DETRENDING AND DENOISING OF IMPEDANCE CYTOMETRY DATA

BY XINNAN CAO

A thesis submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Master of Science

Graduate Program in Electrical and Computer Engineering

Written under the direction of

Professor Mehdi Javanmard

and approved by

_____

_____

_____

New Brunswick, New Jersey

October, 2015

ABSTRACT OF THE THESIS

Detrending and Denoising of Impedance Cytometry Data

by Xinnan Cao

Thesis Director:

Professor Mehdi Javanmard

Bioelectrical signals such as electrocardiogram (ECG) and flow cytometry signals are often affected by both low-frequency and high frequency perturbations during data acquisition. It is necessary to remove these interferences from the acquired data so that the pertinent information needed from these signals can be obtained.

Methods for detrending and denoising in ECG are well established, but in the context of impedance cytometry data, there lacks a body of literature available that provides guidance as to a robust methodology for processing the data. For the first time in the context of impedance cytometry, to the best of our knowledge, in this work we systematically studied and compared the performance of different algorithms for detrending and denoising, and developed a procedure to analyze impedance cytometry data with minimal error. This work can serve as guidance to select the optimal algorithm based on the following parameters: Standard Deviation, Correlation Coefficient, Power Spectral Density (PSD), Root Mean Square Error (RMSE) and Root Mean Square Difference

(RMSD).

The approaches discussed conventional filtering techniques such as: Butterworth Filtering, Chebyshev Filtering, as well as the thresholding techniques in the Discrete Wavelet Transform (DWT). The procedure of selecting the mother wavelet basis functions and the four different thresholding methods are discussed and compared. The performance of the optimized algorithm is compared with the pClamp10 commercially available cytometry analysis software resulting in 19.2% improvement in amplitude preservation, 18.4% improvement in area preservation and more than 50% improvement in the peak-search error rate.

Acknowledgements

I would like to thank my advisor Professor Mehdi Javanmard and co-advisor Professor Waheed U. Bajwa, ECE Department, Rutgers University for providing me with valuable guidance and help throughout this research. This project could not have been accomplished without their assistance. I am grateful to Professor Laleh Najafizadeh for serving on my Master's Thesis Committee and for her evaluation of my thesis research work. I greatly appreciate their valuable contribution! Lastly, I thank my family and friends for believing in me throughout my graduate studies. Your relentless support has been very important for me.

Table of Contents

List of Figures

Chapter 1

INTRODUCTION

1.1 Background

With astounding achievements in recent time, one of the fastest developing fields is biomedical engineering [1]. The ability to effectively detect and quantify various molecular biomarkers (e.g., troponin to diagnose heart disease) is an essential tool for rapid low cost medical diagnostics. But due to the limitations of bulky instrumentation and high costs, the ability to perform the tests in real time and tests for multiple biomarkers is very difficult with current devices. In some cases, the time consumed in clinically measuring those biomarkers is significant, causing the medical professionals to lose the best time window for cure.

Based on these facts, researchers are aiming at designing low-cost wearable biomarker detection devices which could quantify molecular biomarkers in our body to help us identify the diseases and cure them in early stages. Among these low-cost lab-on-chip designs, the microfabricated impedance spectroscopy flow cytometer is a powerful tool for the characterization of cells without the need for fluorescent, magnetic or other cell markers. The main principle of the impedance spectroscopy flow cytometer is as follows: when the particles are passing through the on-chip electrodes, the impedance between those electrodes becomes larger. As a result, a detectable voltage drop is observed, and this drop is turned downside up as peak in further analysis. Ideally the amplitude of

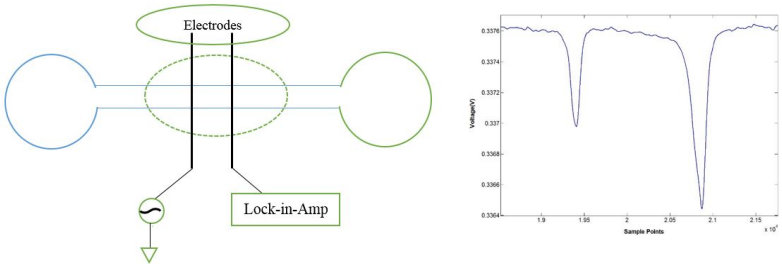

Figure 1.1 Schematic of a microfluidic impedance cytometer. The two different sizes of particles resulting in different peak amplitudes.

the peak is a function of the particle's size, and the larger size of the particles results in larger peaks compared to smaller-sized particles as shown in Fig. 1.1. To demonstrate its applications, our group developed a protein sensing method previously for detecting bead aggregates of different sizes in an impedance cytometer in [2]. By collecting the peak information and calculate its distribution, we can attempt to differentiate various types of cells or bioparticles from each other each for effective diagnosis. The experiments are monitored and recorded both electronically and optically, the ground truth of the peak number is also confirmed with the conductor of the experiments.

1.2 Interferences in Microfluidic Impedance Flow Cytometry

Various kinds of perturbations can affect flow cytometry data during its acquisition. These perturbations can corrupt the cytometry data, thus making analysis more difficult. As a consequence, it is generally accepted that removal of these interferences can help improve the performance of detection [3, 4, 5].

During literature review, we found the Electrocardiogram (ECG) data also suffers from similar perturbations [6, 7, 8, 9]. The Electrocardiogram (ECG) is a tool widely used for recording the electrical behavior of the heart by placing electrodes on a patient's body.

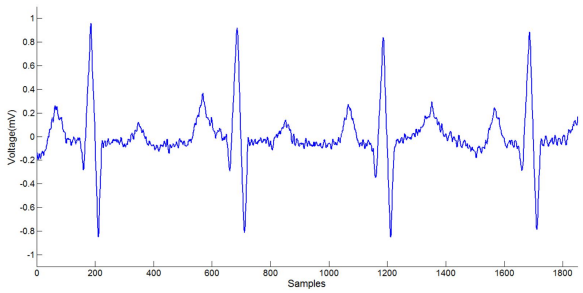

Figure 1.2 Typical ECG waveforms.

These electrodes detect the tiny nerve impulse stimulus during each heartbeat [10]. Comparing impedance cytometry signals with ECG waveforms, ECG waveform reflects the exquisite detail of the heart activity and is thus more complex. Although both of them are often affected by similar artifacts during acquisition, the signal components and analysis focus are different. The two types of perturbations that affect cytometry data are given below.

1.2.1 Baseline Wander (BW)

Baseline wander mainly consists of low-frequency signals caused by temperature fluctuations in the environment, and various electrochemical processes at the electrode surface. The fluid flows through the channel and changes in the concentration between electrodes can result in a slowly varying but non-stationary and non-linear drift.

1.2.2 White Noise

The thermal noise in the electrolyte in the channel, and also the readout instrumentation used in electronic biomarker detection contributes white noise resulting in variations in the data amplitude making the desired peaks difficult to discern. Other instrumentation perturbations and power-line interferences also affect the quality of detection.

1.3 Statement of the Problem

Extracting the peak amplitude and peak area from a noisy recording is formulated as our problem. The unstable recording environment, drifting signals caused by body motions, electrochemical reactions occurring at the electrodes and other low and high frequency perturbations are the main reasons for unwanted noise contamination of the cytometry data, which makes the peak identification and other analysis challenging.

Despite the importance of analyzing cytometry data in electronic biomarker detection, there are little experimental or sufficiently general theoretical results in the literature on how to process the data. The available results are vague [11, 12, 13, 14], and most of the literature in the impedance cytometry domain glosses over these detrending and denoising processes by claiming the data was preprocessed by some methods but without providing further details [15, 3, 4, 5]. Examples like:

"The recorded impedance signals were first preprocessed with data detrending and filtration to improve signal quality. A peak detection approach was developed to detect the spike and calculate the spike amplitude and transit time." [11]

"Data analysis was performed in MATLAB using an in-house program for wavelet denoising, baseline correction, and peak picking." [14]

This phenomenon necessitates an analysis of performance measurement that can be used as a guide on choosing detrending and denoising methods for impedance flow cytometry data.

1.4 Objective and Approach

Our main objective is to evaluate multiple conventional detrending methods and to research the applicability of wavelet based thresholding techniques for denoising the

cytometry data. The reason for selecting wavelets is due to the non-stationary characteristics of the sources of interferences, and they are similar to the interferences in ECG, where the discrete wavelet transform is commonly used for removing these interferences.

The scope of this thesis is to investigate which detrending method gives the best performance, how to select an optimal wavelet filter bank for cytometry data denoising, and comparing different thresholding rules for interferences removal, thus developing a procedure to analyze impedance cytometry data with minimal error. As a result, biosensor designers can apply this procedure for analyzing their data. These tools are also applicable to a wide range of biosensing technologies which have similar undesired interference signals that make data analysis difficult.

1.5 Organization of the Thesis

The thesis is organized as follows: Chapter 2 provides an overview of the methods utilized and presents the basics of wavelet transforms. Detrending and comparison of detrending methods including evaluation of effects on peak information are presented in Chapter 3. By treating both trend and white noise as effective noise, Chapter 4 introduces a method for optimal selection of wavelet basis function for denoising and shows the results of DWT denoising based on different thresholds. Finally in Chapter 5 we conclude our findings.

