University of Alberta

SIGNAL PROCESSING AND AMPLIFIER DESIGN FOR INEXPENSIVE GENETIC ANALYSIS INSTRUMENTS

by

Sheng Heng Choi

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer, Microelectronic Devices, Circuits and Systems

Department of Electrical and Computer Engineering

Fall 2011
Edmonton, Alberta

Examining Committee

Dr. Chris Backhouse, Electrical and Computer Engineering

Dr. Vincent Gaudet, Electrical and Computer Engineering, University of Waterloo

Dr. Bruce Cockburn, Electrical and Computer Engineering

Dr. John Bowman, Mathematics and Statistics

To my wife and my parents

Abstract

The Applied Miniaturisation Laboratory (AML) has recently built a laser-induced fluorescent capillary electrophoresis (LIF-CE) genetic analysis instrument, called the Tricorder Tool Kit (TTK). By using a photodiode instead of photomultiplier tubes in the optical detection, the AML has lowered the cost and size compared to commercial LIF-CE products. However, maintaining an adequate signal-to-noise (SNR) and limit of detection (LOD) is a challenge.

By implementing a multistage amplifier, we increased the bandwidth and voltage swing while maintaining the transimpedance gain compared to the previous design. We also developed signal processing algorithms for post-experiment processing of CE. Using wavelet transform, iterative polynomial baseline fitting, and Jansson's deconvolution, we improved the SNR, reduced baseline variations, and separated overlapping peaks in CE signals. By improving the electronics and signal processing, we lowered the LOD of the TTK, which is a step towards the realisation of inexpensive point-of-care molecular medical diagnosis instruments.

Acknowledgements

I would like to thank Dr. Chris Backhouse and Dr. Vincent Gaudet for being my supervisors and guiding me through my thesis project. They have taught me valuable skills to become a better researcher, engineer and writer. They always encouraged and challenged me to ask the most important question – "Why?" I also thank Dr. Bruce Cockburn and Dr. John Bowman for serving on my thesis committee.

I would like to thank Govind Kaigala for introducing me to the Applied Miniaturisation Laboratory, Mohammad Behnam for his help with electronics, Allison Bidulock for her capillary electrophoresis data and biology expertise, Shane Groendahl for letting me use the laboratory equipment, Sunny Ho for his constant (but fun) distractions, and Ayokunle Olanrewaju for his insights of the world and always finding the positives in every situation.

I also acknowledge the funding I received through National Science and Engineering Research Council, Alberta Ingenuity, and APEGGA. Their generous contribution has allowed me to focus on my research during my time at the university.

I would like to thank my wife Terrilyn Choi for her patience, support and companionship throughout my Masters degree. I am very grateful for my parents' encouragement and unconditional support.

Table of Contents

List of Tables

List of Figures

List of Symbols

Symbol	Definition
V	voltage
Ω	Ohm, SI unit of resistance
A	ampere, SI unit of current
D	duty cycle
F	Farad, SI unit of capacitance
V/V	voltage gain
Av	voltage gain
V/A	current to voltage gain
b	relaxation constant
$h(t)$	point spread function
Hz	Hertz
V_{RMS}	voltage in root mean square
e_n	input referred noise
e_w	white noise specification
f_{nc}	noise corner frequency
f_{3db}	3dB corner frequency
R_f	feedback resistor
C_f	feedback capacitor
$\psi_{j,k}(t)$	wavelet function
ρ	error factor
ρ_{thr}	error factor threshold
k_{max}	maximum iteration count
n	polynomial degree

List of Abbreviations

Abbreviation	Definition
ADC	analog-to-digital converter
AML	Applied Miniaturisation Laboratory
bp	base pairs
CE	capillary electrophoresis
CMOS	complementary metal oxide semiconductor
db	Daubechies
DMM	digital multimeter
DNA	deoxyribonucleic acid
DWT	digital wavelet transform
ENB	equivalent noise bandwidth
ETG	empirically transformed Gaussian
FIR	finite impulse response
FT	Fourier transform
FSD	Fourier self-deconvolution
FWHM	full width at half maximum
IPBF	iterative polynomial baseline fit
LIF	laser induced fluorescents
LPA	linear polyacrylamide
LPF	low pass filter
LOC	lab-on-chip
LOD	limit of detection
JFET	junction gate field-effect transistor
MA	moving average
MCU	micro-controller unit
NEP	Noise Equivalent Product

Abbreviation	Definition
PBR	peak to baseline variation ratio
PCR	polymerase chain reaction
PDMS	polydimethylsiloxane
PDR	peak degradation ratio
PHE	peak height error
PMT	photomultiplier tubes
POC	point-of-care
PRD	peak region detection
PSE	peak shift error
PSF	point spread function
R	resolution
RMSE	root mean square error
RT	region threshold
RTM	region threshold multiplier
SG	Savitzky Golay
SNR	signal-to-noise ratio
SURE	Stein's unbiased risk estimate
sym	Symlet
T	threshold
TTE	Tris TAPS-EDTA
TTK	Tricorder Tool Kit
GBP	gain bandwidth product
WD	wavelet denoising
WT	wavelet transform
ZT	zone threshold
ZTM	zone threshold multiplier
μTK	Microfluidic Tool Kit

Chapter 1

Introduction

Long test times and the high costs associated with molecular medical diagnoses are limiting factors for their widespread usage in our society. Molecular diagnoses are currently done in batches at large centralized laboratories equipped with expensive instruments. To complete a diagnosis, samples are processed in multiple stages and each stage requires a specialized instrument. Due to the complexity of these instruments and the protocols required, highly trained personnel are needed to perform these diagnoses. Consequently, molecular diagnoses are expensive to perform and test results have long turnaround times.

There is a strong need for inexpensive, portable, and fast molecular diagnosis instruments in clinical and hospital settings [1]. Such technology would allow for accurate diagnosis and treatment of potential diseases, virus pandemics, or cancer development. If rapid molecular diagnoses were available, then pandemics could be more easily monitored and controlled by providing suitable treatments. In the recent H1N1 outbreak, medical staff were advised to treat every patient who had flu-like symptoms with antiviral drugs to combat (H1N1 and regular) influenza. These antiviral drugs can have adverse side effects such as nausea, hallucination, and seizure [2]. Antiviral drugs were used on all patients with flu-like symptoms because testing for the H1N1 virus requires the use of real-time polymerase chain reaction (PCR). Treatments for influenza may be ineffective if delayed until labora-

tory test results are returned.

Molecular diagnosis can also help with disease prevention. The Google co-founder Sergey Brin, one of the richest people in the United States, recently underwent a genetic test that revealed a mutation in his LRRK2 gene [3]. Depending on the evaluation metrics used, the results from the genetic test indicated that he has a 20 - 80% chance of getting Parkinson's disease later in life. Genetic testing allowed Mr. Brin to take preventative measures for Parkinson's disease.

Molecular diagnosis can identify cancer – early diagnosis is a key factor in controlling this disease. According to Yager *et al.*, 73% of lung, 57% of colorectal, and 34% of breast cancer patients are diagnosed at a later stage of cancer development [4]. But if cancer is not diagnosed and treated at the beginning of development, the survival rate is below 15% [5, 6].

If molecular diagnoses were offered in a point-of-care (POC) format, inexpensive, portable, and on-site, the health care system could provide treatments tailored to specific needs. For this purpose, research on the application and infrastructure of lab-on-chip (LOC) technology has been on the rise for the last two decades.

The Applied Miniaturisation Laboratory (AML), supervised by Dr. Christopher Backhouse, recently integrated and miniaturized microfluidics, microelectronics, and optics into an instrument that can perform molecular diagnosis. One of the central challenges in such LOC instruments is in maintaining an adequate signal-to-noise ratio (SNR) and limit of detection (LOD) in the detection subsystems as the instruments are miniaturised. In the present thesis, these challenges are addressed by improving the performance of the detection electronics, and by applying signal processing techniques to remove electronic and instrumental artifacts produced in the detection process.

1.1 Thesis Overview

Chapter 2 provides a brief background on microfluidics, standard molecular diagnosis, and capillary electrophoresis (CE). We identify the type and source of noise in the current detection circuitry designed and built by AML. We also review signal processing algorithms that are used to extract information in CE signals.

In Chapter 3, a multistage amplifier is designed and tested for optical detection and amplification of the signal captured by a photodiode in the TTK during CE experiments. This new multistage amplifier improves the bandwidth and voltage swing compared to the previous single-stage amplifier design. We believe that the increase of bandwidth will reveal additional signal information in genetic analysis.

While we are improving the capabilities of our electronic circuits, we are also investigating in this thesis the use of signal processing techniques to better extract signals from recorded measurements compared to current methods used in the TTK. Chapter 4 describes the wavelet transform and how it can be used to remove noise in CE signals. A method is developed to determine a set of reliable wavelet denoising parameters for various synthetic and experimental CE signals. The performance of wavelet denoising is compared to other noise removal methods such as moving average (MA), Savitzky-Golay (SG) smoothing, and low-pass filtering (LPF).

In Chapter 5, a baseline variation removal algorithm that consists of peak region detection and an iterative polynomial baseline fit is implemented and tested.

Chapter 6 introduces a modified parametric Jansson's deconvolution method. This algorithm is used to separate overlapped peaks in CE signals.

In Chapter 7 a complete CE signal processing algorithm is presented and its performance on experimental CE signals is explored.

The project conclusion and future improvements to this thesis work are presented in Chapter 8.

Chapter 2

Background

2.1 Microfluidics

Microfluidic technology possesses key functionalities for the implementation of point-of-care (POC) molecular diagnostic systems. The ability to manipulate and analyse small samples of fluids is popular in molecular biology diagnosis because it greatly reduces reagent usage and analysis time [7]. Microfluidic technology has the ability to perform complex chemical and biological reactions, typically done in laboratories, without the intervention of an expert operator. Thus the term lab-on-a-chip (LOC) has been coined for microfluidics and the terms are used interchangeably in literature and in this thesis. With the impressive advances in microfluidics in the last two decades, molecular diagnostic devices have been reduced in size and cost while increases in throughput, sensitivity, and applications have been achieved [8–10]. With supporting equipment, microfluidic technology can be used to diagnose medical conditions such as herpes simplex viral infection, gene mutations, and muscular dystrophy [11].

2.2 Standard Molecular Diagnosis

A standard molecular diagnosis requires three steps as shown in Fig. 2.1 [12]. The first step is sample preparation, where genetic information such as deoxyribonucleic

acid (DNA) is extracted and purified from raw samples such as blood or urine. The amount of genetic information must then be increased before detection. A popular choice for DNA amplification is the polymerase chain reaction (PCR). The last step in molecular diagnosis is genetic detection. A method that can be employed to perform genetic analysis is electrophoresis. This thesis focuses on improving the signal-to-noise ratio (SNR), resolution (R) and limit of detection (LOD) of genetic detection in a molecular diagnosis.

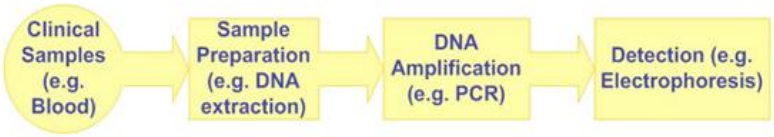


Figure 2.1: Standard steps for molecular diagnosis.

Electrophoresis is the most widely used method for the detection of DNA and protein because it is highly sensitive, flexible, and can be integrated with other molecular biology protocols [11]. Among various electrophoresis methods, capillary electrophoresis (CE) has become a standard due to its speed and low sample requirement [13–17].

