7] to visualize dominating spike waveforms in an assembly of spikes in spike sorting.

2) Evaluation of Simulated Recordings: Four sets of simulated extracellular recordings are generated with EAPSim [2], [3]. Each set consists of five recordings. In two sets, we use the original spike library and in two sets, we use a synthesized library with 2000 spike waveforms. The sets have zero and four target units present respectively. Power spectral density (PSD) is estimated for all recordings using Welch's method and the mean of the PSDs of all recordings at a given setting is compared between the datasets.

3) Double Blind Test: To evaluate the quality of synthesized spikes with respect to general appearance, we present two experienced neuroscientists with a double blind test. A 9×10 matrix of spike figures, each showing either an original or synthesized spike, is shown to the subjects and they are asked to identify synthesized spikes. For each of the spike figures, we first select (with equal probability) either the original or synthesized library. We then select (without replacement) a random spike from that library. The only information given to the subjects is that each figure either shows an original or synthesized spike. The results are evaluated with the *VassarStats* statistical tool [10].

IV. RESULTS

A. Model Parameters

Figure (2) shows the relative contribution of the first N principal components to the variance in the original spike library. A 99% variance threshold is applied and we conclude that the first six principal components capture 99% of the variance in the data. According to the Bayesian information

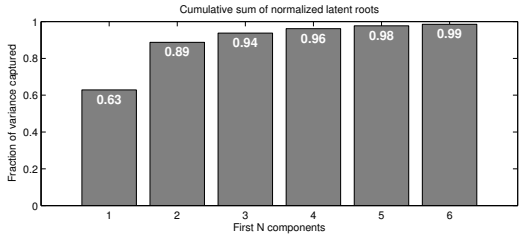

Fig. 2. The relative contribution of the first N principal components to the variance in the original spike library. $N = 6$ principal components capture 99% of the variance in the data.

criterion, we model the weight distribution with two components ($K = 2$). As a result, the weight distribution is assumed to be described by a 6-dimensional 2-component Gaussian mixture model. Table (I) summarizes the results from the parameter estimation. Figure (3) shows the first six principal components and their individual weight distributions in the original spike library.

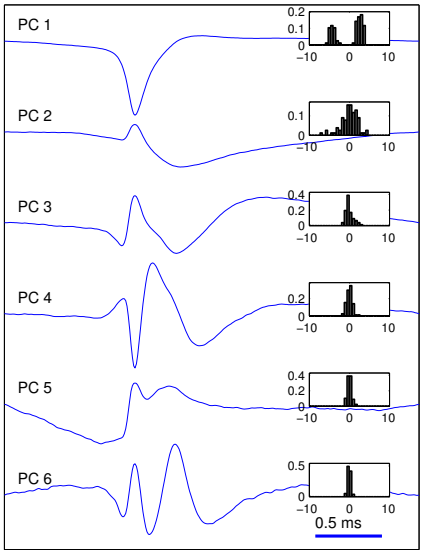

Fig. 3. The first 6 components and their original weight distributions (inset histograms).

B. Feature Comparison: Original vs. Synthesized Spike Libraries

1) Feature Comparison: Figure (4) shows the comparison between features of the original and synthesized spike libraries. A qualitative analysis of the figures shows that we obtain close matches between original and synthesized spike libraries in all cases.

2) Evaluation of Simulated Recordings: Figure (5) shows means of power spectral densities for five simulated recordings with four target units, using original and synthesized spike libraries. The results for background noise only (zero target units) are very similar and are not shown here. In [2], we showed that a good match in power spectral densities of simulated and in vivo recordings could be obtained with our original spike library. The close match between the curves in Figure (5) among with the previously mentioned observations shows that realistic spectral features in simulated extracellular recordings can be obtained even when using simulated spike libraries.

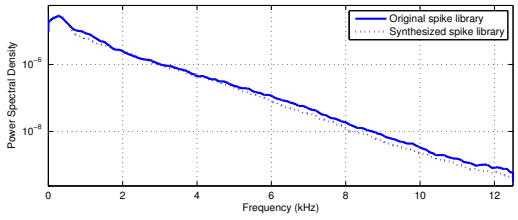

Fig. 5. Power spectral density of simulated extracellular recordings with original and simulated spike libraries.