Chapter 2

METHODOLOGY

2.1 An Overview of Methods for Detrending and Denoising

The overall system design for detrending and denoising of flow cytometry data is shown in Fig. 2.1 and the arrow shows the procedure of the experiments. The major steps are: performing denoising and detrending on raw data acquired from a microfluidic impedance cytometer, and extracting the peak information (amplitude and area under peak).

As for detrending, researchers usually apply baseline restoration algorithms as preprocessing tools before further analysis, but in the work referenced, the actual techniques used were not specified [11, 13, 16]. Song et al. [11] claimed the data was preprocessed by detrending and filtration, but with no details about those methods. Gawad et al. [13] have used a High Pass Filter (HPF) for the data detrending, Valenti et al. [16] also used a filter for noise reduction, but they all did not provide further information on the filter type or the parameters used. The detrending methods compared and analyzed in this thesis are Butterworth LPF, Butterworth HPF, Chebyshev LPF and Chebyshev HPF. In order to compare the performance of different detrending methods, we added artificial white Gaussian noise to the low noise dataset to compare which method can fight against noise best.

As for denoising, wavelet denoising has already been studied thoroughly in the ECG

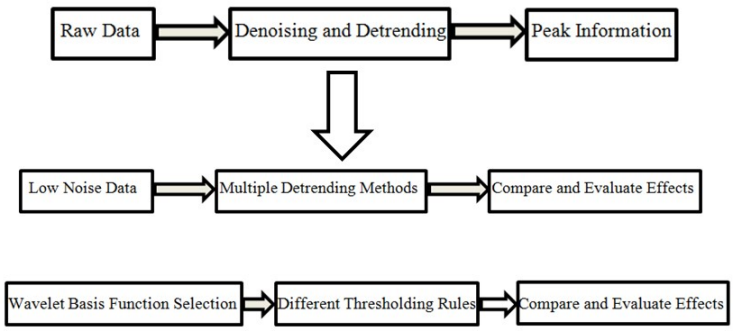

Figure 2.1 System Diagram.

domain [6, 7, 8, 9], but there is not as much discussion in the context of cytometry data analysis. As we discussed in Chapter 1, ECG signals and impedance cytometry signals are acquired from different physical devices and are different in research focus, so we explored the practicality of wavelet denoising in cytometry data by treating both trend and white noise interferences as effective perturbations and evaluated different wavelet thresholding techniques for removal of these perturbations.

2.2 Experiments and Data Acquisition

The cytometry signals are detected and recorded by a micro-electronic biochip consisting of two electrodes on a glass substrate and the impedance is measured using a lock-in amplifier [17]. The particles are injected into the micro biochip, and as the beads pass through the electrodes, the voltage drop will be detected.

From Fig. 2.2, we derive the mathematical model of our raw data acquired as $y[n]$,

$$y[n] = x[n] + f[n] + r[n], 0 \leq n \leq N-1 \tag{2.1}$$

In this assumption, the raw data consists of three components as shown in equation (2.1), where $x[n]$ is the useful information containing the peaks, $f[n]$ is the baseline function and $r[n]$ is the white noise perturbations and n is the sample points. The

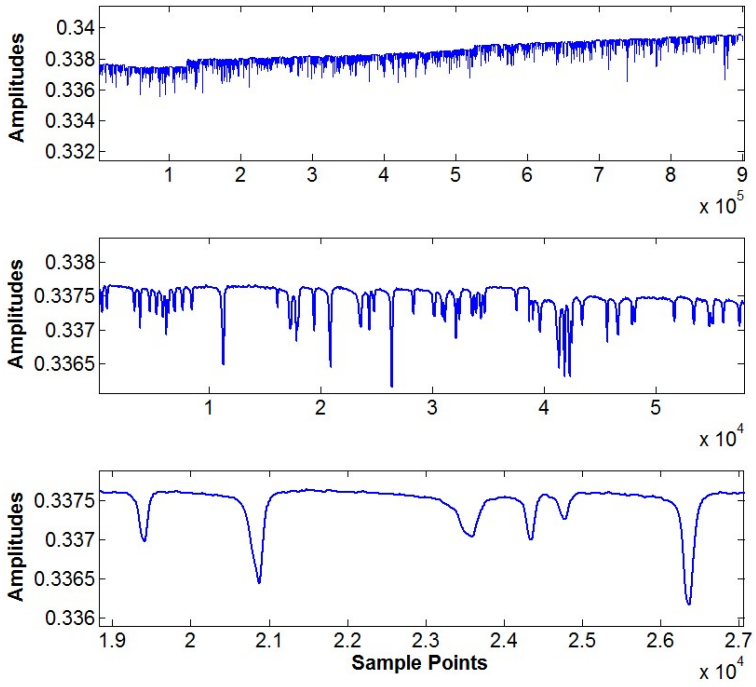

Figure 2.2 Raw data in the low noise situation.

data acquired is affected by non-stationary trend interferences. And we used the Autocorrelation Function (ACF) as a tool to match the non-stationary behavior in the Appendix.

2.3 Detrending Methods

By observing the spectrum domain of the raw data, we can acquire the related frequency information and determine the filter parameters such as the cutoff frequency. After choosing the appropriate parameters empirically, directly applying highpass filter can remove the trend easily, or we can also use the raw data to subtract the lowpass filtered data for detrending. The two types of filtering techniques discussed are: Butterworth Filtering and Chebyshev Filtering.

2.3.1 Butterworth Filter

The Butterworth filter is designed to have the maximum flat frequency response in the passband, so it's also called the maximally flat magnitude filter [18]. In the stopband, the frequency response rolls off towards zero. Comparing the Butterworth filter with other filters, the magnitude function is changing monotonically with angular frequency, while others are having ripples in their passband or the stopband. Since the low-pass and high-pass can convert to each other by switching the parameters, we give the example of low-pass filter as prototype. The transfer function of an n-order Butterworth low pass filter H(j) is shown in equation (2.2).

$$|H(j\omega)|^2 = 1/1 + (\omega/\omega_c)^{2n}. \tag{2.2}$$

where n is the order of filter, ω_c is the cut-off frequency, ω is the angular frequency.

2.3.2 Chebyshev Filter

The Chebyshev filter's mathematical characteristics are derived from Chebyshev polynomials [19]. Comparing the Chebyshev filter with Butterworth filter, the Chebyshev filter has a sharper roll-off but more ripples in the passband (type I) or stopband (type II).

The transfer function of the nth order low-pass type I Chebyshev filter is:

$$|H_n(j\omega)|^2 = 1/1 + \varepsilon^2 T_n^2(\omega/\omega_0). \tag{2.3}$$

The transfer function of the nth order low-pass type II Chebyshev filter is:

$$|H_n(j\omega)|^2 = 1/1 + \varepsilon^2 T_n^2(\omega_0/\omega). \tag{2.4}$$

where ε is the ripple factor, ω_0 is the cutoff frequency and T_n is the nth order Chebyshev polynomial. Notice the ω and ω_0 are inversed in (2.3) and (2.4), so the type II Chebyshev

filter is also known as the inverse Chebyshev filter, but this type is less common because it does not roll off as fast as type I and requires more components.

2.4 Wavelet Transform Revisited

2.4.1 Wavelet Filters

In the signal processing domain, there are many impressive tools for signal analysis. The Fourier transform is one of the most well known tools, which produces the signal into sinusoids of multiple frequencies. This method transforms the signal from time domain to frequency domain to analyze the frequency components of the signal. However, while we transform the signal to frequency domain, the information in time domain is lost. In order to correct this deficiency, Short Time Fourier Transform (STFT) was introduced by Dennis Gabor (1946) and it maps a signal by both time and frequency using a moving window function [20]. Since the size of the window is determined, we could only reach limited precision and it does not give us multi-resolution information of the signal [21]. But the wavelet analysis starts using variable-sized windowing techniques for multi-resolution information, while producing both time domain as well as frequency domain information. The wavelet transform represents the signal by a scaled and shifted version of the mother wavelet ψ.

Since calculating every possible scale of wavelet coefficients requires huge amount of work and time. In order to simplify the computation, selecting positions and scales based on powers of two so called dyadic positions and scales can also result in an efficient and accurate analysis [22, 23]. The implementation of this scheme was introduced by Mallat [22] in 1988 to calculate the wavelet decomposition of a given sequence of discrete numbers by repetitively using high pass and low pass filters as shown in Fig. 2.3.

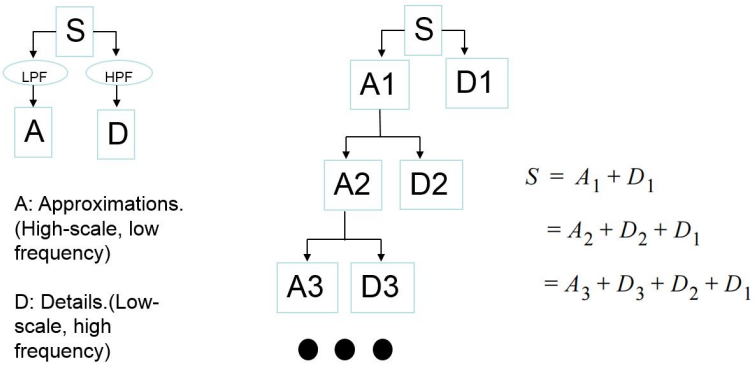

Figure 2.3 Multiple-level decomposition: tree structure DWT.

