Using FPGAS to accelerate the training process of a Gaussian mixture model based spike sorting system
by Yongming Zhu

A thesis submitted in partial fulfillment of the requirement for the degree of Master of Science in Electrical Engineering
Montana State University

Abstract:
The detection and classification of the neural spike activity is an indispensable step in the analysis of extracellular signal recording. We introduce a spike sorting system based on the Gaussian Mixture Model (GMM) and show that it can be implemented in Field Programmable Gate Arrays (FPGA). The Expectation Maximization (EM) algorithm is used to estimate the parameters of the GMM. In order to handle the high dimensional inputs in our application, a log version of the EM algorithm was implemented. Since the training process of the EM algorithm is very computationally intensive and runs slowly on Personal Computers (PC) and even on parallel DSPs, we mapped the EM algorithm to a Field Programmable Gate Array (FPGA). It trained the models without a significant degradation of the model integrity when using 18 bit and 30 bit fixed point data. The FPGA implementation using a Xilinx Virtex II 3000 was 16.4 times faster than a 3.2 GHz Pentium 4. It was 42.5 times faster than a parallel floating point DSP implementation using 4 SHARC 21160 DSPs.

USING FPGAS TO ACCELERATE THE TRAINING PROCESS OF A GAUSSIAN MIXTURE MODEL BASED SPIKE SORTING SYSTEM

by

Yongming Zhu

A thesis submitted in partial fulfillment
of the requirement for the degree

of

Master of Science

in

Electrical Engineering

MONTANA STATE UNIVERSITY
Bozeman, Montana

November 2003

APPROVAL

of a thesis submitted by

Yongming Zhu

This thesis has been read by each member of the thesis committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

Dr. Ross K. Snider Ross Snider 12-1-03
(Signature) Date

Approved for the Department of Electrical and Computer Engineering

Dr. James Peterson James N. Peterson 12-1-03
(Signature) Date

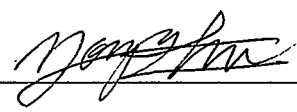
Approved for the College of Graduate Studies

Dr. Bruce McLeod Bruce R. McLeod 12-8-03
(Signature) Date

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library.

If I have indicated my intention to copyright this thesis by including a copyright notice page, copying is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for permission for extended quotation from or reproduction of this thesis in whole or in parts may be granted only by the copyright holder.

Signature 

Date 12/1/2003

TABLE OF CONTENTS

TABLE OF CONTENTS - CONTINUED

LIST OF FIGURES

FIGURE PAGE

LIST OF FIGURES - CONTINUED

FIGURE PAGE

LIST OF TABLES

TABLE PAGE

ABSTRACT

The detection and classification of the neural spike activity is an indispensable step in the analysis of extracellular signal recording. We introduce a spike sorting system based on the Gaussian Mixture Model (GMM) and show that it can be implemented in Field Programmable Gate Arrays (FPGA). The Expectation Maximization (EM) algorithm is used to estimate the parameters of the GMM. In order to handle the high dimensional inputs in our application, a log version of the EM algorithm was implemented. Since the training process of the EM algorithm is very computationally intensive and runs slowly on Personal Computers (PC) and even on parallel DSPs, we mapped the EM algorithm to a Field Programmable Gate Array (FPGA). It trained the models without a significant degradation of the model integrity when using 18 bit and 30 bit fixed point data. The FPGA implementation using a Xilinx Virtex II 3000 was 16.4 times faster than a 3.2 GHz Pentium 4. It was 42.5 times faster than a parallel floating point DSP implementation using 4 SHARC 21160 DSPs.

CHAPTER 1

INTRODUCTION

What is Neural Spike Sorting?

Most neurons communicate with each other by means of short local perturbations in the electrical potential across the cell membrane, called action potentials or spikes [1]. By using extracellular glass pipettes, single etched (sharp) electrodes, or multiple-site probes, scientists have been able to record the electrical activity of neurons as they transmit and process information in the nervous system. It is widely accepted that information is coded in the firing frequency or firing time of action potentials [2,3]. It is further believed that information is also coded in the joint activities of large neural ensembles [4]. Therefore, to understand how a neural system transmits and processes information, it is critical to simultaneously record from a population of neuronal activity as well as to efficiently isolate action potentials arising from individual neurons.