Biological components in solution can be separated based on their mobility differences when an electric current is applied. This is the principle of CE separation. In a fluid mixture, the velocity of a particle is dependent on its charge and size; thus an electric field can be used to separate smaller molecules from larger molecules. The separation channel is filled with a polymer sieving matrix and a buffer solution. Since longer DNA fragments are bigger and have a lower electrophoretic mobility than shorter DNA fragments, they pass more slowly through the sieving matrix [18]. The buffer solution provides ions to move the molecules along the channel when an electric field is applied [19].

CE detection techniques such as electrochemical detection [20], mass spectrometry [21], and laser-induced fluorescence (LIF) [22] have been developed. Among

these methods, LIF is the technique used most often because of its sensitivity, reliability, and compatibility with other biological systems [16, 23–28].

LIF works as follows: PCR produces multiple copies of product DNA (typically millions or more) of a specific small region of the sample DNA. The sample DNA may consist of billions of bases, and the product DNA is typically on the order of 300 bases long. The PCR process can be used to fluorescently label the product DNA so that it is readily detectable by fluorescence at the near-single copy level. By applying an electric field, the movement of the DNA can provide information on the size and sequence of the product DNA (and hence also upon a specific region of the sample sequence). From this information (an electropherogram), a clinical diagnosis can often be made.

2.3 Miniaturisation Challenges

Significant advances in applications using microfluidic technology have been made in recent years [8], but large supporting equipment are required to facilitate the operations of microfluidic chips [29]. The expense and size of the these supporting equipment have limited the application of the technology.

2.3.1 Commercial Capillary Electrophoresis Instruments

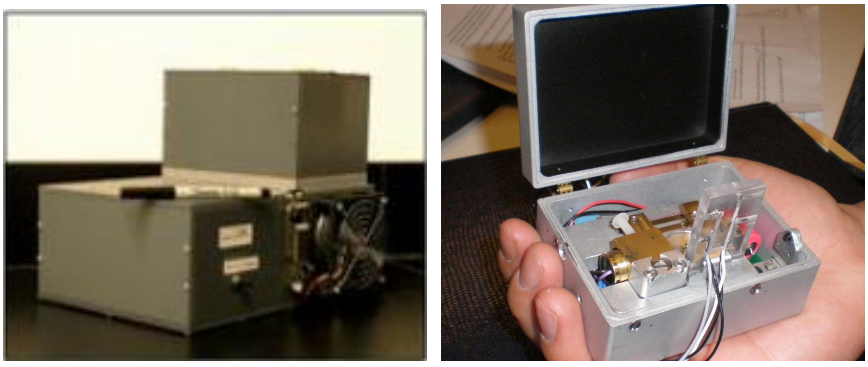
Commercial CE products, such as the μTK (Micralyne) and the Biofocus 3000 (Biorad) use a combination of a high-gain photomultiplier tube (PMT) and confocal optics for LIF. A confocal optic system focuses and captures the exact location where the DNA emits fluorescence, which is a very small area of the microfluidic chip [30]. This method eliminates most of the scattered light of the microfluidic chip and therefore has low baseline offset and variation. Although these instruments have impressive throughput and sensitivity, they are large, expensive, and cannot perform the entire molecular diagnosis. For these reasons, they are not suited for automated POC applications.

If POC molecular diagnoses were available, samples would not have to be sent to off-site laboratories and diagnoses could be done on-site. This would allow for quick diagnoses and treatments of medical conditions. In order to provide POC molecular diagnosis, the support equipment must be portable, inexpensive, automated, easy to use, and reliable [1, 31]. Miniaturisation and integration of optics, microfluidics, and microelectronics can eliminate expensive facilitating equipment; however, as devices shrink in size, integration becomes more complex and new challenges arise.

2.3.2 Optical Setup and Noise

To address the cost and size challenges in miniaturising commercial CE systems, AML has recently demonstrated that the building cost of a LIF-CE instrument can be reduced by using a photodiode, an interference filter, and a gradient index lens instead of a confocal optical setup [29, 32, 33]. Fig. 2.2 shows two instruments capable of performing LIF-CE, designed and built by AML. Photodiodes are suitable for POC applications because they are low cost, small, durable, and compatible with existing complementary metal oxide semiconductor (CMOS) technology (for future integration). One problem with this non-confocal approach is that the area of detection is much larger. Hence, more scattered light is captured off the microfluidic chip walls, causing a higher baseline.

A light source with constant intensity is required for a LIF-CE system. Laser diodes are suitable for POC applications because they are inexpensive and compact. However, laser diodes exhibit a phenomenon called mode hopping, where the wavelength of the emitted light varies slightly. According to a report by Heumier and Carlsten, mode hopping is caused partially by temperature and injection current variations [34]. As a result, the output intensity of laser diodes often fluctuates. In the TTK, slight laser intensity fluctuation due to temperature variation could cause baseline variations as high as the signal peaks [29]. This is because non-confocal



(a) A complete molecular diagnosis platform [29].

(b) A compact instrument capable of performing CE using a custom high voltage generating integrated circuit [32].

Figure 2.2: Non-confocal LIF CE instruments designed and built by AML.

systems like the TTK captures light of a large area and the baseline is very sensitive to laser intensity fluctuations. The authors of [29] minimized the temperature variation by attaching a large metal slab onto the laser diode to increase the temperature coefficient. The metal slab combined with a warm up period of 10 minutes reduced the light intensity fluctuation by 66%.

Although the authors of [29] successfully lowered the laser intensity variations by attaching a large metal slab and warming up the laser, these methods are not permanent solutions. The stability and performance are different for every laser diode and the same strategy will have a different effect on each laser diode. Also, long continuous warm up periods shorten a laser diodes's life span and a large metal slab adds unnecessary bulk to a POC genetic analysis system.

2.3.3 Microfluidics

Microfluidic devices were originally fabricated with silicon and glass substrates. To reduce cost, polymeric materials such as polydimethylsiloxane (PDMS) have been used to fabricate microfluidic chips [35]. A drawback is that polymer substrates typically emit higher autofluorescence than glass, which results in higher baseline

offset and variation.

The resolution of a CE diagnosis depends on the channel length and sieving matrix properties [36]. Peak resolution is defined as the ratio of the distance between two peaks and the average peak width [37]. Peak resolution decreases as the length of the microfluidic chip is reduced to improve compactness and separation speed. Low peak resolution causes peaks to overlap, limiting the number of CE applications.

2.3.4 Amplifier Noise and Bandwidth

Because emitted DNA fluorescence causes a photodiode to output a very low current (nA range), the TTK uses a single-stage transimpedance amplifier configured in negative feedback with a 1×10^{9} (G) Ω feedback resistor (R_f) and a 1×10^{-12} (p) F feedback capacitor (C_f) as shown in Fig. 2.3 [29, 32, 38]. This circuit is used to amplify and condition the signal. There are three major sources of noise associated with this design: shot noise, thermal noise, and interference noise. Shot noise originates from electrons (in electronics) and photons (in optical systems) that not only carry the signal but also carry energy which gives rise to fluctuations in measurements [39]. Photodiodes are PN junctions and therefore are prone to shot noise in low light conditions [40]. Thermal (or Johnson's) noise stems from electron movement due to thermal agitation [41]. Thermal noise causes all resistive elements to fluctuate in voltage as a function of temperature and resistance and increases as temperature or resistance increases. Thermal and shot noise have Gaussian probability functions and thus they are known as source of white noise [42]. Interference noise can arise from magnetic interference and digital switching. Noise generated from electrons (shot), heat (Johnson's), and interference generally has higher frequency than the signal; thus these types of noise are classified as high-frequency noise in this thesis.

Another problem with this single-stage amplifier is that its bandwidth (of 1.6

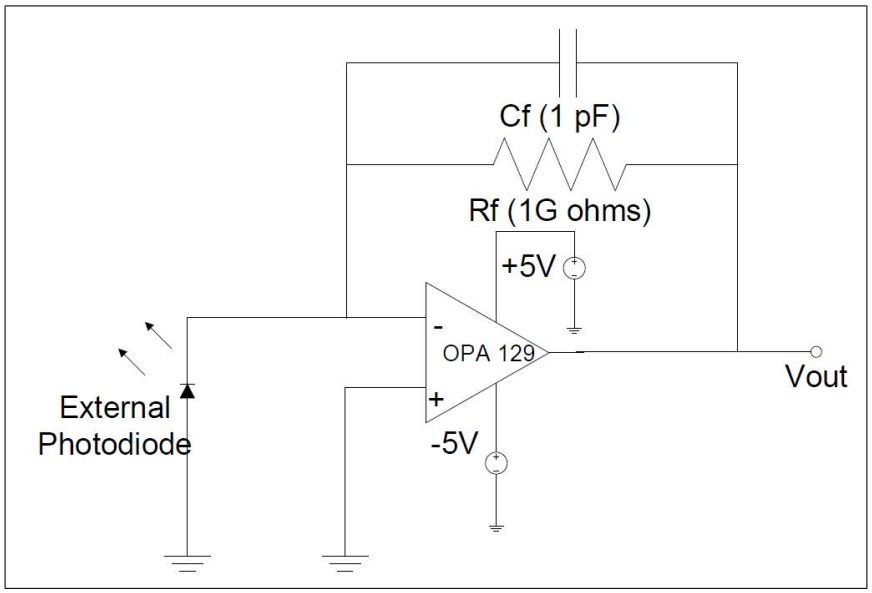


Figure 2.3: Transimpedance amplifier currently used in the TTK.

Hz) is severely limited by the op amp's gain bandwidth product (GBP). Because the bandwidth of this amplifier is low, it cannot provide insight on noise such as laser intensity fluctuations. In this thesis, we will design and test a multistage amplifier to be used for LIF-CE in the TTK. Such amplifier could provide us with more insight on the signal and noise.

2.4 Noise Sources

2.5 Signal Processing

We identified high-frequency noise, baseline variations, and low resolution as three key challenges to miniaturizing LIF-CE diagnostic instruments. Improvements in chemistry, microfluidics, optics, and microelectronics can solve these problems, but these improvements may increase design complexity, size and manufacturing costs. We take a different approach in this thesis to address these issues: in addition to working to improve the electronics in the TTK, we develop signal-processing algorithms to reduce baseline variations, increase the SNR, and lower the LOD for

Signal Condition	Cause	Method
High-frequency noise	Shot, thermal and interference noise in circuits	Wavelet denoising
Low-frequency baseline variation	Non-confocal photodiode-based detection and laser mode hopping	Iterative polynomial baseline fit and peak region detection
Low signal resolution	Short microfluidic channels	Jansson's deconvolution

Table 2.1: Signal conditions observed in the TTK and proposed methods

the TTK. Table 2.1 summarises the signal condition problems observed in the TTK and the signal processing methods used in this thesis to address these issues.

2.5.1 High-Frequency Noise

High-frequency noise in CE signals is commonly removed with Fourier low-pass filtering (LPF) [20], moving average (MA) [43,44] and Savitzky-Golay (SG) smoothing [45].

For many years, Fourier LPF has been the workhorse in signal processing in analytical chemistry and other fields [20, 46]. Based on the frequency spectrum of the Fourier transform (FT), LPF can be designed to remove the high-frequency noise. FT is useful for signals where signal peaks have the same shapes and widths. Because CE signal characteristics (peak widths, shapes, mean etc.) vary from signal to signal (non-stationary random processes and time dependent), FT is inefficient in removing noise for CE signals [47]. Furthermore, if a Fourier filter's cut off frequency is too low, it distorts the signal; but if it is too high, it does not remove noise well enough. Another disadvantage of Fourier filtering is that it produces ripples in the filtered signal when there are sharp edges in the signal [48].

MA is a technique used to remove random white noise. In MA, each data point in the signal is compared and averaged to its adjacent points. Higher order smooth-

ing decreases noise but increases the likelihood that the peaks are flattened [49].