In the wavelet analysis, we define the Approximation Coefficients as the low-frequency, high-scale components of the signal and Detailed Coefficients as the high-frequency, low-scale components. As Fig. 2.3 shows, the Approximation Coefficients can be continually divided into new Approximation Coefficients and Detailed Coefficients, so the signal is broken down into many lower-resolution components. In theory this process can be iterated indefinitely, but we can only proceed until the individual details consist of only one sample point. In our case, the level of decomposition was selected by empirical experiments.

2.4.2 Wavelet Thresholding

The thresholding of wavelet is used for signal denoising, and can be divided into two types: hard thresholding Eq. 2.5 and soft thresholding Eq. 2.6 as shown in Fig. 2.4 [22, 23]. The performance of thresholding depends on the type of thresholding and the rule applied. The hard threshold function may have bigger variance and is more sensitive to tiny changes in signal, but unstable at the same time, while soft thresholding may cause bigger bias in larger wavelet coefficients.

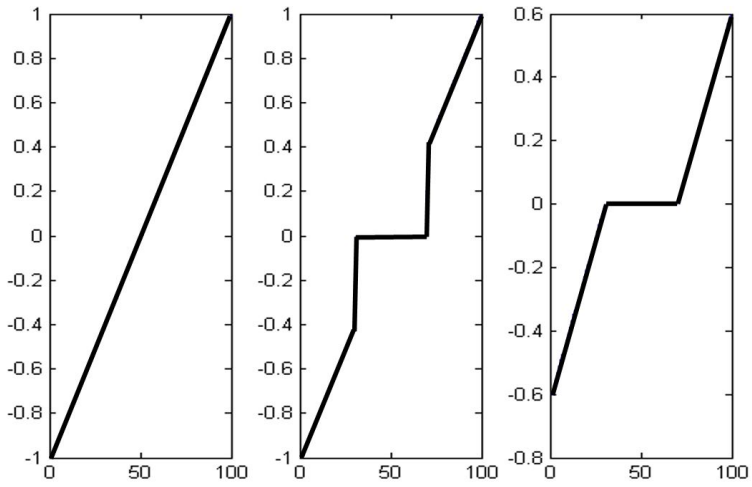

Figure 2.4 (Left) Original signal; (Mid) After hard thresholding; (Right) After soft thresholding.

The hard thresholding equation is:

$$W_H = \begin{cases} w, & |w| \geq t \\ 0, & |w| < t \end{cases} \tag{2.5}$$

And the soft thresholding equation is as follows:

$$W_S = \begin{cases} [sign(w)](|w| - t), & |w| \geq t \\ 0, & |w| < t \end{cases} \tag{2.6}$$

In Eq. 2.5 and Eq. 2.6 the w is the wavelet coefficient and t is the threshold value applied. Specifically speaking, there are four types of thresholding rules mostly used on denoising applications. We will compare each one of them in Chapter 4.

A. Steins Unbiased Risk Thresholding (Rigrsure)

This adaptive threshold selection rule is based on Steins unbiased risk estimate, and was proposed by Donoho and Jonstone in [21]. This method first uses a given threshold to estimate the risk, and then provides the threshold by minimizing this risk.

B. Universal Thresholding (Sqtwolog)

This method is a fixed thresholding method using a fixed form threshold yielding minimax performance multiplied by the log value of the number of the wavelet coefficients and it's given as:

$$w_A = \sqrt{2log(n)} \tag{2.7}$$

where n is the number of wavelet coefficients.

C. Heuristic Variant of the Steins Unbiased Risk Thresholding (Heursure)

This threshold is a combination of the previous two threshold options. As a result, if the SNR is very small, then the A option will be more noisy. In this case, option B will be selected as the threshold.

D. Minimax Thresholding (Minimaxi)

Minimax thresholding uses a fixed threshold but yields minimax performance for mean square error against ideal procedures. So the threshold value will be chose by getting a minimum of the maximum mean square error between the original signal and the wavelet coefficients of the noisy signal.

$$w_D = \begin{cases} 0.3936 + 0.1829\frac{log(n)}{log(2)}, & |n| > 32 \\ 0, & |n| \leq 0 \end{cases} \tag{2.8}$$

2.4.3 Wavelet Denoising

By applying wavelet transform to the raw signal, the signal can be represented by Eq. 2.9, W as the Wavelet Transform matrix, and there exists W^{-1}, which $W^{-1}W = I$. In order to recover the x, we give $\hat{X}$ as the estimate of X from Y in wavelet domain, also $\hat{x}$ as the estimate of x from y in time domain. Define a diagonal linear projection in Eq. 2.11. The way of recovering the useful information is given in Eq. 2.12.

$$Wy = Y = X + F + R \tag{2.9}$$

$$\hat{X} = \Delta Y \tag{2.10}$$

$$\Delta = diag(\delta_0, \delta_1, ..., \delta_{N-1}), \delta_i \in [0, 1] \tag{2.11}$$

$$\hat{x} = W^{-1}\hat{X} = W^{-1}\Delta Y = W^{-1}\Delta W y \tag{2.12}$$

The procedure is given as:

1. Get rid of the high frequency perturbations.

a. Choose an optimal wavelet filter to decompose the input signal using DWT at level L. L is chose by empirical experiments which gives us the best performance.

b. Apply the four different thresholding rules for wavelet coefficients.

c. Reconstruct the signal S_1 using the inverse DWT of the wavelet coefficients.

2. Get rid of the baseline components.

a. Choose a optimal wavelet filter to decompose the input signal using DWT at level M. Theoretically, M is the larger the better, because at this time only the lowest frequency component (the slow varying trend) is useful.

b. Apply the four different thresholding rules for wavelet coefficients.

c. Reconstruct the baseline f_1 using the inverse DWT of the wavelet coefficients.

d. The x is extracted by $x = S_1 - f_1$;

2.5 Peak Search and Information Extraction

In order to extract the amplitude and area under each peak, it is first necessary to identify and locate those peaks. The findpeaks function in Matlab returns the local maximums by comparing every three elements. The details of this peak search process is given:

1. Traverse through the dataset from the beginning to the end to identify and locate the preliminary peaks by the following algorithm:

a. Compare sample point(i) with sample point (i-1) and sample point(i+1), Only if point(i) is greater than both of them, then treat point(i) as a peak.

i. If there is a plateau, then place the peak in the center of the plateau. And If point(i+2)>point(i+1)=point(i)>point(i-1), choose either point(i+1) or point(I) as the peak.

ii. Move on to compare point(i+2) with point(i+1) and point(i+3) after each step. iii. By defining a minimum peak height as a threshold to shorten the traverse time.

b. Take i from the beginning to the end of the dataset, we will be able to identify and locate the peaks.

The minimum peak height is predefined as the mean value of data set. After locating all the peaks above the average, then we process another peak search to screen out the small peaks due to baseline variations. The peak height threshold in the second peak search is defined as $\frac{1}{3}$(mean value of the peaks from the first time peak search). The choice of $\frac{1}{3}$ is due to the statistical property of the data set which will be discussed in appendix.

After locating the peaks, we define the width of the peak as the full width at half maximum (FWHM): the distance from the front slope of the peak to the back slope of the peak measured at 50% of the maximum peak height as shown in Fig. 2.5. This definition is widely used in optical communications and it's similar to the idea of the -3 dB attenuation point in signal processing. After defining the peak widths, by integrating over each pulse width we can extract the area information under each peak easily.

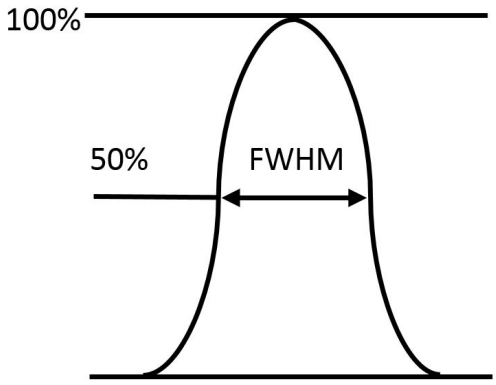

Figure 2.5 Peak width as Full Width at Half Maximum.

Chapter 3

DETRENDING

For filter design, the cutoff frequency is one of the most important parameters. In order to get this information, a Fast Fourier Transform of the low-noise signal is performed. The time domain and frequency domain of the signal are shown in Fig. 3.1. It shows there is a DC component nearly 100 dB. And if we zoom into the spectrum, the trend signal occurs approximately around 6-9 Hz, so the cutoff frequency chose is 6 Hz with no more than 3 dB of ripple in the passband, and with 20 dB attenuation in the stopband at 14 Hz. Then the minimum order of the filter can be calculated. We experimented with Butterworth lowpass filter, Butterworth highpass filter, Chebyshev lowpass filter and Chebyshev highpass filter. The lowpass filter will let the low frequency go through and filter the useful information x, so in order to get x back, the raw data should be applied to subtract the filtered signal.