Single nerve cell recording is possible by using intracellular electrodes. Glass pipettes and high-impedance sharp electrodes are used to penetrate a particular nerve cell and monitor the electrical activities directly. This has been used to understand the mechanism of action potentials and many other neuronal phenomena [5,6]. However, these electrodes are not practical in multi-neuron recording. For intact free-moving animals, a small movement of the intracellular electrodes will easily damage the nerve cell tissue. Furthermore, it is very difficult to isolate a large number of neurons in a small local region. In awake animals, the isolation sometimes lasts for a very short

period of time. Fortunately, since all that we need is the timing of action potential, it is possible to use extracellular electrodes to acquire this information. With larger tips than intracellular electrodes, extracellular electrodes can simultaneously record signals of a small number (3 to 5) of neurons from a local region. Figure 1 shows the waveform comprised of action potentials recorded by one extracellular microelectrode. Each voltage spike in the waveform is the result of an action potential from one or more neurons near the electrode. The process of identifying and distinguishing these spikes that arise from different neurons is called "spike sorting".

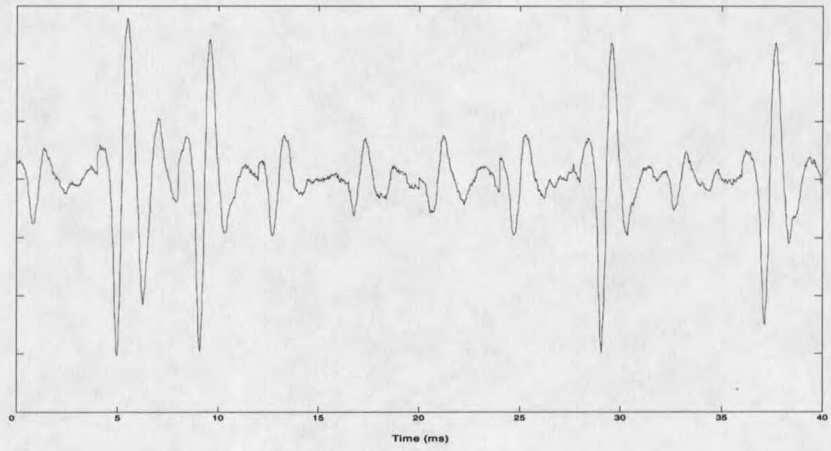


Figure 1-1 This extracellular recording waveform shows different action potentials from an unknown number of local neurons. The data were recorded from an adult female cricket's cercal sensory system. (Data courtesy of Alex [7])

Spike sorting provides an alternative to physical isolation for multi-neuron recording. In this approach, no effort is made to isolate a single cell; rather the spikes due to several cells are recorded simultaneously and sorted into groups according to their waveforms. Each group is presumed to represent a single cell since their waveforms change as a function of position relative to the electrode. The closer an

electrode is to a particular neuron, the greater the amplitude of the signal will be compared with other waveforms.

Spike Sorting System

A typical system that measures and analyses extracellular neural signals is shown in Figure 1-2. There are four stages between the electrode and the identification of spike trains from individual neurons.