SG filters were first published in 1964 [45]. Since then, SG filters have been popular in removing noise in CE signals [50,51]. Because SG filtering requires multiple parameters (polynomial degree and smoothing span), distortions in signals are very common when these parameters are not chosen correctly [47]. Vivo-Truyols *et al.* automated part of SG filtering by combining SG with the Durbin-Watson criterion [52]. In their method, the span length was calculated automatically for a pre-specified polynomial degree. However, a fundamental disadvantage with SG filtering remains: when there are random spikes (due to noise) in the signal, SG filters overcompensate by trying to fit the polynomial over the spike regions, which leads to ripples in the post-processed signal [48].

Recently, use of the wavelet transform (WT) to process analytical chemistry signals, including CE signals, has increased significantly. According to a study by Shao [53], WT became popular after Daubechies [54] and Mallat [55] published influential papers in 1989 on fast computational algorithms for wavelet basis functions. Perrin *et al.* were the first to use WT to remove noise (wavelet denoising) in CE signals [46]. Since then, WT have been shown to improve SNR better than FT [56] and to outperform both SG and FT in preserving peak shape and height in the post-processing signal [47, 57]. Since CE signals are non-stationary signals and localised in time, WT is superior to other high-frequency noise removal methods because wavelets have a time component in the statistics of the transformed signal [46].

The problem with wavelet denoising is that suitable choices of wavelet type, decomposition level, and other threshold settings depend on the signal type and noise level. For example, Perrin *et al.* found that the Haar wavelet yielded the lowest root mean square error (RMSE) between ideal and denoised signals. In another study, Liu *et al.* claimed that the Daubechies (db) 5 wavelet with level 6 decomposition resulted in the lowest root mean square error (RMSE) between the

denoised signal and the ideal signal [47]. Both Cao and Zhang claimed that the Symlet (sym) 4 wavelet with higher than level 5 decomposition gave satisfactory results for their application using a PMT-based instrument [58, 59]. No one has researched a set of wavelet denoising parameters suitable for CE signals recorded by non-confocal CE instruments. Our goal in this thesis is to find a set of wavelet denoising parameters suitable for removing the high-frequency noise in TTK's CE signals.

2.5.2 Baseline Variation Removal

A varying baseline is one of the most common problems in the measurement of CE signals [60]. Baseline removal is important because a stable reference for the background signal must be established to provide an accurate CE diagnosis. The main difference between baseline variation and actual peaks is that the frequency of the baseline variation is lower than the frequency of the peaks [53]. Ordinary Fourier frequency analysis cannot distinguish between the actual signal peaks and baseline variations because the spectral difference is small [60].

Methods such as median subtraction (MS) [61,62], FT [63,64], and WT [13,46, 47,65] have been reported to address the baseline variation problem in CE signals, but, as indicated by Schulze [66], these methods all have shortcomings. For example, MS fails to remove the baseline variations if there are wide or overlapping peaks in the signal. Fourier filtering cannot be automated because the frequency contributions of baseline variation and signal peaks vary from signal to signal. Schulze also argued that WT's baseline removal ability depends on the wavelet type and the decomposition level, and the optimal choice depends on the nature of the signal.

More recently, Gan *et al.* showed that an iterative polynomial baseline fitting (IPBF) method removed high baseline variations in CE signals [60]. For this reason, we decided to implement the IPBF in our signal processing algorithm. In their method, the beginning and end coordinates for the peak regions must be manually

specified, which is not possible when automatic analysis of unknown samples is required. Detection of peak regions is easy by visual inspection, but automatic detection of peak regions is much more complicated. Peak location detection research is extensive [19,52,67,68], but to our knowledge, peak region detection algorithms for CE signals have received little attention. Landers suggested that CE peak regions can be found by finding the zero crossing of the first derivative of the signal [69]. This approach works well in an ideal world where noise is absent. In reality, using the zero crossings of the first derivative alone to determine the peak region often leads to false peak region detection. To address this problem, we developed a peak region detection algorithm to assist IPBF in removing baseline variations from CE signals.

2.5.3 Overlapping Peak Separation

Low peak resolution is another challenge facing portable CE devices today. Low resolution leads to overlapping peaks, which hinders the ability to distinguish signal peaks. Resolution can be increased by increasing peak spacings, which effectively reduces peak widths. This can be accomplished by either lengthening the microfluidic chip channel or by using a longer CE separation time. However these options do not align with the goals of a POC molecular diagnostic tool.

Instead of modifying the CE system or protocol to improve resolution, post signal-processing of CE signals can separate overlapping peaks to increase resolution. The most common peak separation technique is curve fitting, but this can only separate overlapping peaks if noise is low because artifacts are easily created with the curve fitting method [13]. Initial estimates of the number of peaks, peak shape, and peak width are required for the curve fitting method [70, 71]. Fourier self-deconvolution (FSD) with a Weiner smoothing filter has been shown to separate overlapping peaks in [72–74]; however, this method also requires a high SNR because artifacts are easily created [75].

Various deconvolution algorithms have been developed for chromatography and infrared signals. For example, Olazabal *et al.* showed a method where signals were deconvolved using WT; but although peaks were effectively separated, the deconvolution shifted peak locations and distorted peak shapes from the original signals [13, 76].

Numeric deconvolution methods, such as Jansson's deconvolution, have been reported to separate overlapping peaks in chromatographic signals [77–81]. To our knowledge, no one has used Jansson's deconvolution to separate the peaks in CE signals. The main benefit of Jansson's deconvolution is that it does not require knowledge of the noise characteristics. The only parameters required are the impulse response of the system and the amplitude bound of the signal [82]. To increase the usability and the automation of Jansson's deconvolution in a CE signal processing algorithm, we incorporated the use of normalization, peak detection, and a deconvolution factor.

Chapter 3

Multistage Amplifier Design and Testing

3.1 Introduction

The AML has designed and built an inexpensive medical diagnostic instrument called the Tricorder Tool Kit (TTK). The TTK is capable of performing genetic amplification and detection. In the TTK's optical detection subsystem, the amplifier currently used to amplify the current from the photodiode is a single-stage operational amplifier (op amp) connected in (unbiased) photovoltaic mode (Fig. 3.1(a)) with a large feedback resistor of 1×10^{9} Ω and a feedback capacitor of 1×10^{-12} F. This circuit produces a gain of 1×10^{9} V/A. A problem with this single stage high-gain amplifier is that the bandwidth is severely limited by the gain bandwidth product (GBP) of the OPA129 op amp (Texas Instruments), which is 1 MHz [83]. With a low bandwidth amplifier, it is very difficult to distinguish between the signal and noise (such as laser intensity fluctuations). Photodiode can also be configured in photoconductive mode, where it is reverse biased with a voltage source. Amplifiers in this configuration tend to have higher speed but they have higher leakage current and noise.

The bandwidth could be increased by splitting the single-stage amplifier into multiple stages. A multistage amplifier has the potential to achieve higher gain

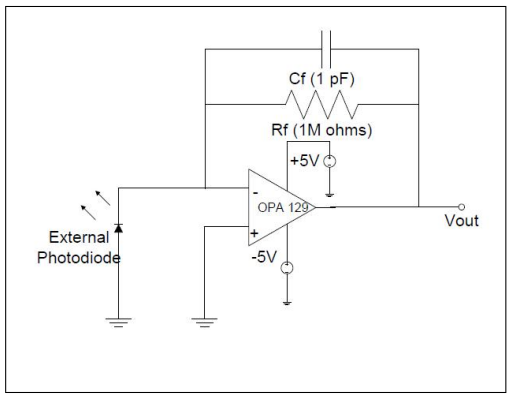
and bandwidth than a single-stage amplifier and reveal additional signal content. The passage of a DNA peak (of a given size) usually takes about a second and typical sampling rates for DNA detection are about 10 Hz to obtain all available peak information. Or goal is to design an amplifier with similar gain (1×10^9 V/A) as the current amplifier and increase the bandwidth to about 10 Hz. In this chapter we present the design and test of a multistage transimpedance amplifier that can be used in the TTK to convert photodiode current into a voltage for laser-induced fluorescent (LIF) - capillary electrophoresis (CE).

3.2 Multistage Amplifier Design

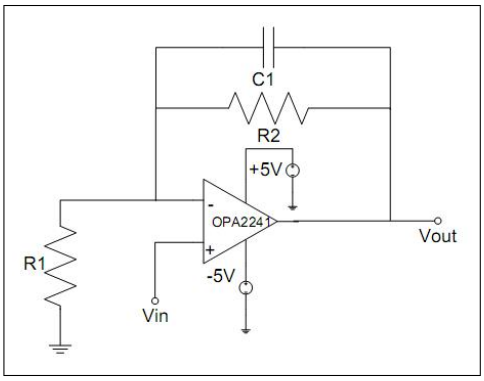
In our initial design of the multistage amplifier, as shown in Fig. 3.1, the first stage (Fig. 3.1(a)) is a transimpedance amplifier in photovoltaic mode. The purpose of the first stage is to convert the photodiode current into a voltage. It has the same configuration as the single-stage amplifier currently used in the TTK optics board, except a smaller feedback resistor R_f (1 MΩ) is used. The OPA129 op amp was chosen again because of its low input bias currents (100 fA max) which is required for low current applications. Because photodiodes are capacitive devices, they have a built-in capacitance (*Cin*). In the photodiode (NT57-506, Edmund Optics) used in the TTK, C_{in} is reported to be 4 pF [84]. *Cin* combined with R_f create a phase lag (between the input and the output of the amplifier) which causes gain peaking and can destabilise a circuit [85]. Gain peaking is a phenomenon where the gain of the amplifier increases sharply at a high frequency region. We used a feedback capacitor (C_f) parallel to R_f to reduce the gain peaking effect. R_f and C_f in parallel also act as a low-pass filter and can be used to control the bandwidth of the amplifier.

The second stage (Fig. 3.1(b)) provides a voltage gain of 1001 V/V by setting the ratio of R_2 to R_1 to 1000. The third stage (Fig. 3.1(c)) is a buffer circuit connected in a voltage follower configuration. The last stage is implemented to facilitate impedance matching with the analog-to-digital converter (ADC) and to

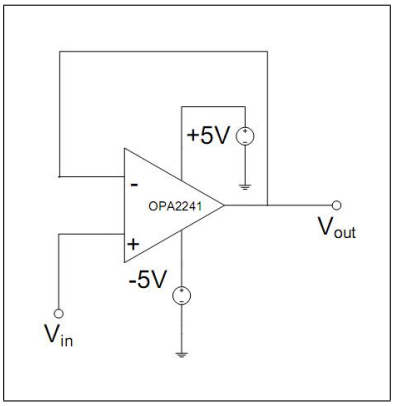
provide the current required for the ADC. We used the OPA2241 to implement the second and third stages because it has low noise, high open loop gain and multiple op amps can be placed in a single integrated circuit to reduce parasitics.



(a) The first stage is a transimpedance amplifier. This amplifier is in the same configuration (photovoltaic) as the existing TTK, except that a smaller 1×10^6 Ω resistor is used.



(b) The second stage of the amplifier is used to provide voltage gain.



(c) Third stage of the multistage amplifier is a unity gain voltage follower. It will provide isolation to the op amp circuit. The last stage helps with impedance matching on the ADC.