From the figures, it is hard to tell which method gives us the best performance. From Fig. 3.3, by observing the trend, it can be noticed that Chebyshev LPF gives a slightly smoother baseline which seems to be the best solution, but a criteria or parameter is needed to quantify the performance. Therefore, two parameters are introduced here: standard deviation and the correlation coefficient with original data. The reasons are presented as follows:

1. A high standard deviation indicates the data are spread out and far from the mean value, but a low standard deviation indicates the data tends to be close to the mean.

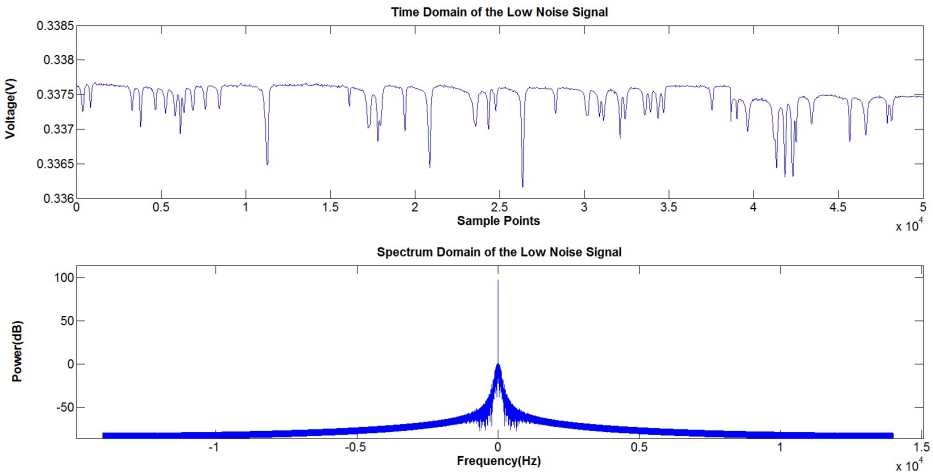

Figure 3.1 Time and frequency domain of the low-noise signal.

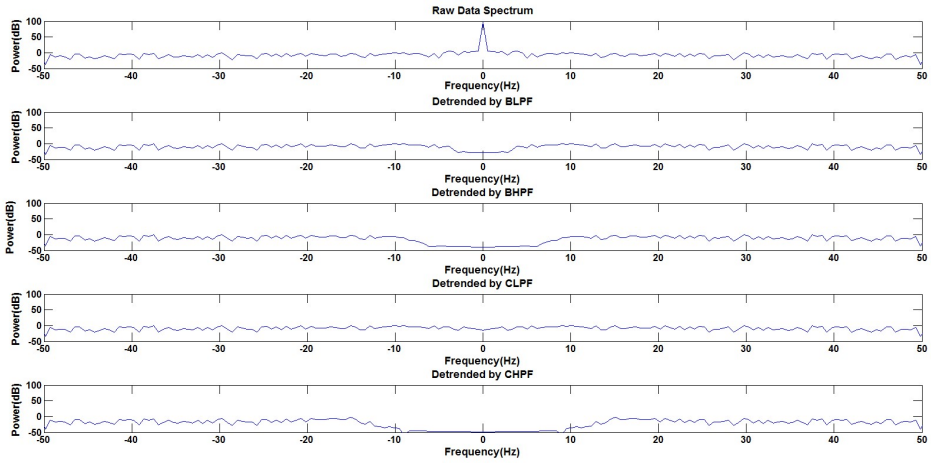

Figure 3.2 Frequency domain of the peaks after four detrending filters.

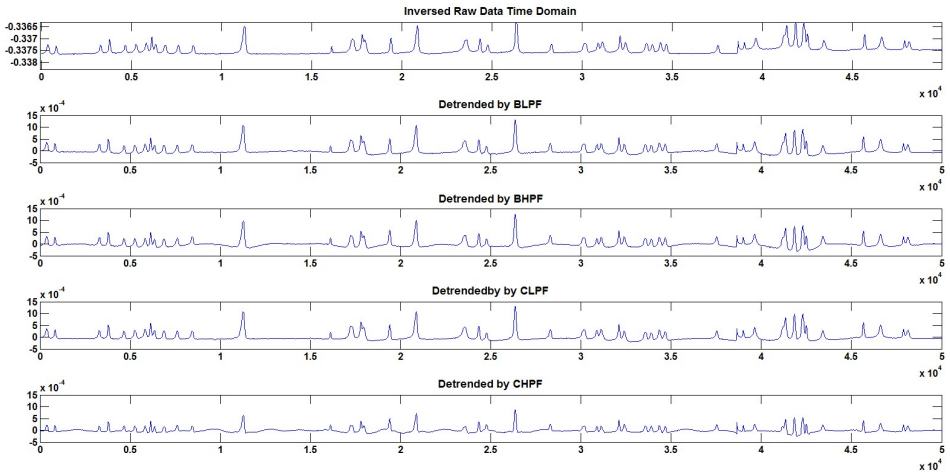

Figure 3.3 Time domain of the peaks after four detrending filters.

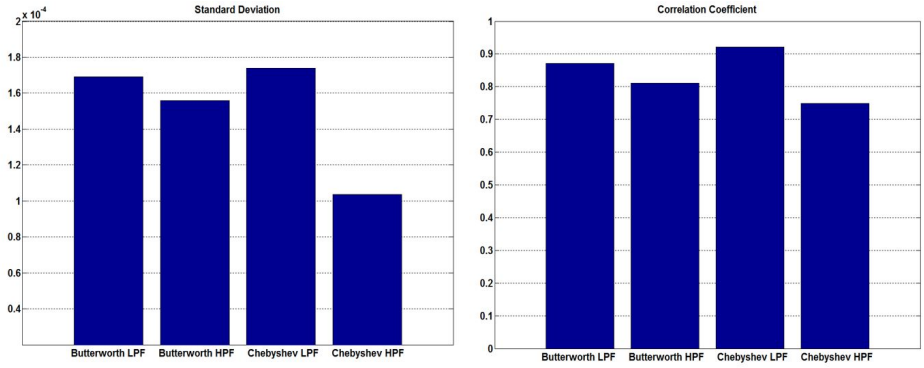

Figure 3.4 Standard deviation and correlation coefficient of the four detrending methods.

Therefore, if the standard deviation is higher, it suggests better preservation of the peaks.

2. By doing cross-correlation of the detrended signal with the original signal, we can measure the degree of similarity between those waveforms. The larger the correlation coefficient is, the higher degree of similarity it has, which also suggests the information of peaks is preserved better. From Fig. 3.4, the standard deviation and correlation coefficient of the detrended results by Chebyshev lowpass filter both returned the largest value, which suggests it is the best solution. Then the information extraction experiment will be done, and it will further strengthen the findings by the Root Mean Square Error (RMSE). In our case, the error rate ε and SNR are defined as:

$$\varepsilon = \frac{|Num(True\ Peaks) - Num(Detected\ Peaks)|}{Num(True\ Peaks)} * 100\% \tag{3.1}$$

$$Average\ Peak\ Power = (Average\ Peak\ Amplitude)^2 \tag{3.2}$$

$$SNR = \frac{Average\ Peak\ Power}{Noise\ Power} = 10log\frac{(Average\ Peak\ Power)}{(Noise\ Power)}(dB) \tag{3.3}$$

By adding White Gaussian Noise to the original signal, and performing Monte Carlo Testing for 1000 trials we formed the results seen in Table 3.1. This experiment shows which detrending method performs best against noise. In order to prove that the error in the noise-added situation is caused by the noise added, not the detrending operation,

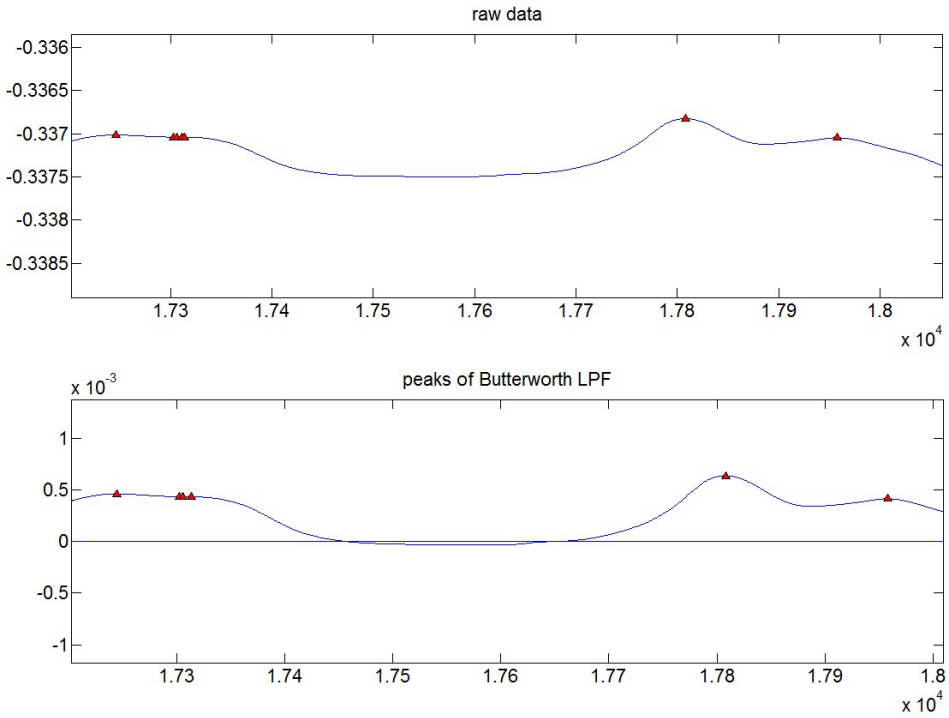

Figure 3.5 Peak search on noise-added data.