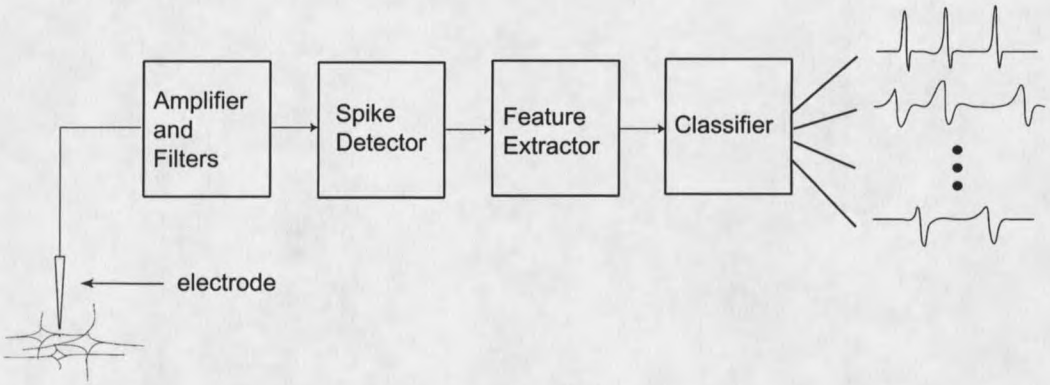


Figure 1-2 A typical neural signal recording, detecting and spike sorting system.

In the first stage neural signals are picked up by extracellular electrodes and amplified and then filtered. Low-pass filters are used to smooth the spike waveforms and provide an anti-aliasing filter before sampling. The frequency content of the waveform is typically less than 10 KHz. Other methods, such as wavelet denoising [8] can also be used to remove the recording noise. The second stage is spike detection. This is usually achieved by a simple threshold method, in which spikes are picked up when the maximum amplitude is bigger than a manually set threshold value. However, other methods including non-linear energy operators [9], wavelet-based detectors [10] and slope-shape detectors [8] have been used to improve the detection performance,

especially in the low SNR situations. A new approach, in which a noise model is first generated and then the distance from this noise model is computed to identify spikes [7] is used to obtain the spikes for this paper. However, even in this approach a threshold must be chosen.

Once the data has been collected, features must be extracted to be used in classification. The features can be simple user-defined features such as peak spike amplitude, spike width, slope of initial rise, etc. Another widely used method for feature extraction is Principle Components Analysis (PCA) [11]. PCA is used to find an ordered set of orthogonal basis vectors that can capture the directions in the data of largest variation. The first K principal components will describe the most variation of the data and can be used as features to classify the spikes. Wavelet decomposition has also been used to define spike features in a combined time-frequency domain. Under high background noise, another linear transforming method based on entropy maximization has been reported to generate good features for classification [12].

People use extracted features instead of the full sampled waveform because of the "Curse of Dimensionality", which means the computational costs and the amount of training data needed grow exponentially with the dimension of the problem. In high dimensional space, more data will be needed to estimate the model densities. Hence, with limited computational ability and training data, high dimensional problems are practically impossible to solve, especially in regards to statistical modeling. A person's inability to view high dimensional data is another reason to decrease the data dimension to two or three for manual clustering.

In this paper, we present a log version of the Expectation Maximization (EM) algorithm which solves the numerical problem arising from the high dimensional input based on the Gaussian Mixture Model. In this case, we can use the full sampled spike waveforms which preserve all the information of the neural spikes, as the classifier input.

The last stage is the clustering of spikes. In this stage, spikes with similar features are grouped into clusters. Each cluster represents the spikes that arise from a single neuron. In most laboratories, this stage is done manually with the help of visualization software. All the spikes are plotted in some feature space and the user simply draws ellipses or polygons around sets of data which assigns data to each cluster. This process is extremely time-consuming and is affected by human bias and inconsistencies.

The development of microtechnology enables multi-tip electrode recording, such as strereotrode [13] or tetrode [14] recording . By adding spatial information of action potentials, multi-electrode recording can improve the accuracy of spike sorting [15]. Obviously, the amount of data taken from multi-channel recording can be too overwhelming to deal with manually. To solve this problem, some type of automated clustering method is required. During the past three decades, various unsupervised classification methods have been applied to the spike sorting issue. The applications of general unsupervised clustering methods such as K-means, fuzzy C-means, and neural network based classification schemes have achieved some success. Methods such as Template Matching and Independent Component Analysis have also been used. A complete review can be seen in [16].