Figure 3.1: Multistage amplifier design

Eric Cheong, a M.Sc. graduate, had previously attempted to design a multistage amplifier for Spectrometry applications for the AML. His multistage amplifier was

also a three-stage design but he observed oscillation in the output. We speculate that the cause for this ringing phenomenon was due to the gain peaking effect, where the amplifier oscillates in a high frequency range. In his multistage design, the feedback resistor and capacitor he used were 100 MΩ and 33 pF, respectively [86]. We believe that because the photo-detector he used was a collector plate (presumably with a large built-in capacitance), the feedback capacitor he used was not large enough to compensate for the collector plate's built-in capacitance, hence causing the amplifier to oscillate. We have to attempted to calculate the build-in capacitance of the collector plate used but his thesis only describes the plate diameter to be about 1 to 1.5cm with no information about the separation between the plates [86]. The feedback capacitor should have equal or greater capacitance than the input capacitance of the photo-detector to eliminate gain peaking [40].

3.3 Test Setup

The test setup used to test our multistage amplifier is shown in Fig. 3.2. The micro-controller unit (MCU) board sends amplifier output sampled by the ADC to a PC. The power board provides power and various voltages to other subsystems. The MCU and the power boards were designed by other students and staff at the AML. The modified TTK optics board supplies +5 V and -5 V to the multistage amplifier. The multistage amplifier board consists of a three-stage amplifier and an ADC. The multistage amplifier was designed on a daughter board to provide easy modifications to the amplifier. A Keithley 236 source-measurement unit was used to inject current to test the amplifiers' transimpedance (current to voltage) gain.

3.4 Test Overview

Several tests were performed on the new multistage amplifier and the TTK's single-stage amplifier. In the first set of tests, a photodiode was connected to the amplifier

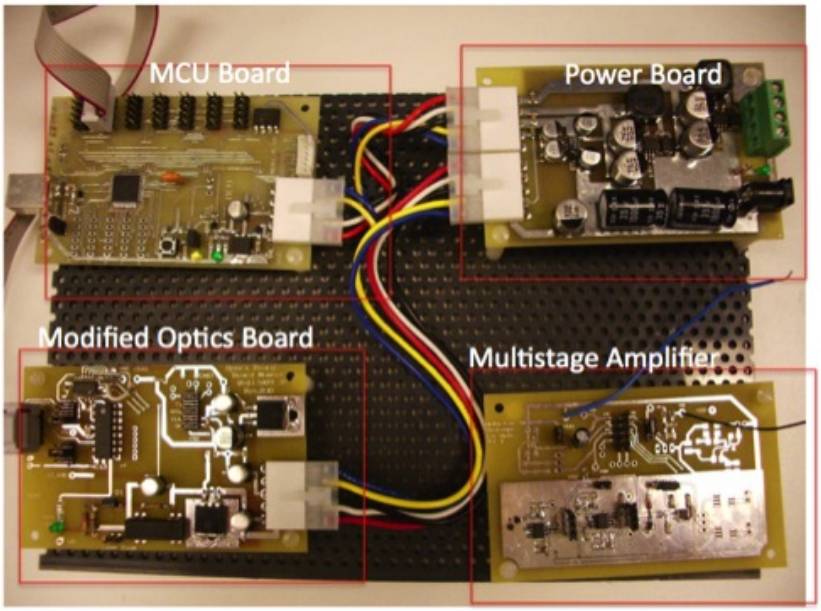


Figure 3.2: Test setup for multistage amplifier.

and measurements were taken when the photodiode was placed in a dark box (no current) and out in the open (photodiode saturation). These tests were used to characterize the noise and the output voltage swing of the amplifiers. We used baseline offset and baseline variation as metrics to evaluate the performance of the amplifiers in these tests. Baseline offset was obtained by calculating the mean ($\bar{x}$) of the signal whereas baseline variation was obtained by calculating the standard deviation (σ). σ is defined by (3.1),

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}}, \tag{3.1}$$

where n is the number of samples recorded.

In the second set of tests, various levels of current were injected into the amplifiers using a Keithley 236 source measurement unit and the DC voltages were measured using a digital multimeter (DMM). This set of tests determines the DC gain of the amplifier. The AC frequency response of the amplifiers was simulated in PSPICE.

DC voltage gain tests were performed separately on the second and third stages of the amplifier. Stepped down voltages (using a power supply and a voltage divider) were used as inputs to the amplifier and the output voltages were measured using a DMM.

3.5 Test Results

3.5.1 TTK Single-Stage Amplifier

We tested the TTK optics board's single-stage amplifier to compare its performance with our multistage amplifier. When the photodiode was placed in the dark, the baseline offset was calculated to be 7.5 mV with a standard variation of 0.4 mV. When the photodiode was saturated with light, the baseline offset was calculated to be 3.59 V with a standard variation of 0.45 mV. The TTK optics board has a voltage swing of 3.59 V between no light and saturated light levels. Various currents were injected into this amplifier and the output voltages were measured as shown in Fig. 3.3. The current to voltage gain for the TTK optics amplifier is about -1×10^9 V/A with less than 10% error from the theoretical design.

The bandwidth of the single-stage amplifier was found by simulation in PSPICE. The circuit simulation model can be found in Appendix C. As shown in Fig. 3.4, we found the bandwidth of the single-stage amplifier to be around 1.6 Hz. The bandwidth of an amplifier is defined by the -3 dB frequency [87]. This simulated result is very close to a published work by Mohammad Behnam [29], a recent Ph.D. graduate from the AML. He found that the bandwidth of this single-stage amplifier to be 1.5 Hz.

We analysed the TTK single-stage amplifier's performance and signal artifacts by looking at a CE run with no DNA product performed by a M.Sc. student Allison Bidulock. The microfluidic chip was filled with water and the injection wells were taped to prevent evaporation. As shown in Fig. 3.5(a), this CE run has a baseline variation of 0.7 mV and a baseline offset of 0.974 V. Fig. 3.5(b) shows the Fourier

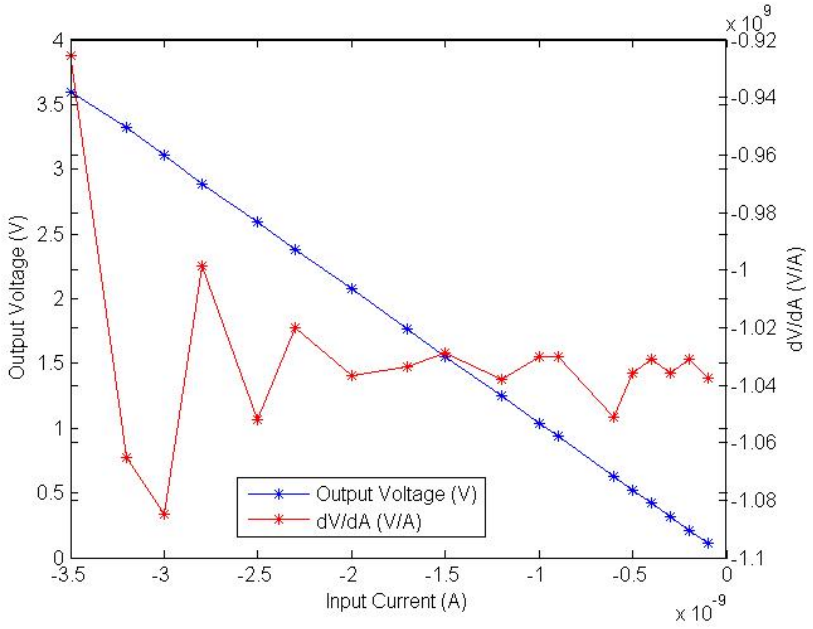


Figure 3.3: TTK optics board amplifier with a current to voltage gain of -1$\times 10^9$ V/A. The scatter in the voltage gain plot is less than 10% error from the theoretical design.

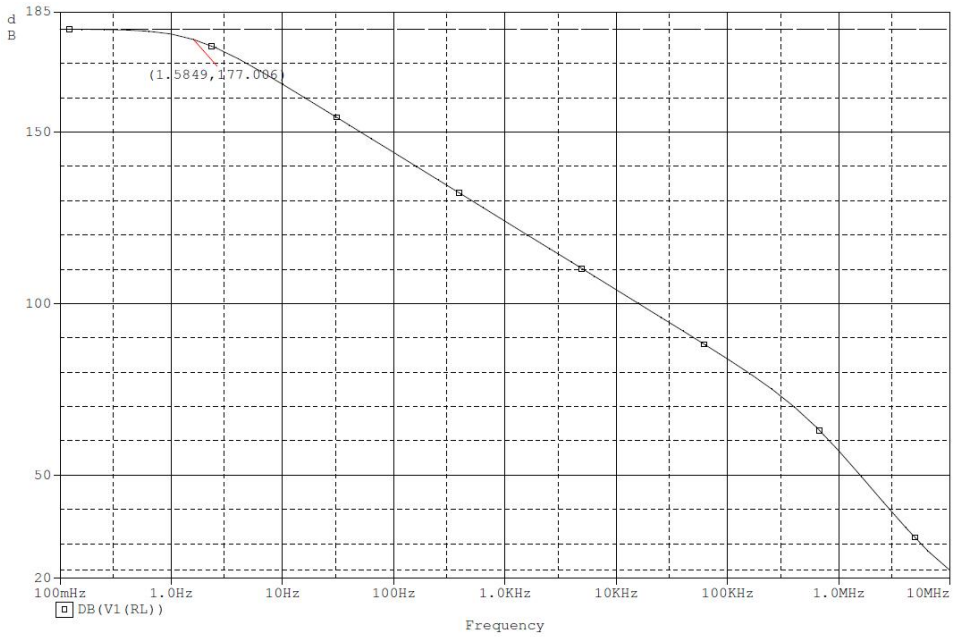


Figure 3.4: Magnitude frequency response of the single-stage amplifier in the existing TTK.

transform (FT) of this signal in the window of 0 $\pm$ 1.5 Hz. FT showed that the majority of the spectral contribution of this signal is low, confirming the amplifier's low bandwidth. On top of the-high frequency noise, we noticed a slow varying baseline. Since there are no DNA product in this run, we speculate this is due to laser intensity fluctuation caused by laser temperature variation. Behnam *et al.* of AML demonstrated that the laser variation can be suppressed by warming up the laser for 10 minutes before CE experiments and attaching a heatsink to the laser [29].

3.5.2 Multistage Amplifier

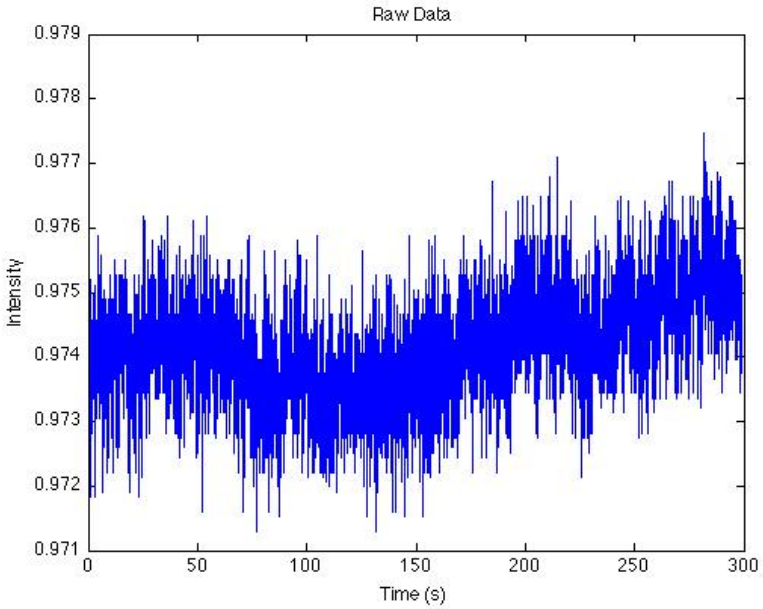
3.5.2.1 First Stage

When the photodiode was placed in the dark, the output DC voltage from the first stage of the multistage amplifier should have remained low. But when the photodiode was exposed to light, the output DC voltage should have been higher than the previous case, but not saturating the amplifier. When there was no current (or light detected by the photodiode), the first stage of the amplifier outputted a signal with a very small baseline offset voltage of 4.04 mV and a small baseline variation of 0.11 mV. When the photodiode was exposed, the first stage of the amplifier outputted signal with a baseline offset voltage of 61.9 mV with a standard variation of 0.23 mV.