Table 3.1 Performance comparison by RMSE of the 4 different detrending methods based on 1000 M-C tests.

4 Types of Filters	Butterworth LPF	Butterworth HPF	Chebyshev LPF	Chebyshev HPF
No Noise Added	0 %	0 %	0%	0%
SNR=20dB	0%	2.0408%	0%	4.08%
SNR=16dB	0.3535%	4.3364%	0.2886%	8.8341%
SNR=13dB	1.8139%	9.8489%	1.7571%	10.4637%
SNR=10dB	10.1265%	22.6204%	8.3136%	21.1858%

it is necessary to show that the peak search on raw data will have the same error as the detrended data. Take Butterworth LPF as example in Fig. 3.5, with the smoothing effect by LPF, the errors are even reduced.

Take a closer look at the noise-added scenario as an example, the errors are caused by the small variations of noise added as shown in Fig. 3.6.

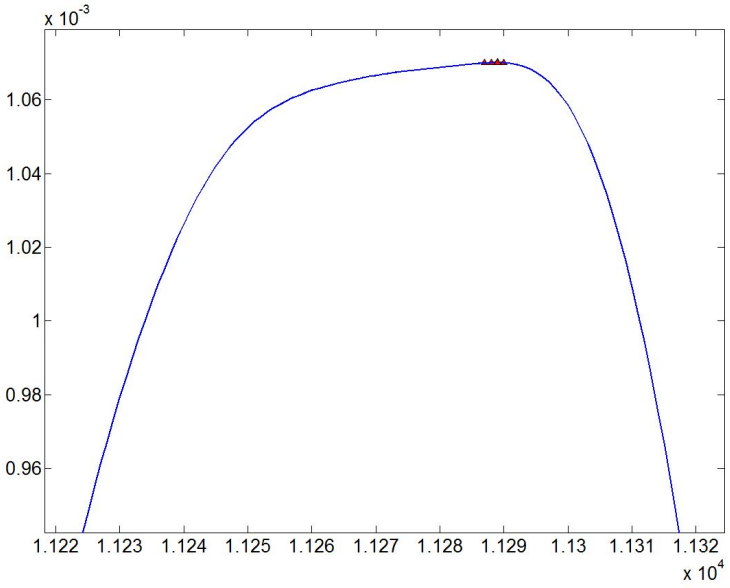

Figure 3.6 The error caused by small variations of noise.

Chapter 4

WAVELET METHODS

4.1 Optimal Wavelet Basis Function Selection

The properties of wavelets include vanishing moments, reproduction of polynomials, the order of approximation and smoothness of the basis functions. These properties are important because they provide an explanation about its performance for the signal reconstruction (the approximations and the details). So the selection of an optimal basis mother wavelet filter is important for the cytometry signal processing.

During literature review, it can be found that the self-similarity rule is a widely adopted method for the selection of the suitable basis mother wavelet filter, which means that the wavelet should be as similar as possible to the target signal. The optimal wavelet basis in ECG domain is chosen by the one leads the cross correlation coefficient reaching the maximum value [1, 24, 25]. Thus similarly the steps for choosing the optimal wavelet basis are applied to the cytometry signal:

1. Choose the basis wavelet filter from the Matlab wavelet filter bank;

2. Compute the cross-correlation coefficient between the standard peak signal and the selected wavelet basis.

3. Choose the wavelet basis which results in the maximum cross-correlation coefficient.

Table 4.1 General Properties of Wavelet Families.

Wavelet Family	Haar Wavelet	Daubechies Wavelets	Symlets	Coiflets	Biorthogonal Wavelets
Property/Short Name	haar	dbN	symN	coifN	biorNr.Nd
Orthogonal	Yes	Yes	Yes	Yes	No
Biorthogonal	Yes	Yes	Yes	Yes	Yes
Support Width	1	2N-1	2N-1	6N-1	2Nr+1, 2Nd+1
Filters Length	2	2N	2N	6N	Max(2Nr,2Nd)+2
Symmetry	Yes	Far From	Near From	Near From	YEs
Vanishing Moments for Ψ	1	N	N	2N	Nr

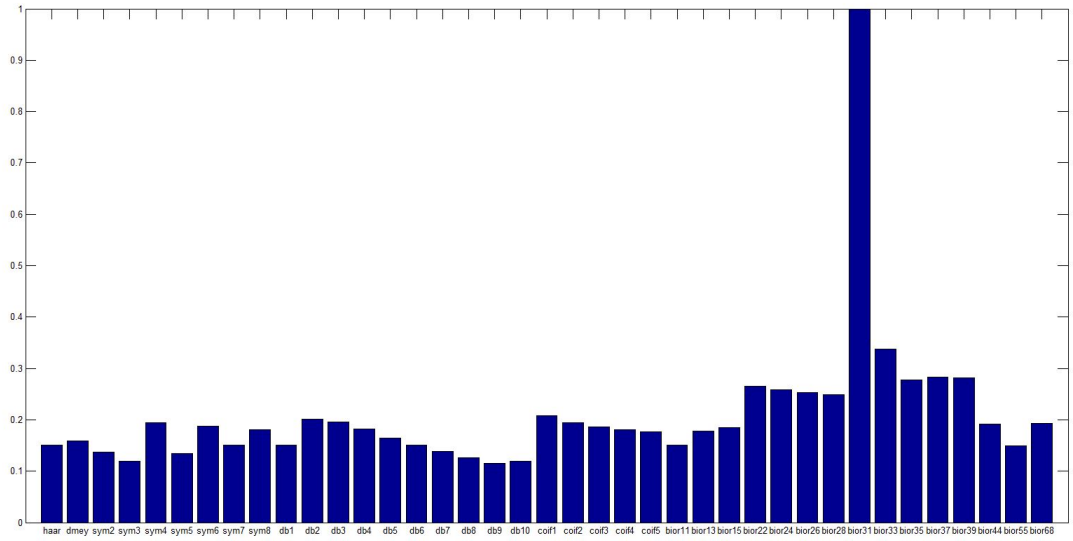

Figure 4.1 Comparison of the normalized correlation coefficients with selected mother wavelet filter for cytometry signal.

Table 4.1 summarizes the properties of the wavelet basis functions in Matlab wavelet filter banks can be used for DWT and investigation. In Fig. 4.1, the normalized cross-correlation coefficient of the standard peak signal with different wavelet filters (available in Matlab wavelet filter bank) was shown. Results show that compared with other filters, the synthesis filter of biorthogonal 3.1 filters gives better performance, because the symmetric filters are more close to the cytometry signal peak form.

4.2 Wavelet Thresholding

The performance of the wavelet thresholding techniques on cytometry signals have been measured by Power Spectral Density (PSD) and Root Mean Square Difference

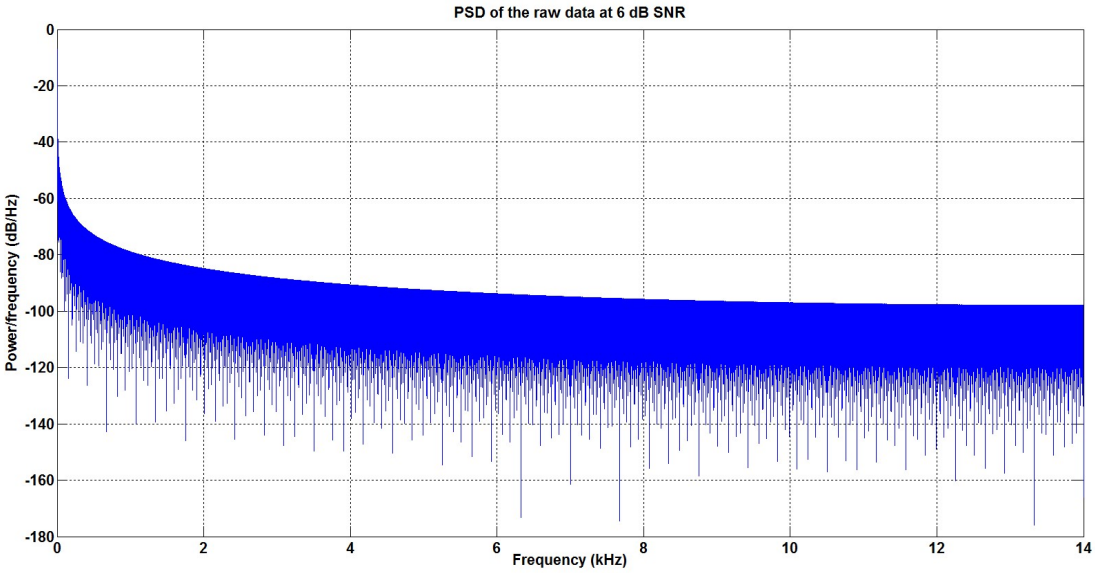

Figure 4.2 The PSD of raw data.