In most cases, the classification is done off-line. There are two reasons: 1. Most clustering methods need user involvement. 2. Most automated clustering methods are very computational intensive and general hardware, such as personal computers or work stations, can't meet the real-time requirement of the data streams. However, online spike sorting is needed for many applications of computational neuroscience [8]. Implementing a classification method on high-speed hardware will reduce the training time needed to develop sophisticated models, which will allow more time to be devoted to data collection. This is important in electrophysiology where limited recording times are the norm.

We implement a Gaussian Mixture Model based spike sorting system on both a DSP system and a Field Programmable Gate Array (FPGA) platform. The FPGA implementation speeds up the system dramatically which will allow an online spike sorting system to be implemenated.

Gaussian Mixture Model

If we view the measured spike signal as the combination of the original action potentials with the addition of random background noise, we can use a statistical process to model the spike signal generating process. Clustering can then be viewed as a model of the statistical distribution of the observed data. The whole data set can then be modeled with a mixture model with each cluster being a multivariate Gaussian distribution.

Gaussian mixture models have been found very useful in many signal processing applications, such as image processing, speech signal processing and pattern recognition [17,18]. For the spike sorting problem, a number of studies, based on the

Gaussian Mixture Model, have also been tried to provide statistically plausible, complete solutions. These studies have shown that better performance can be obtained by using a GMM than other general clustering methods, such as K-means [19], fuzzy c-means [20] and neural-network based unsupervised classification schemes [21]. With some modifications, the GMM based approaches also show promise in solving the overlap and bursting problems of spike sorting [22].

The Expectation Maximization (EM) algorithm is normally used to estimate the parameters of a Gaussian Mixture Model. EM is an iterative algorithm which updates mean vectors and covariance matrices of each cluster on each stage. The algorithm is very computational intensive and runs slowly on PCs or workstations.

In this paper, we present a log version of the EM algorithm which can handle high dimensional inputs. We also implement this revised EM algorithm on three different platforms, which include the PC, Parallel DSP and FPGA platforms. By comparing the different implementations, we found that, in terms of speed, the FPGA implementation gives the best performance.

Overview of the Thesis

The thesis will mainly focus on the following three points: the Gaussian Mixture Model, the Expectation Maximum algorithm, and the hardware implementation of the EM algorithm. Chapter 2 introduces the Gaussian Mixture Model and how to use the EM algorithm to estimate the mixture parameters. In Chapter 3, a log version of the EM algorithm and its performance on a PC is presented. The details of the implementation of the EM algorithm on a parallel DSP system are described in Chapter 4. Chapter 5 describes the parallel implementation of the EM algorithm on a

Xilinx Virtex II FPGA and compares the performance of FPGA with the previous implementations. Finally, Chapter 6 summarizes the work of this thesis and suggests possible future research directions.

CHAPTER 2

THE GAUSSIAN MIXTURE MODEL AND THE EM ALGORITHM

Introduction

This chapter introduces the Gaussian Mixture Model (GMM) and a maximum likelihood procedure, called the Expectation Maximization (EM) algorithm, which is used to estimate the GMM parameters. The discussion in this chapter will focus on the motivation, advantage and limitation of using this statistical approach. Several considerations for implementing the EM algorithm on actual hardware will be discussed at the end of this chapter.

The chapter is organized as follows: Section 2.2 serves to describe the form of the GMM and the motivations to use the GMM in clustering the action potentials from different neurons. In section 2.2, first a description of the Gaussian Mixture Model is presented. Then several assumptions and limitations of using the GMM in spike sorting are discussed. Section 2.3 introduces the clustering of the action potentials based on Bayesian decision boundaries. Finially, the EM algorithm and derivation of the GMM parameters estimation procedure are presented.

Mixture Model for Spike Sorting

The Gaussian Mixture Model (GMM)

A Gaussian mixture density is a weighted sum of a number of component densities. The Gaussian Mixture Model has the ability to form smooth approximations to arbitrarily shaped densities. Figure 2-1 shows a one-dimensional example of the

GMM modeling capabilities. An arbitrary non-Gaussian distribution (shown in solid line) is approximated by a sum of three individual Gaussian components (shown by dashed lines).