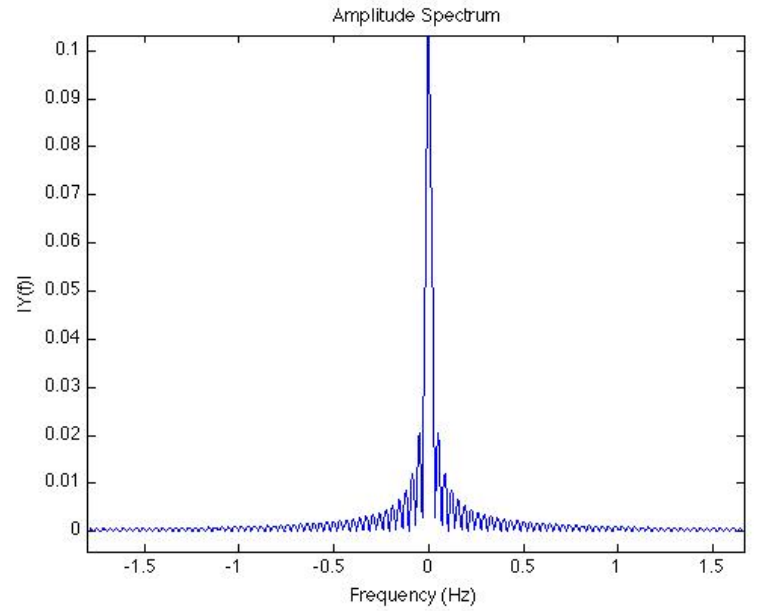
A Keithley 236 was used to source various amounts of current into the first-stage amplifier and a DMM was used to measure the output voltage. As shown in Fig. 3.6, the first stage amplifier behaved as expected, providing a gain of -1$\times 10^6$ V/A with less than 0.3% error from the theoretical design gain.

3.5.2.2 Second Stage

The second-stage amplifier was designed to provide a voltage gain (A_v) of 1001 V/V. A_v is defined by (3.2),



(a) CE run with no DNA product.



(b) Fourier transform

Figure 3.5: Single-stage amplifier's background response.

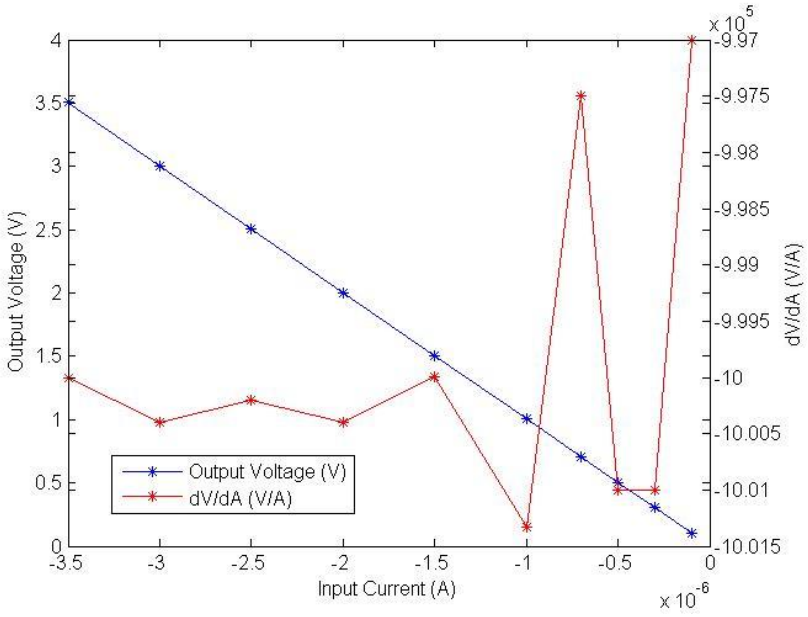


Figure 3.6: First stage (-1×10^6 V/A) voltage output and current to voltage gain. The scatter in the voltage gain plot is only 0.3 % from the theoretical design gain.

$$A_v = \frac{\Delta V_{out}}{\Delta V_{in}}, \tag{3.2}$$

where ΔV_{out} and ΔV_{in} are the change of the output voltage and input voltage, respectively. Using a benchtop power supply and a resistor voltage divider, various DC voltages were applied to the second-stage amplifier and the output voltages were measured. The DC testing results of the second-stage amplifier are shown in Fig. 3.7. We believe the large gain error is due to the imprecision in our recorded data (i.e. one significant digit in our input voltages.)

3.5.2.3 Third Stage

The third stage of the multistage amplifier was designed to follow the input with a voltage gain of 1. Various DC voltages were applied to the third stage of the amplifier and output voltages were measured as shown in Fig. 3.8. The DC test verified that the third stage behaved as designed. Because of op amp clipping, there

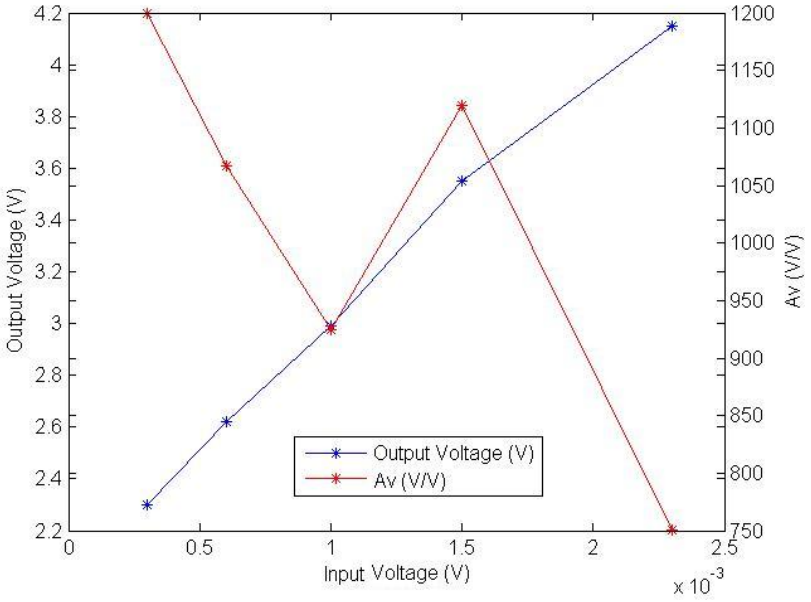


Figure 3.7: DC testing of the second stage amplifier.

was a large voltage gain error of -66.7% when the input voltage was high.

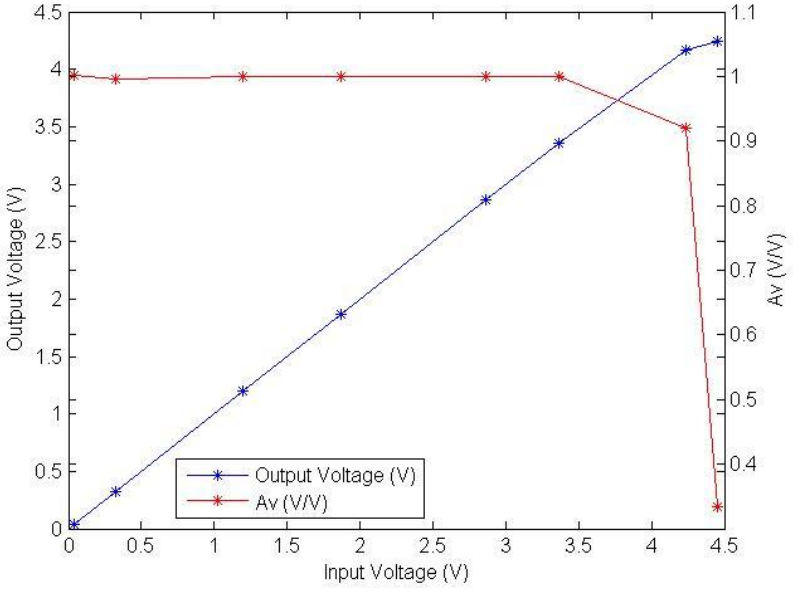


Figure 3.8: DC testing of the third stage amplifier.

3.5.2.4 First and Second Stages Combined

The first two stages of the multistage amplifier were combined and connected to a photodiode. When the photodiode was placed in the dark, the baseline offset (or mean) was 3.52 V and the baseline variation (or standard deviation) was 0.0019 V. With the photodiode exposed to light, the amplifier saturated at 4.08 V. Although the first two stages of the amplifier were responsive to different lighting levels, the high offset of 3.52 V in the dark gave a voltage swing of less than 0.5 V.

3.5.2.5 Second-Stage Gain Reduction

A second stage with a smaller voltage gain of 101 V/V was built on a breadboard to lower the baseline offset. Various DC voltages were applied to the second stage amplifier and the output voltages were measured, as shown in Fig. 3.9. The voltage gain for this amplifier was almost constant over the DC range.

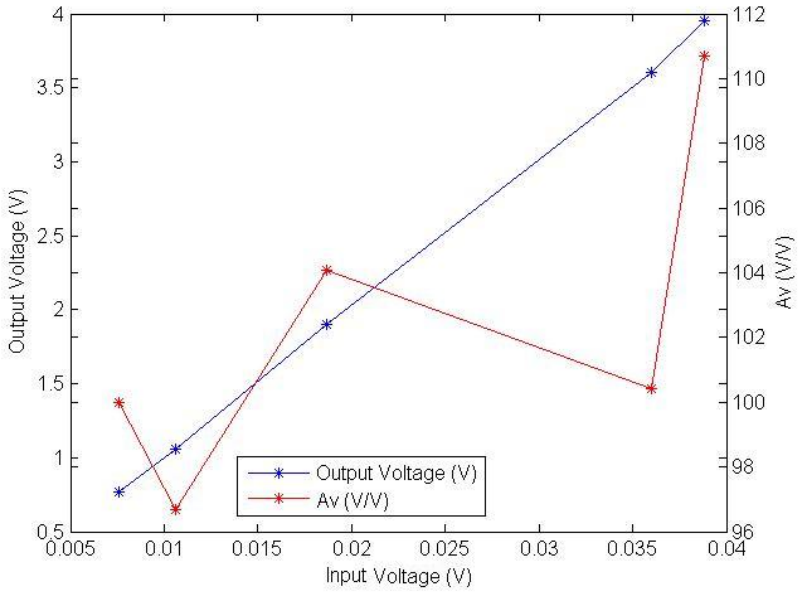


Figure 3.9: DC testing of a modified second-stage amplifier with a gain of 101 V/V (implemented on a breadboard).

When the photodiode was connected and placed in the dark, the baseline offset

remained low at 0.304 V with a baseline variation of 3.3 mV. When saturated with light, the baseline offset was calculated to be 4.08 V and the baseline variation was found to be 94 μV. Lowering the gain in the second-stage amplifier improved the amplifier's voltage swing to 3.7 V. Fig. 3.10 shows the DC response of the amplifier when the source measurement unit was used to source current into this amplifier. For the majority of the data points, the amplifier had a gain around -1$\times 10^8$ V/A, as designed.