(RMSD). RMSE in this case is not sensitive enough to detect the difference of denoised performance between each thresholding rule.

4.2.1 Power Spectral Density(PSD)

The power spectral density function shows how the variance (energy) of the data is distributed as a function of frequency and it reveals the details regarding which frequency the energy is high or low. The PSD was calculated by Matlab using Fast Fourier Transform (FFT). Fig. 4.2 and Fig. 4.3 showed the PSD of the raw data and the PSD after processing. From the figures, it can be seen that both the trend signal and the white noise perturbations are suppressed.

4.2.2 Root Mean Square Difference(RMSD)

The computation of RMSD is defined as:

$$RMSD = \sqrt{\frac{[x - \hat{x}]^2}{[x]^2}} \tag{4.1}$$

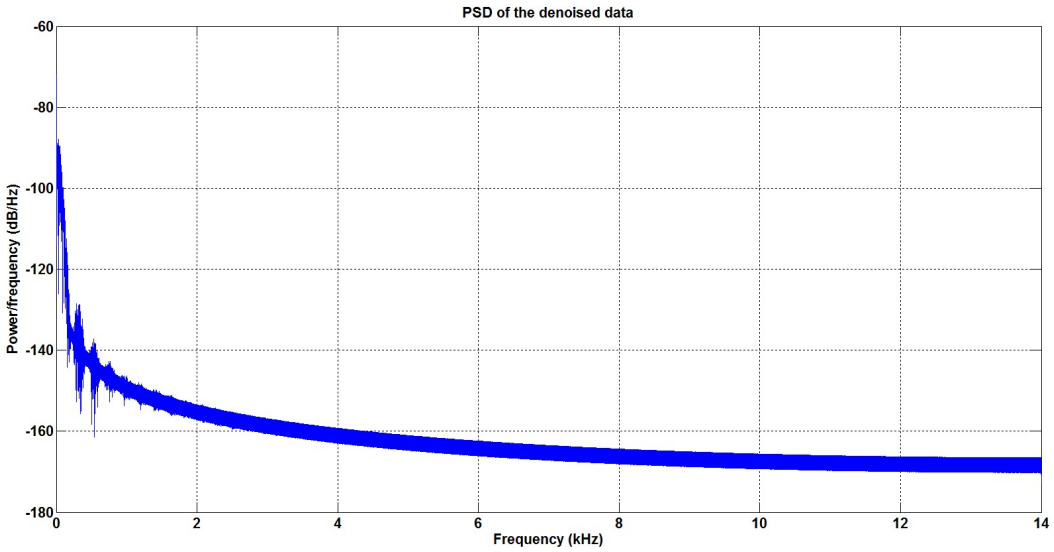

Figure 4.3 The PSD of wavelet denoised data.

where the x is the raw data before denoising, and $\hat{x}$ is the signal after wavelet denoising. This parameter is used for measuring the difference between the raw data and denoised data, so the larger the RMSD is, the better performance it has.

The Table 4.2 shows the RMSD value for the four different thresholding rules by performing Monte Carlo Testing 1000 times on 10 sets of data samples. The level of decomposition in this discussion is chosen by empirical experiments and remains the same during the comparison. For ECG signal, the soft thresholding technique is proven by many researchers to have better performance over the hard thresholding technique. In this work, we analyzed both soft and hard thresholding techniques to test whether this conclusion still holds in cytometry data. From Table 4.3, it can be concluded that the minimaxi thresholding rule is slightly better than the universal threshold and gives the best performance. Compared with hard thresholding, the soft thresholding performs better.

Table 4.2 Selection of thresholding rule for wavelet denoising the cytometry signals using RMSD.

Thresholding Rules/RMSD	Rigrsure	Heursure	Sqtwolog	Minimaxi
Data Sample Set	Soft/Hard	Soft/Hard	Soft/Hard	Soft/Hard
1	1.0006/1.0006	1.0006/1.0005	1.0009/1.0003	1.0008/1.0006
2	1.0005/1.0005	1.0005/1.0005	1.0006/1.0005	1.0006/1.0005
3	1.0003/1.0003	1.0003/1.0003	1.0002/1.0001	1.0000/1.0002
4	1.0005/1.0005	1.0005/1.0005	1.0004/1.0004	1.0004/1.0005
5	1.0005/1.0006	1.0006/1.0005	1.0007/1.0006	1.0007/1.0007
6	1.0005/1.0005	1.0005/1.0005	1.0005/1.0004	1.0005/1.0005
7	1.0005/1.0004	1.0004/1.0004	1.0005/1.0005	1.0006/1.0005
8	1.0003/1.0004	1.0004/1.0004	1.0004/1.0004	1.0004/1.0004
9	1.0003/1.0004	1.0004/1.0004	1.0003/1.0004	1.0003/1.0002
10	1.0002/1.0003	1.0003/1.0003	1.0005/1.0004	1.0005/1.0003
Average	1.00042/1.00045	1.00045/1.00043	1.00048/1.00040	1.00050/1.00044

Table 4.3 Performance comparison by RMSE.

5 Methods	Butterworth LPF	Butterworth HPF	Chebyshev LPF	Chebyshev HPF	Wavelet
No Noise Added	0 %	0 %	0%	0%	0%
SNR=20dB	0%	2.0408%	0%	4.08%	0.20%
SNR=16dB	0.3535%	4.3364%	0.2886%	8.8341%	0.81%
SNR=13dB	1.8139%	9.8489%	1.7571%	10.4637%	1.47%
SNR=10dB	10.1265%	22.6204%	8.3136%	21.1858%	2.24%

Chapter 5

CONCLUSIONS

By taking 187 peaks from 200,000 sample points in three data sets (A, B, C) to evaluate the detrending effects from the 5 methods (BLPF, CLPF, BHPF, CHPF, Wavelet), the percentages of differences between the detrended data and the raw data are averaged over the number of peaks in each data set (see Table 5.1). It suggests that wavelet detrending has the least effect on both amplitude and area.

In this thesis, the procedure of extracting the cytometry data information from the noisy electronic biomarker detection device based on filters and Discrete Wavelet Transform was discussed. In order to quantify the detrending and denoising performance, several parameters including Standard Deviation, Cross-Correlation Coefficient, Root Mean Square Error, Power Spectrum Density and Root Mean Square Difference were employed. For conventional filtering methods, the Chebyshev lowpass filter gives the best performance. For wavelet denoising method, a procedure of selecting the optimal wavelet filter was presented. Results showed the biorthogonal 3.1 filters are optimal to use in cytometry data analysis. Similar to the results in ECG signal processing,

Table 5.1 Effects evaluation on peak amplitudes and area.

5 Methods	Amp (A)	Area (A)	Amp (B)	Area (B)	Amp (C)	Area (C)	Ave Amp	Ave Area
BLPF	23.36%	47.27%	4.370%	17.22%	4.190%	16.0%	10.64%	26.83%
CLPF	22.59 %	46.40 %	1.970 %	14.77 %	2.390%	14.20%	8.983%	25.12%
BHPF	30.90%	60.18%	11.99%	48.95%	13.00%	21.00%	18.63%	43.38%
CHPF	34.18%	63.94%	31.15%	64.95%	26.00%	32.80%	30.44%	53.89%
Wavelet	2.020%	27.45%	0.700%	7.050%	1.670%	8.100%	1.463%	14.2%

the soft thresholding technique is better than the hard thresholding technique. Both the universal thresholding rule and the minimax thresholding rule are excellent for the purpose of denoising. However, the minimax thresholding rule performs slightly better. The procedure discussed in this work can be applied as guidance for the future analysis of impedance cytometry data.

Chapter 6

APPENDIX

6.1 The Non-stationarity

From the statistical perspective, the non-stationary series will have an Auto Correlation Function (ACF) that has long existing residue and remains outstanding for six lags or even more, rather than declining to zero quickly. By calculating the ACF of the raw data, the result is shown in Fig. 6.1; it remains significantly even after 20 lags, which suggests the data is non-stationary. By conducting an Augmented Dickey-Fuller Test against a trend-stationary alternative, augmenting the model with 0, 1 and 2 lagged difference terms. The result is: 0, 0 and 1, which suggests that if we applied a second order difference to the raw data sample, the residual is stationary. The results in Fig. 6.2 show the ACF of the first order difference data still remains large, while in second order differenced data decays fast.