TABLE I

GAUSSIAN MIXTUREMODEL PARAMETERS

Parameter	Component 1	Component 2
Mixing proportion	0.62	0.38
Mean	$\begin{bmatrix} 2.66 & -0.11 & -0.08 & 0.08 & -0.02 & -0.01 \end{bmatrix}$	$\begin{bmatrix} -4.4 & 0.18 & 0.13 & -0.12 & 0.03 & 0.02 \end{bmatrix}$
Covariance matrix	$\begin{bmatrix} 0.46 & 0.67 & 0.27 & -0.23 & 0.15 & 0.14 \\ 0.67 & 5.29 & -0.64 & -0.49 & 0.00 & 0.13 \\ 0.27 & -0.64 & 0.57 & -0.05 & 0.16 & 0.11 \\ -0.23 & -0.49 & -0.05 & 0.19 & 0.00 & -0.05 \\ 0.15 & 0.00 & 0.16 & 0.00 & 0.29 & 0.00 \\ 0.14 & 0.13 & 0.11 & -0.05 & 0.00 & 0.08 \end{bmatrix}$	$\begin{bmatrix} 0.53 & 0.18 & 0.46 & -0.49 & -0.03 & -0.05 \\ 0.18 & 4.49 & 1.02 & 0.84 & -0.01 & -0.22 \\ 0.46 & 1.02 & 1.62 & 0.11 & -0.27 & -0.19 \\ -0.49 & 0.84 & 0.11 & 0.9 & 0.01 & 0.09 \\ -0.03 & -0.01 & -0.27 & 0.01 & 0.31 & 0.00 \\ -0.05 & -0.22 & -0.19 & 0.09 & 0.00 & 0.29 \end{bmatrix}$

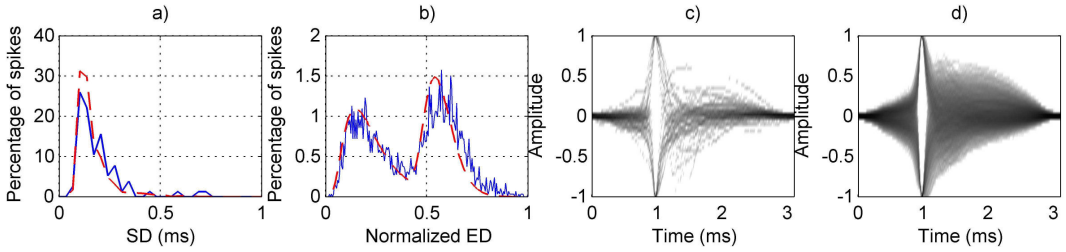

Fig. 4. Feature comparison between the original (whole lines in a) and b)) and synthesized (dashed lines in a) and b)) spike libraries. Figures a) and b) show distributions of spike durations and interspike distances respectively. Figures c) and d) show sample intensity for the original and synthesized spike libraries respectively.

3) Double Blind Test: Analysis of the double blind tests show, within a 95% confidence interval, that none of the subjects performed significantly better than chance when discriminating between original and synthesized spikes.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we have described a method in which we use principal component analysis to obtain a statistical model to describe the waveforms in an experimentally obtained spike library. The statistical model, among with the originally obtained principal components, is used to synthesize a spike library of arbitrary size. Our results show that the modelling and synthesis result in spikes with realistic features, usable for realistic simulation of extracellular recordings in the cerebellum.

The model will be implemented into EAPSim [3] for common use. By allowing principal component weights to move within the modelled distribution, we can model variations in spike shapes over time within or between recordings. These variations would facilitate studies on algorithms for spike tracking and spike sorting under dynamic conditions.

Our results show that six principal components are sufficient to describe 99% of the variance in the original spike library of cerebellum recordings. This result is consistent with the independent results reported in [11] where the authors performed principal component analysis on a large ensamble of spikes coming from different neurons and concluded that 99% of the variance was described by the first six principal components. These results give us reason to suspect that spike sorting algorithms with correlation against a constant set of basis shapes (PCs) might be feasible.

REFERENCES

[1] H. Bokil et al. A Method for Detection and Classification of Events in Neural Activity, *IEEE Transaction on Biomedical Engineering*, vol. 53, no. 8, pp. 1678-1687, 2006.

[2] P. T. Thorbergsson et al. Spike Library Based Simulator for Extracellular Single Unit Neuronal Signals, *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6998-7001, 2009.

[3] Website: *EAPSim*, http://www.eit.lth.se/staff/Palmi_Thor.Thorbergsson

[4] D. A. Henze et al. Intracellular Features Predicted by Extracellular Recordings in the Hippocampus In Vivo, *Journal of Neurophysiology*, vol. 84, 2000, pp 390-400.

[5] M.S. Fee, P. Mitra, and D. Kleinfeld. Variability of Extracellular Spike Waveforms of Cortical Neurons, *Journal of Neurophysiology*, vol. 76, pp. 3823-3833, 1996.

[6] H. Jörntell and C.F. Ekerot, Reciprocal bidirectional plasticity of parallel fiber receptive fields in cerebellar Purkinje cells and their afferent interneurons. *Neuron*, vol. 34, pp. 797-806, 2002.

[7] Website: *Chronux Analysis Software*, http://www.chronux.org.

[8] P. Mitra, H. Bokil. *Observed Brain Dynamics*, Oxford University Press, USA, 2008.

[9] M. S. Lewicki. A review of methods for spike sorting: the detection and classification of neural action potentials, *Network: Comput. Neural Syst.*, vol. 9, pp. R53-R78, 1998.

[10] Website: *VassarStats*, http://faculty.vassar.edu/lowry/clin1.html

[11] M. S. Fee et al. Automatic sorting of multiple unit neuronal signals in the presence of anisotropic and non-Gaussian variability, *Journal of Neuroscience Methods*, vol. 69, pp. 175-188, 1996.