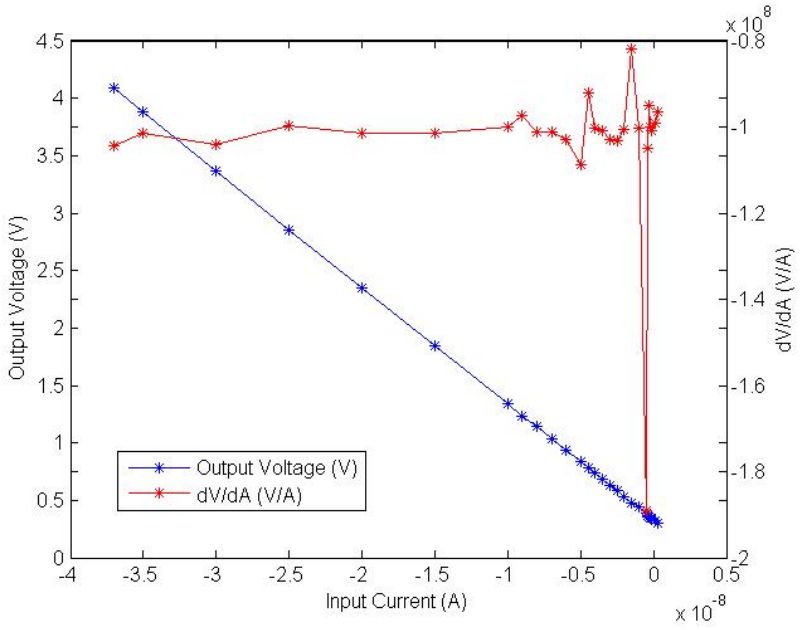


Figure 3.10: Two-stage amplifier (-1$\times 10^8$ V/A) current to voltage gain.

3.5.2.6 First-Stage Gain Increase

Because we lowered the gain of the second stage, the overall gain for the multistage amplifier was only -1$\times 10^8$ V/A, which is about ten times less than the gain of the current TTK amplifier. To compensate for this problem, we increased the gain of the first stage by 10. Increasing the first stage gain was accomplished by using a 1×10^7 Ω feedback resistor in the first stage instead of a 1×10^6 Ω resistor.

When there was no light, the baseline offset and variation were 3.4 mV and 1.9

mV, respectively. When the photodiode was exposed to light, the baseline offset and variations were 0.179 V and 4.1 mV. Various currents were injected to the amplifier and the output voltages were measured, as shown in Fig. 3.11. The gain measured was close to the theoretical gain of -1×10^{7} V/A, with the exception of very low current levels between 0 and 1×10^{-7} A. We believe this is due to the leakage of current into the op amp's terminals.

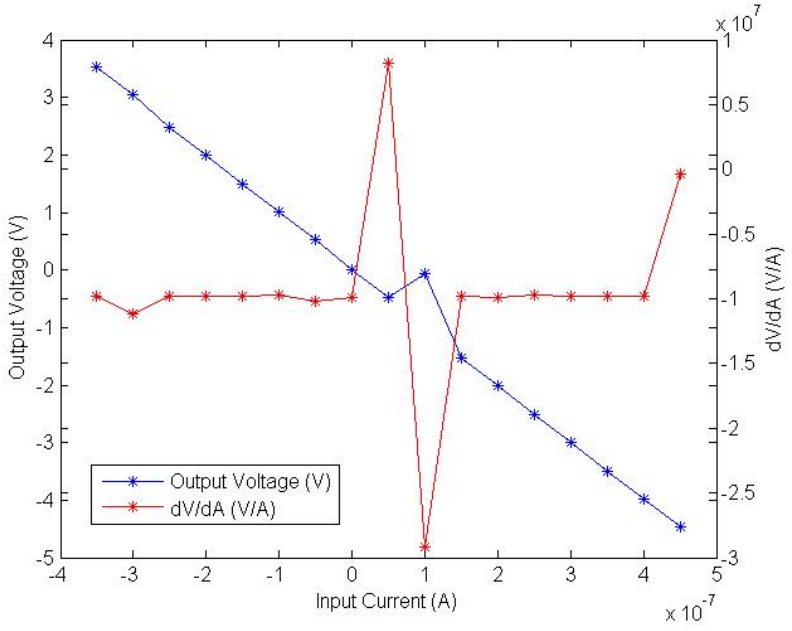


Figure 3.11: Single-stage amplifier (-1×10^{7} V/A) current to voltage gain.

When a photodiode was connected and placed in the dark, the baseline offset and variation were calculated to be 0.30 V and 2.6 mV, respectively. When the photodiode was exposed to light, the baseline offset and variation were calculated to be 4.0 V and 0.158 mV, respectively. The DC current to voltage gain of this amplifier is shown in Fig. 3.12. The gain of multistage amplifier was found to be around -1×10^{9} V/A, which is very close to the TTK optics board's single-stage amplifier.

The bandwidth of this multistage amplifier was simulated. Without sacrificing any gain, the multistage amplifier improved the bandwidth to 160 Hz, as shown in

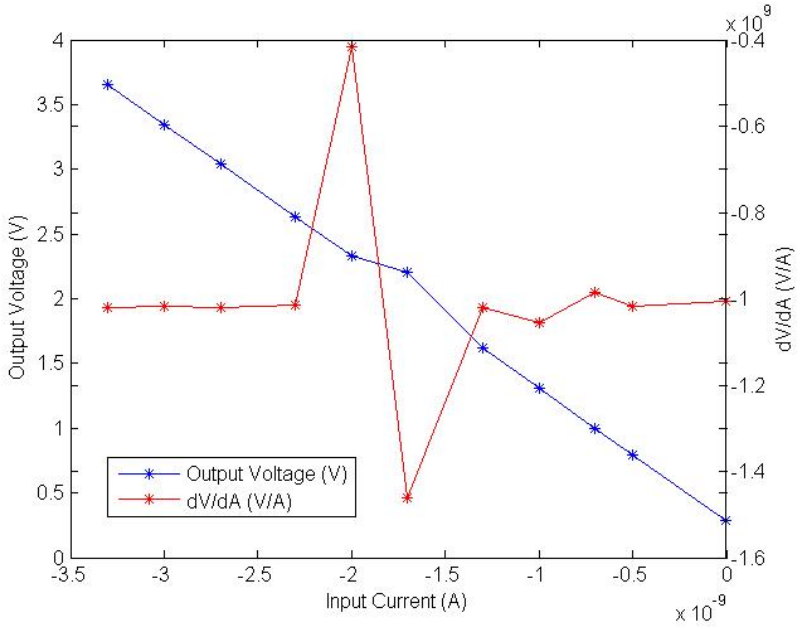


Figure 3.12: Two-stage amplifier (-1$\times 10^9$ V/A) current to voltage gain.

Fig. 3.13 via PSPICE simulation.

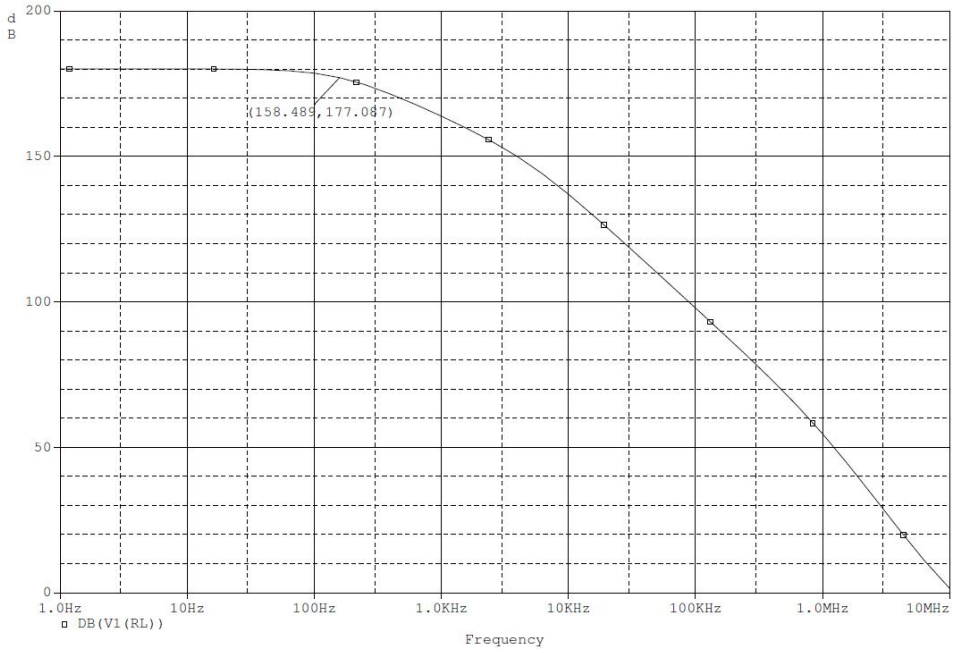


Figure 3.13: Magnitude response of the multistage amplifier.

3.6 Performance Discussion

The performances of the various multistage amplifier prototypes and the TTK's single-stage amplifier are summarized in Table 3.1. As discussed, the multistage amplifier possesses the same current to voltage gain (-1×10^9 V/A) as the single-stage amplifier with a bandwidth increase from 1.6 Hz to 160 Hz.

One disadvantage of the multistage amplifier is that it has higher baseline offset (300 mV) compared to the single-stage amplifier (7.5 mV). This is due to the amplification of the baseline DC offset voltage from the first stage. We lowered the baseline offset in the first iteration (3.52 V) of the multistage amplifier by increasing the gain of the first stage and lowering the gain of subsequent stage. Although the multistage amplifier has higher baseline offset compared to the single-stage amplifier, it is compensated by the increase of voltage headroom to 4 V; therefore, improving the voltage swing. The voltage swing could be further increased by a circuit modification discussed in the recommendation section of this chapter.

The multistage amplifier has higher baseline variation (2.6 mV) than the single-stage amplifier (0.4 mV). We speculate that this is due to thermal noise. As described in section 2.3.2, the thermal noise is a product of temperature, resistance, and the bandwidth of the measured signal. As the bandwidth of the amplifier increased, thermal noise should increase as well. Another contribution to the noise in the multistage amplifier could be because the final prototype tested was partially implemented on a breadboard. The parasitics between traces and the long wires used to connect components could have increased interference noise.

3.7 Noise Analysis

3.7.1 Op Amp Noise

A real op amp can be modelled by the op amp model shown in Fig. 3.14(a) [42]. It is modelled by the combination of input referred noise sources which consist of a

Specifications	TTK	Multistage 1	Multistage 2	Multistage 3
First-Stage Gain (V/A)	-1×10^{9}	-1×10^{6}	-1×10^{6}	-1×10^{7}
Second-Stage Gain (V/V)	-	1001	101	101
Current to Voltage Gain (V/A)	-1.00×10^{9}	-1.00×10^{9}	-1.01×10^{8}	-1.01×10^{9}
Bandwidth (Hz)	1.6	1.47×10^{3}	1.47×10^{3}	1.61×10^{2}
Baseline Variation (V)	4×10^{-4}	1.9×10^{-3}	1×10^{-4}	2.6×10^{-3}
Baseline Offset in the Dark (V)	7.50×10^{-3}	3.52	0.304	0.3
Baseline Offset Exposed (V)	3.59	4.08	4.08	4.0
Voltage Swing (V)	3.58	0.56	3.776	3.7

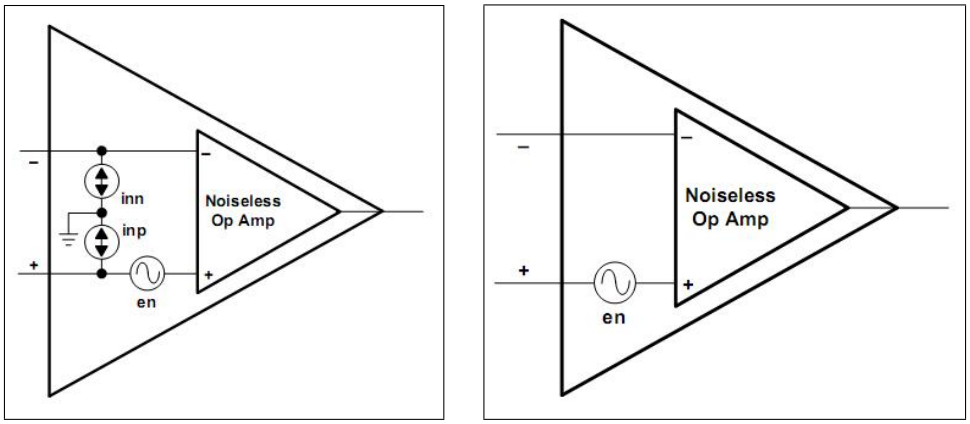
Table 3.1: Comparison between TTK single-stage amplifier and multistage amplifier.

voltage noise source and two current noise sources at its terminals and a noiseless op amp. Input referred noise sources are used to represent the total noise in the op amp by placing noise sources at its inputs. But for a low input impedance junction gate field-effect transistor (JFET) op amp, only the voltage noise is important. This is because the input noise sources are dominated by the input impedance of the op amp [88]. Thus, op amps are often characterised by only the input referred voltage noise, as shown in Fig. 3.14(b), for hand analysis.