6.2 The Choice of Threshold

From the histogram in Fig. 6.3 it can be seen that the data is mainly focused on $0.2e^{-3}$ $0.4e^{-3}$; which makes the average peak amplitude locate in the same region. And one third of the average peak amplitude would be a strong condition for the peak search height. If we imagine there will be coincide between the threshold and peaks, then threshold ($AveP/3$) is between $0.2e^{-3}$ $0.4e^{-3}$, which will make the Ave P to be

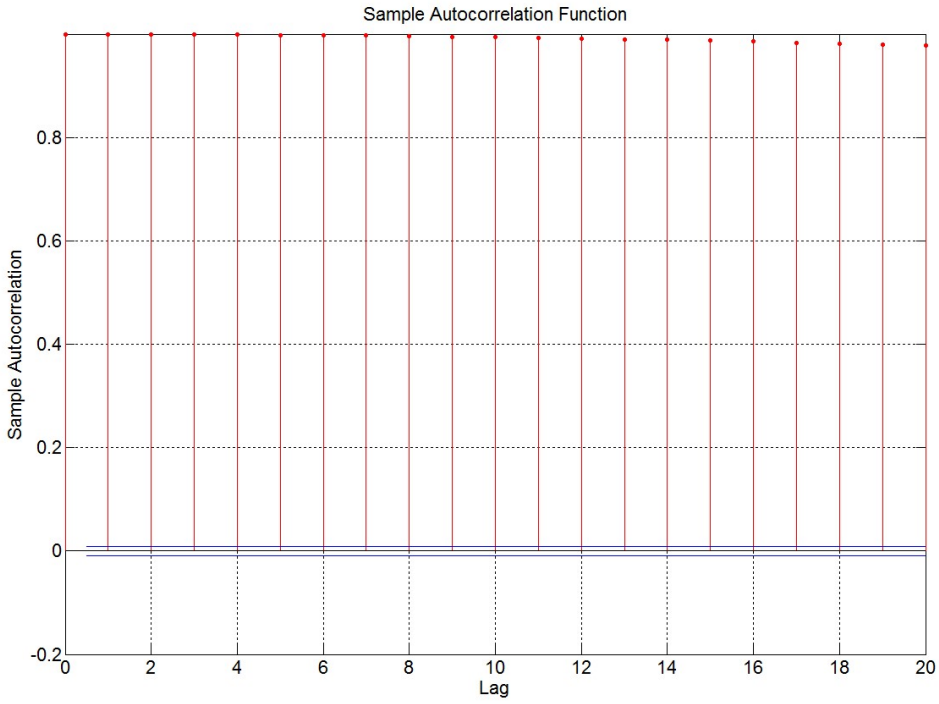

Figure 6.1 ACF of the raw data.

$0.6e^{-3}$ $1.2e^{-3}$. And this is inconsistent with the fact that the average peak amplitude will be in the $0.2e^{-3}$ $0.4e^{-3}$ region. So this suggests that the threshold ($AveP/3$) is good enough, and the later experiments from 1M sample points shown in Fig. 6.4 and Fig. 6.5, all the true peaks are above this threshold.

6.3 pClamp10 Software Comparison

pClamp 10 is a commercial software designed for electrophysiology data acquisition and analysis experiments. In the last 10 years, nearly 3,000 papers have been published by using pClamp 10 software to analyze their data.

This software is integrated with filter design and spectrum analysis package. It also provides the auto baseline correction option which is compared with methods in this paper. The peak search algorithm is defined by giving a universal threshold, and picks

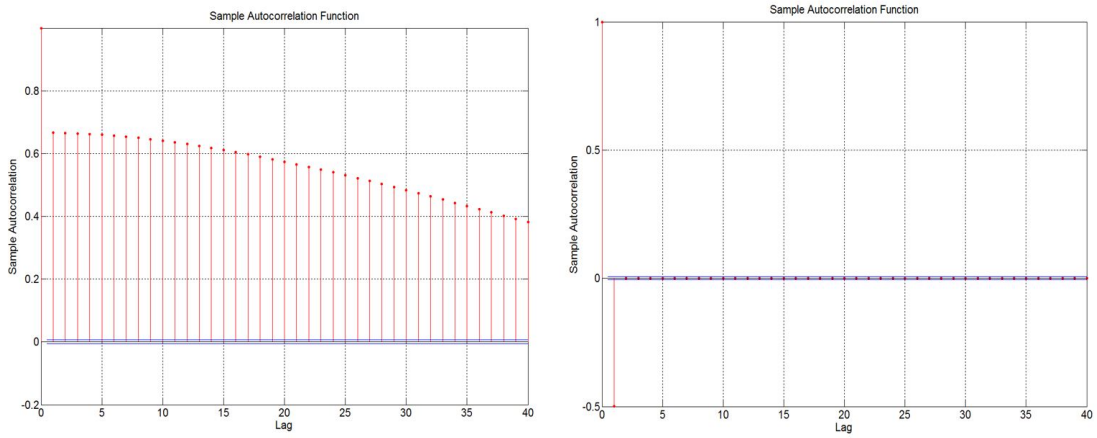

Figure 6.2 ACF of the first order differenced data (left) and second order differenced data (right).

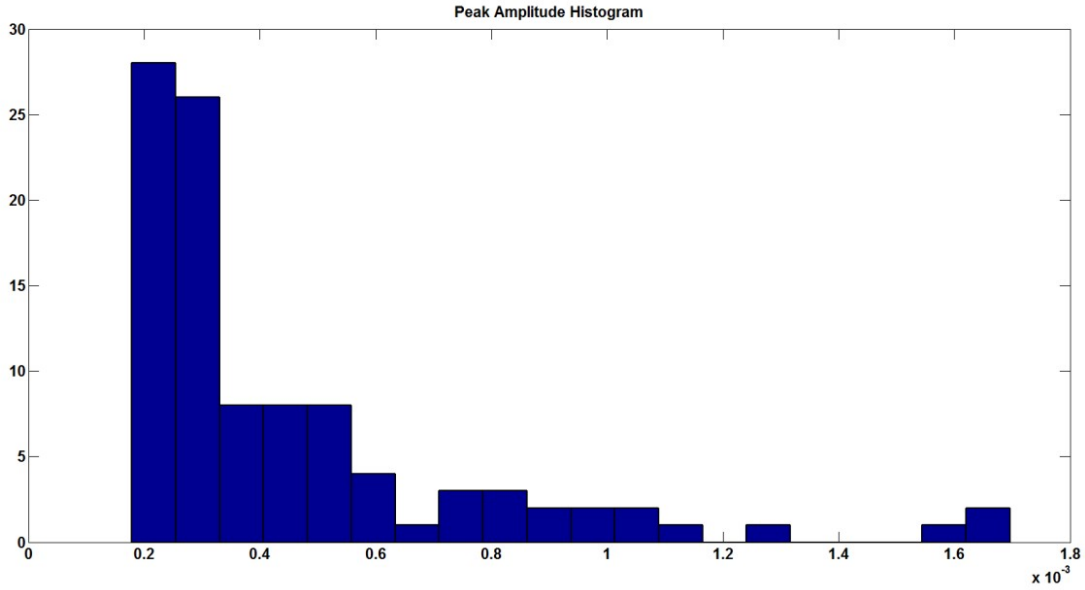

Figure 6.3 Histogram of the dataset.

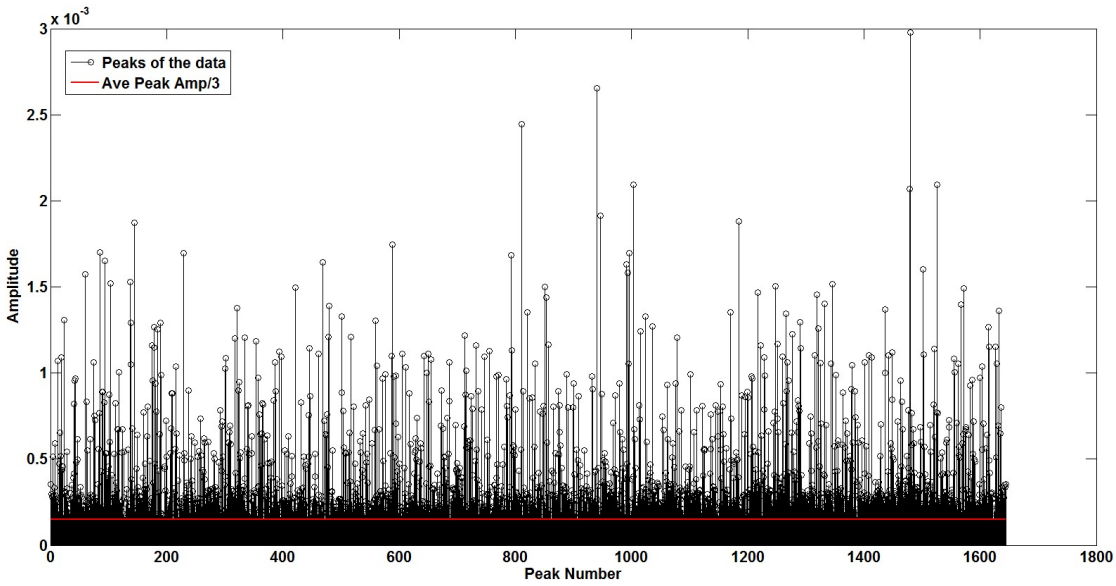

Figure 6.4 True peaks and threshold (red line) comparison.

Table 6.1 Effects evaluation on peak amplitudes and area.

6 Methods	Ave Amp	Ave Area	Error Rate at Noise-Free
BLPF	10.64%	26.83%	0%
CLPF	8.983%	25.12	0%
BHPF	18.63%	43.38%	0%
CHPF	30.44%	53.89%	0%
Wavelet	1.463%	14.2%	0%
pClamp10	20.663%	32.6%	14.3%

the local maximum value above this threshold, which is similar to the findpeak function in Matlab.