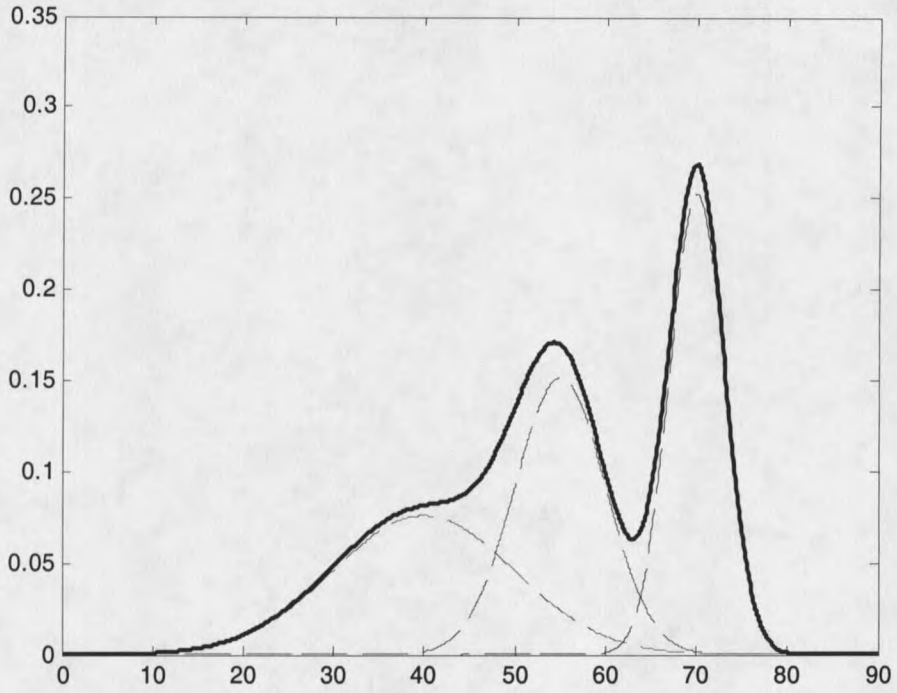


Figure 2-1 Non-Gaussian density estimation using a GMM

The complete Gaussian mixture density is parameterized by the mean vectors (μ), covariance matrices (Σ) and weights (α) from all component densities which are represented by the notations $O = \{\alpha, \mu, \Sigma\}$. The model is given by Eq. 2-1,

$$p(x \mid O) = \sum_{l=1}^{M} \alpha_l p(x \mid o_l) \qquad \text{Eq. 2-1}$$

where $p(x|O)$ is the probability of x given model O, x is a D-dimensional random vector, $p(x \mid o_l)$ (i = 1,…,M) are the component densities and α_i (i = 1,…,M) are the

mixture weights. Each component density is a D-variant Gaussian function given by Eq. 2-2,

$$p(x \mid o_l) = \frac{1}{(2\pi)^{R/2} \mid \sum_l \mid^{1/2}} e^{-\frac{1}{2}(x-\mu_l)^T \sum_l^{-1} (x-\mu_l)} \qquad \text{Eq. 2-2}$$

where μ_l is the mean vector and $\sum_l$ is the covariance matrix of each individual Gaussian model o_l.

The mixture weights α_l satisfy the constraint that $\sum_{l=1}^{M} \alpha_l = 1$ ($\alpha_l \geq 0$), which ensures the mixture is a true probability density function. The α_l's can be viewed as the probability of each component.

In spike sorting, action potentials generated by each neuron are represented by each component in a Gaussian Mixture Model. Details about using GMMs to model the spike sorting process will be described in the next section.

GMMs for spike sorting

Since action potentials are observed with no labeling from neurons they arise from, spike generating can be considered to be a hidden process. To see how the hidden process leads itself to modeling the spike waveforms, consider the generative process in Figure 2-2.