The op amp input referred voltage noise $\mathbf{e}_n$ in V_{RMS} can be described by equation (3.3) [42],

$$\mathbf{e}_n^2 = \mathbf{e}_w^2[f_{nc} ln\left(\frac{f_h}{f_l}\right) + ENB] \tag{3.3}$$

where $\mathbf{e}_w^2$ is the op amp's white noise specification, which can be found in the op amp datasheet. f_{nc} is the noise corner frequency, where the white noise and $1/f$ noise are equal. It can be found by finding the frequency where the total noise is



(a) Op amp model with current and voltage noise sources

(b) Equivalent op amp noise model

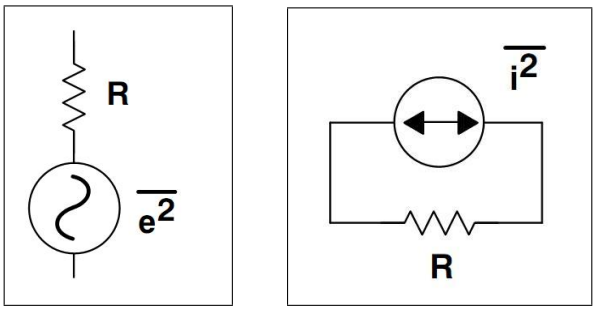
Figure 3.14: Op amp noise model

$\sqrt{2}$ times of the white noise. For OPA 129, the white noise specification is approximately equal to the noise density at 10 kHz, which is 15 nV/$\sqrt{(Hz)}$. This is because from the input voltage noise spectral density graph in the datasheet, the noise voltage density does not increase after 10 kHz. To find the corner frequency, we used the same input voltage noise spectral density graph and found a frequency of 200 Hz where the noise voltage density is at about 21 nV/$\sqrt{(Hz)}$ ($\sqrt{2}$ times of the white noise).

Equivalent noise bandwidth (ENB) is used to account for the extra noise outside of the bandwidth of the amplifier, which is 1.57 times of the 3-dB frequency for first order systems with a single pole such as our amplifier [42] . f_h and f_l are the upper and lower bound for the bandwidth of interest. Equation (3.3) accounts for thermal and shot noise, two dominating noise sources in op amps.

3.7.2 Resistor Noise

Resistors can be modelled by a voltage noise source in series with the resistor or a current noise source in parallel to the resistor, as shown in Fig. 3.15 [42]. The noise voltage and current sources are defined by equations (3.4) and (3.5), respectively.



(a) Voltage noise source model

(b) Current noise source model

Figure 3.15: Resistor noise model.

$$\mathbf{e}^2 = \int 4KTRdf \tag{3.4}$$

$$\mathbf{i}^2 = \int \frac{4KT}{R} df \tag{3.5}$$

where K, T, and R represents Boltzmann's constant, the temperature in Kelvin, and the resistance in Ω, respectively. f is the frequency range of the bandwidth which noise is measured or the equivalent noise bandwidth. We chose to use a series voltage source in our analysis.

3.7.3 Photodiode Noise

Shot noise is dominant in a photodiode if it is configured in high speed reverse-biased photoconductive mode. But because we implemented our photodiode in low noise photovoltaic mode with zero biasing, dark current approaches zero and shot noise is essentially eliminated. However, thermal noise still exists. To calculate the thermal noise contribution from our photodiode, we modelled our photodiode by the combination of a junction capacitance (C_j), a shunt resistor (R_{sh}) and a series resistor (R_s), as shown in the equivalent photodiode noise model (Fig. 3.16) [40]. $\mathbf{e}_1$ and $\mathbf{e}_2$ are the thermal noise associated with R_{sh} and R_s, respectively.

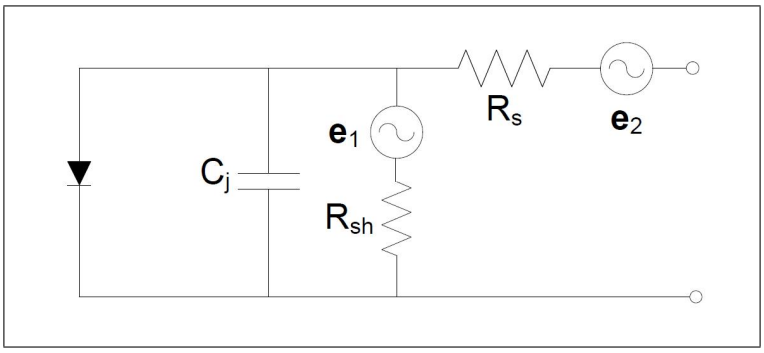


Figure 3.16: Photodiode noise model

Because the resistance (and thermal noise) of R_{sh} is always much greater than that of R_s [40], we will only investigate the effect of R_{sh}. Using a Keithley 236, we applied a DC voltage (1V) to the photodiode and measured about 1 nA of current going through the photodiode; therefore, the R_{sh} is around 1×10^9 Ω. It is ideal to use low bias voltages; however, we used a bias voltage of 1V because at lower bias voltages, the Keithley 236 could not measure the current going through the photodiode.

We calculated the photodiode current noise ($\mathbf{e}_{pd}$) using the photodiode's Noise Equivalent Power (NEP) and responsivity (R). The photodiode current noise can be calculated by equation (3.6),

$$\mathbf{e}_{pd}^2 = (NEP \times R)^2 \times ENB. \tag{3.6}$$

The NEP and R for the NT57-506 in photoconductive mode is 2.8×10^{-15} W/$\sqrt{Hz}$ and 0.65 A/W, respectively [84]. The NT57-506 photodiode datasheet does not state the NEP and R values for the amplifier configured in photovoltaic mode. However, the analysis method and the characterizing equations do not change and only the NEP and R values change. For completeness, we will calculate and compare the photodiode's noise contribution from the NT57-506 photodiode configured in photoconductive mode to another suitable photodiode configured in photovoltaic mode.

3.7.4 Total Noise

To obtain the total noise, noise sources are added to the circuit and the input signals are disconnected. The first stage of the multistage amplifier becomes the circuit shown in Fig. 3.17. When there are multiple noise sources in the circuit and if they are independent, the total noise in the circuit is the sum of the contribution from each noise source.

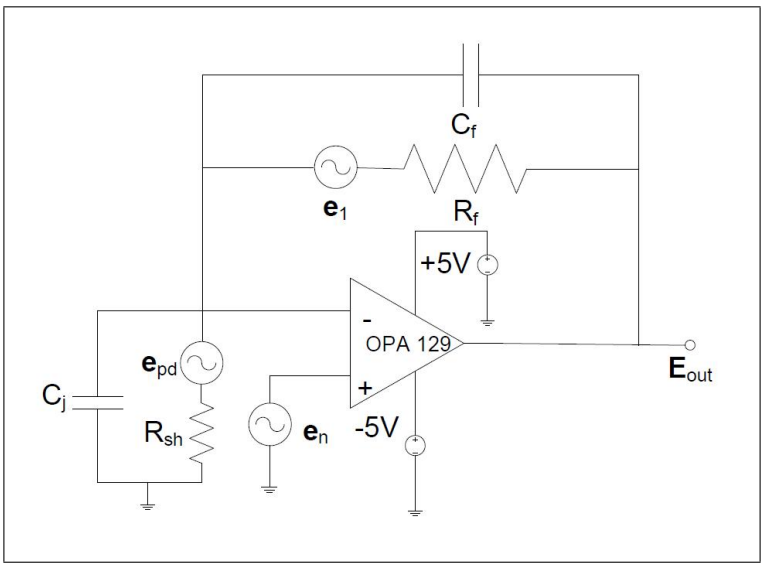
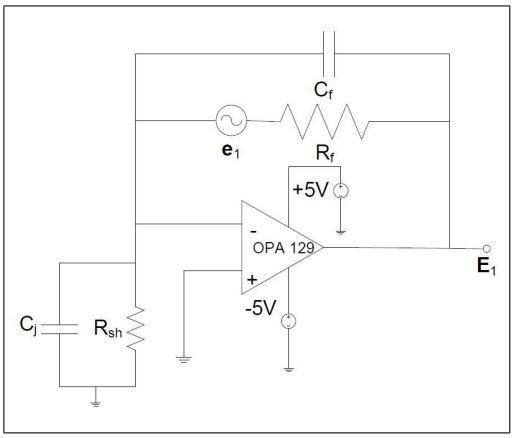
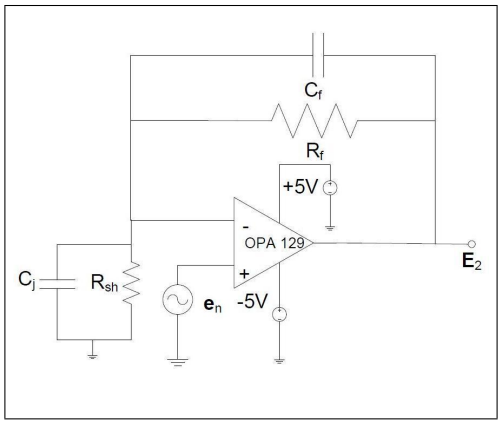


Figure 3.17: First-stage amplifier noise model

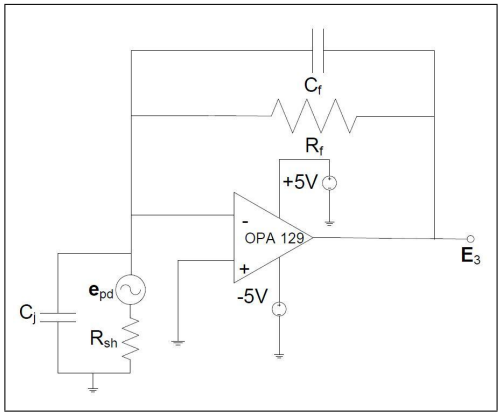
We analysed the noise in our amplifier circuit by separating the noise sources and used superposition to obtain the total noise [42]. As shown in Fig. 3.18(a), the op amp was assumed to be noiseless, but configured in negative feedback with a noisy resistor R_f. The noise contributed by the noisy resistor is represented by a voltage noise source $\mathbf{e}_1$. The noise contribution from the noisy R_f is represented by $\mathbf{E}_1$. Because the current going into the op amp's terminals can be assumed to be zero and the negative terminal of the op amp is a virtual ground, $\mathbf{E}_1$ is equal to $\mathbf{e}_1$, as shown in equation (3.7). The (thermal) voltage noise $\mathbf{e}_1$ in V_{RMS} from R_f is shown by equation (3.8).