This integrated peak search process will automatically reject the subpeak near the main peak which will cause errors as shown in Fig. 6.5. If we load the detrended data to Matlab, and use the peak search process introduced before, the result is in Fig. 6.6, and brings more than 60% error rate. And it shows that the detrending process will cause a serious sawtooth effect to the dataset and result in difficulties as well as huge error rates for peak search process. In Table 6.1, all the methods together are compared, and wavelet method results in 19.2% improvement in amplitude preservation, 18.4% improvement in area preservation.

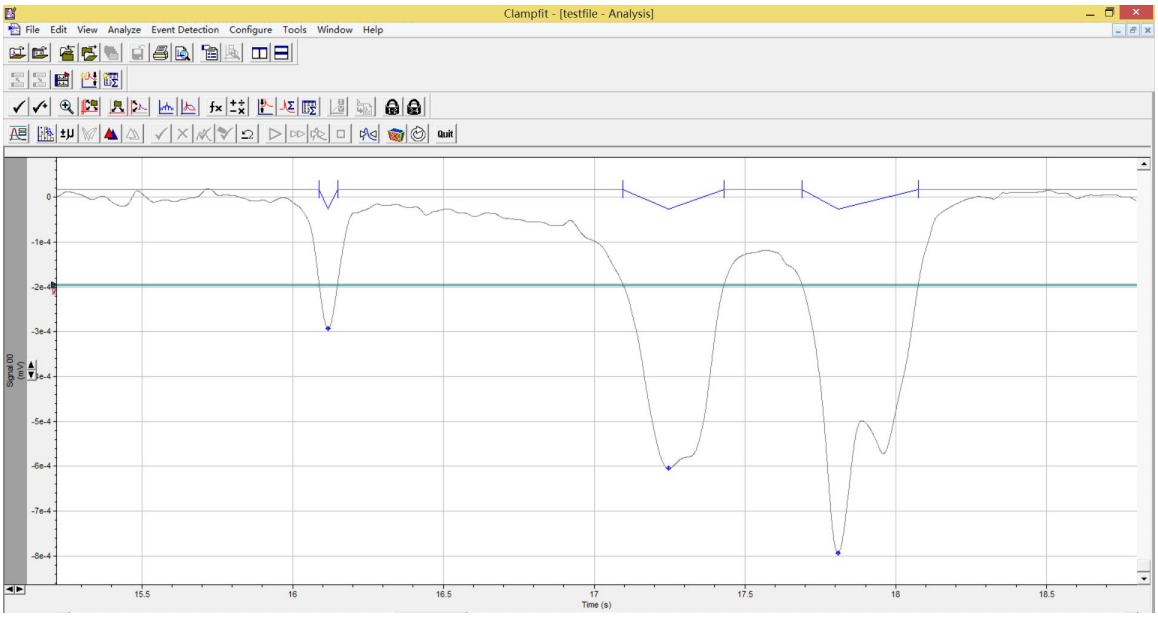

Figure 6.5 pClamp 10 peak search analysis.

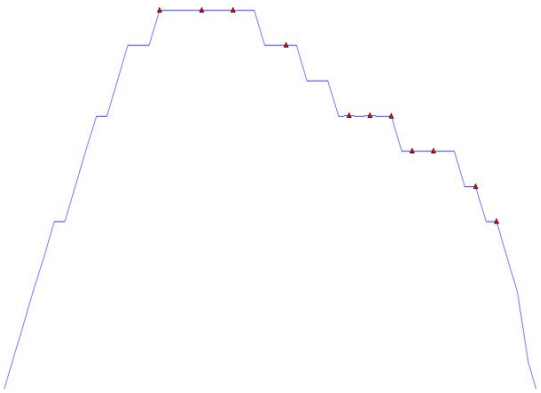

Figure 6.6 Sawtooth effect caused by pClamp 10 auto baseline correction.

Bibliography

[1] Brij N Singh and Arvind K Tiwari. Optimal selection of wavelet basis function applied to ecg signal denoising. *Digital Signal Processing*, 16(3):275–287, 2006.

[2] Zhongtian Lin, Pengfei Xie, Xinnan Cao, and Mehdi Javanmard. Electronic quantification of protein biomarkers based on bead aggregate sizing.

[3] Nick A Zilmer, Mahesh Godavarti, Jeffrey J Rodriguez, Timothy A Yopp, Georgina M Lambert, and David W Galbraith. Flow cytometric analysis using digital signal processing. *Cytometry*, 20(2):102–117, 1995.

[4] Shuxiang Guo, Fang Wu, Wei Wei, Jian Guo, Yuehui Ji, and Yunliang Wang. A novel wavelet denoising method used for droplet volume detection in the microfluidic system. In *Mechatronics and Automation (ICMA), 2013 IEEE International Conference on*, pages 1732–1737. IEEE, 2013.

[5] David Damm, Chaofeng Wang, Xunbin Wei, and Axel Mosig. Cell counting for in vivo flow cytometer signals using wavelet-based dynamic peak picking. In *Biomedical Engineering and Informatics, 2009. BMEI'09. 2nd International Conference on*, pages 1–4. IEEE, 2009.

[6] HG Rodney Tan, AC Tan, PY Khong, and VH Mok. Best wavelet function identification system for ecg signal denoise applications. In *Intelligent and Advanced Systems, 2007. ICIAS 2007. International Conference on*, pages 631–634. IEEE, 2007.

[7] S Poornachandra. Wavelet-based denoising using subband dependent threshold for ecg signals. *Digital signal processing*, 18(1):49–55, 2008.

[8] Brij N Singh and Arvind K Tiwari. Optimal selection of wavelet basis function applied to ecg signal denoising. *Digital Signal Processing*, 16(3):275–287, 2006.

[9] Mikhled Alfaouri and Khaled Daqrouq. Ecg signal denoising by wavelet transform thresholding. *American Journal of applied sciences*, 5(3):276–281, 2008.

[10] Manuel Blanco-Velasco, Binwei Weng, and Kenneth E Barner. Ecg signal denoising and baseline wander correction based on the empirical mode decomposition. *Computers in biology and medicine*, 38(1):1–13, 2008.

[11] Hongjun Song, Yi Wang, Jenna M Rosano, Balabhaskar Prabhakarpandian, Charles Garson, Kapil Pant, and Eva Lai. A microfluidic impedance flow cytometer for identification of differentiation state of stem cells. *Lab on a Chip*, 13(12):2300–2310, 2013.

[12] Karl Hülber, Michaela Sonnleitner, Ruth Flatscher, Andreas Berger, Rainer Dobrovsky, Sophie Niessner, Thomas Nigl, Gerald M Schneeweiss, Magdalena Kubešová, Jana Rauchová, et al. Ecological segregation drives fine-scale cytotype distribution of senecio carniolicus in the eastern alps. *Preslia*, 81(3):309, 2009.

[13] Shady Gawad, Laurent Schild, and Ph Renaud. Micromachined impedance spectroscopy flow cytometer for cell analysis and particle sizing. *Lab on a Chip*, 1(1):76–82, 2001.

[14] Paul L Biancaniello, Anthony J Kim, and John C Crocker. Long-time stretched exponential kinetics in single dna duplex dissociation. *Biophysical journal*, 94(3):891–896, 2008.

[15] Kaori Yamakawa, Masahiro Matsunaga, Tokiko Isowa, Kenta Kimura, Kunio Kasugai, Masashi Yoneda, Hiroshi Kaneko, and Hideki Ohira. Transient responses of inflammatory cytokines in acute stress. *Biological psychology*, 82(1):25–32, 2009.

[16] Roberta Valenti, Veronica Huber, Paola Filipazzi, Lorenzo Pilla, Gloria Sovena, Antonello Villa, Alessandro Corbelli, Stefano Fais, Giorgio Parmiani, and Licia Rivoltini. Human tumor-released microvesicles promote the differentiation of myeloid cells with transforming growth factor-β–mediated suppressive activity on t lymphocytes. *Cancer research*, 66(18):9290–9298, 2006.

[17] Mehdi Javanmard, Amirali H Talasaz, Mohsen Nemat-Gorgani, Fabian Pease, Mostafa Ronaghi, and Ronald W Davis. Electrical detection of protein biomarkers using bioactivated microfluidic channels. *Lab on a Chip*, 9(10):1429–1434, 2009.

[18] Stephen Butterworth. On the theory of filter amplifiers. *Wireless Engineer*, 7(6):536–541, 1930.

[19] Wikipedia. Chebyshev filter — wikipedia, the free encyclopedia, 2015. [Online; accessed 1-September-2015].

[20] P Karthikeyan, M Murugappan, and S Yaacob. Ecg signal denoising using wavelet thresholding technique in human stress assessment. *International Journal on Electrical Engineering and Informatics*, 4(2):306–319, 2012.

[21] David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[22] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(7):674–693, 1989.

[23] Martin Vetterli. Wavelets, approximation, and compression. *Signal Processing Magazine, IEEE*, 18(5):59–73, 2001.

[24] Hua Cui and Guoxiang Song. Study of the wavelet basis selections. In *Computational Intelligence and Security*, pages 1009–1017. Springer, 2007.

[25] Aaushi Kumari and Megha Bisht. Optimal wavelet filter maximizes the cross correlation coefficient with an ecg signal. *IJITR*, 1(2):191–193, 2013.