(a) Noise due to the resistor



(b) Noise due to the op amp



(c) Noise due to the photodiode

Figure 3.18: Noise equivalent circuits

$$\mathbf{E}_1 = \mathbf{e}_1 \tag{3.7}$$

$$\mathbf{e}_1^2 = \int 4KTR_f\,df \tag{3.8}$$

The noise contribution from the op amp's voltage noise ($\mathbf{e}_n^2$) can be analysed using Fig. 3.18(b). Assuming the voltage between the positive and negative terminals of the op amp is zero, then the voltage drop across R_f is the potential difference between $\mathbf{E}_2$ and $\mathbf{e}_n$. The current flowing into the op amp's terminals is assumed to

be zero, therefore the same current goes through R_f. Using these two conditions, the noise contribution from the op amp ($\mathbf{E}_2^2$) can be shown by equation (3.9).

$$\mathbf{E}_2^2 = \mathbf{e}_n^2 \left(1 + \frac{R_f}{R_{sh}}\right)^2 \tag{3.9}$$

Similarly, the noise contribution from the photodiode ($\mathbf{E}_3^2$) can be analysed using Fig. 3.18(c) and represented by equation (3.10),

$$\mathbf{E}_3^2 = \mathbf{e}_{pd}^2 R_f^2, \tag{3.10}$$

where $\mathbf{e}_{pd}^2$ is the photodiode current noise.

The overall noise ($\mathbf{E}_{out1}^2$) in the first stage amplifier represented by equation (3.11), which is the summation of the noise contribution from R_f, the op amp and the photodiode.

$$\mathbf{E}_{out1}^2 = R_f^2[\frac{4KT}{R_f} + NEP \times R] \times ENB + e_w^2 \left(1 + \frac{R_f}{R_{sh}}\right)^2 [f_{nc} ln\left(\frac{f_h}{f_l}\right) + ENB] \tag{3.11}$$

Noise analysis for the second stage of the multistage amplifier can be performed in a similar manner. Fig. 3.19 shows the circuit noise sources for the second stage amplifier. $\mathbf{e}_1$ and $\mathbf{e}_2$ are the voltage noise sources associated with R_1 and R_2, respectively. $\mathbf{e}_3$ is the noise from the first stage of the amplifier and $\mathbf{e}_n$ is the input referred voltage noise source used to model the op amp. Using the same analysis method as described in the first stage, it can be shown that equations (3.12 - 3.15) can be used to represent the noise contribution from these noise sources. Using superposition and substituting for the variables, equation (3.16) can be used to calculate the total noise voltage ($\mathbf{E}_{out2}^2$) from both stages of the op amp in V_{RMS}.

$$\mathbf{E}_1 = \mathbf{e}_1 \left(\frac{R_2}{R_1}\right) \tag{3.12}$$

$$\mathbf{E}_2 = \mathbf{e}_2 \tag{3.13}$$

$$\mathbf{E}_3 = \mathbf{e}_3 \left(1 + \frac{R_2}{R_1}\right) \tag{3.14}$$

$$\mathbf{E}_n = \mathbf{e}_n \left(1 + \frac{R_2}{R_1}\right) \tag{3.15}$$

$$\mathbf{E}_{out2}^2 = 4KT \times ENB \times R_2 \left(1 + \frac{R_2}{R_1}\right) + \left(1 + \frac{R_2}{R_1}\right)^2 [\mathbf{E}_{out1}^2 + e_w^2 [f_{nc} ln \left(\frac{f_h}{f_l}\right) + ENB]] \tag{3.16}$$

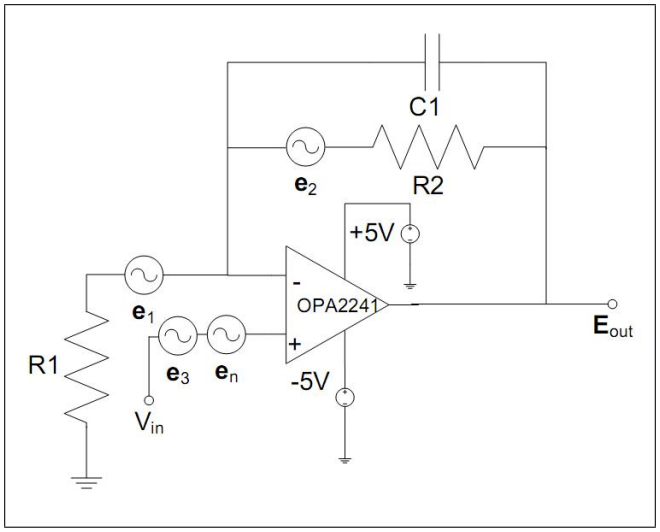


Figure 3.19: Second stage amplifier noise model

As discussed, the total noise in the TTK amplifier consists of feedback resistor noise, op amp noise, and the photodiode noise. Using equation (3.11), we calculated the noise in the TTK's current single-stage amplifier circuits. A summary of the noise analysis parameters and the total noise from 0.1 Hz to 1.6 Hz is shown in Table 3.2. R_f is the feedback resistor used in the amplifier. R_{sh} is the measured shunt resistance. f_{3dB} is the simulated bandwidth of the op amp. ENB is the equivalent

Parameter	Value
R_f (Ω)	1×10^{9}
R_{sh} (Ω)	1×10^{9}
f_{3dB} (Hz)	1.6
ENB (Hz)	2.51
$\mathbf{e}_w$ ($V/\sqrt{Hz}$)	15×10^{-9}
f_{nc} (Hz)	200
f_h (Hz)	1.6
f_l (Hz)	0.1
Photodiode NEP ($W/\sqrt{Hz}$)	2.8×10^{-15}
Photodiode Reponsivity (A/W)	0.65
$\mathbf{e}_{Rf}$ (V_{RMS})	6.45×10^{-6}
$\mathbf{e}_{opamp}$ (V_{RMS})	7.08×10^{-7}
$\mathbf{e}_{photodiode}$ (V_{RMS})	2.88×10^{-6}
$\mathbf{E}_{total}$ (V_{RMS})	7.10×10^{-6}

Table 3.2: Noise analysis parameter for the single stage amplifier in the TTK from 0.1Hz to 1.5 Hz.

noise bandwidth for a f_{3dB} frequency of 1.6 Hz. $\mathbf{e}_w$ and f_{nc} were obtained from the voltage noise density graph in the OPA 129's datasheet. The values for f_h and f_l were chosen because of the bandwidth of the amplifier. The photodiode's NEP and responsivity values were obtained from the specifications from the photodiode's datasheet. The total noise in the TTK single stage amplifier is 7.10×10^{-6} V_{RMS} and is dominated by the noise contributions from the feedback resistor, which is 6.45×10^{-6} V_{RMS}.

Consequently, and the total noise ($\mathbf{E}_{out2}$) in our multistage amplifier can be calculated using equations (3.11) and (3.16). A summary of the noise analysis parameters and the total noise from 0.1 Hz to 160 Hz is shown in Table 3.3. Similar to the notations used for the single stage amplifier, R_f is the feedback resistor used in the first stage of the amplifier, R_{sh} is the photodiode's measured shunt resistance, f_{3dB} is the simulated bandwidth of the op amp, ENB is the equivalent noise bandwidth for the f_{3dB} frequency and $\mathbf{e}_w$ and f_{nc} were obtained from the voltage noise

density graph in the datasheets of the OPA 129 and OPA 2241. The values for f_h and f_l were chosen because of the bandwidth of the amplifier. The photodiode's NEP and responsivity values were obtained from the specifications from the photodiode's datasheet. R_1 and R_2 were the resistors used in the second gain stage of the op amp. Our analysis showed that the multistage amplifier has a total noise of 2.79×10^{-4} V_{RMS}, which is higher than the single-stage amplifier used in the TTK (7.10×10^{-6} V_{RMS}). This is due to the increased of bandwidth from 1.6 Hz to 160 Hz and the amplification of the first-stage amplifier noise in the second stage. The main source of noise in the first stage of the amplifier is due to the thermal noise of R_f. Whereas in the second stage, the main source of noise is from the amplification of $\mathbf{E}^2_{out1}$, the output noise from the first stage. Our noise analysis provides valuable noise metrics to compare noise associated with different design and showed that multistage design are noisier than single stages.

3.7.5 Photovoltaic and Photoconductive Photodiode Comparison

As discussed, the NEP and R values for the NT57-506 photodiode configured in photovoltaic mode are not stated in its datasheet. Since the analysis method remains the same, and the only difference is the change in NEP and R values in calculating the noise contribution for photodiodes configured in photoconductive and photovoltaic modes, we will compare the noise contribution from the NT57-506 in photoconductive mode to another suitable photodiode in photovoltaic mode. A suitable photodiode that can be configured in photovoltaic mode is the PIN-2DPI (OSI Optoelectronics). This photodiode has a NEP of 2.1×10^{-15} $W/\sqrt{Hz}$ and a responsivity (R) of 0.55 A/W [89]. Using equation (3.6) and the parameters in Table 3.2, we calculated the noise contribution from the PIN-2DPI photodiode configured in photovoltaic mode to be 1.83×10^{-6} V_{RMS}. This is a 35% noise reduction compared to that of the NT57-506, which is 2.88×10^{-6} V_{RMS}. Using equation (3.11), we cal-

Parameter	Multistage amplifier
R_f (Ω)	1×10^{7}
R_{sh} (Ω)	1×10^{9}
Photodiode NEP ($W/\sqrt{Hz}$)	2.8×10^{-15}
Photodiode Reponsivity (A/W)	0.65
$\mathbf{e}_{w1}$ ($V/\sqrt{Hz}$)	15×10^{-9}
$\mathbf{e}_{w2}$ ($V/\sqrt{Hz}$)	45×10^{-9}
f_{3dB} (Hz)	160
ENB (Hz)	250
f_{nc1} (Hz)	200
f_{nc2} (Hz)	1.5
f_h (Hz)	160
f_l (Hz)	0.1
R_1 (Ω)	100
R_2 (Ω)	1×10^{4}
$\mathbf{E}_{R1}$ (V_{RMS})	2.05×10^{-7}
$\mathbf{E}_{R2}$ (V_{RMS})	2.05×10^{-8}
$\mathbf{E}_{opamp2}$ (V_{RMS})	1.75×10^{-4}
$\mathbf{E}_{FirstStageAmp}$ (V_{RMS})	2.18×10^{-4}
$\mathbf{E}_{out2}$ (V_{RMS})	2.79×10^{-4}

Table 3.3: Noise analysis parameter for the multistage stage amplifier from 0.1Hz to 160 Hz. Subscripts 1 and 2 of f_{nc} and $\mathbf{e}_w$ indicate the first stage and the second stage op amp specifications, respectively.

culated the total noise in the single stage photovoltaic amplifier to be 6.74×10^{-6} V_{RMS}, a slight reduction of noise compared to 7.1×10^{-6} V_{RMS}, the total noise in the photoconductive amplifier with the NT57-506. The reason that the improvement in noise is small in the amplifier is because the thermal noise from the feedback resistor dominates the noise contribution from the photodiode.

3.8 Recommendations

In the next amplifier design for the TTK optics board, I recommend using a two-stage amplifier. The first stage (Fig. 3.1(a)) provides a current-to-voltage gain and the second stage (Fig. 3.1(b)) provides a voltage gain. The third buffer stage is not

necessary because the op amp from the second stage is capable of driving the ADC. Hence adding a third stage only adds more noise from the the input referred noise of the third op amp. The gain of the first stage of the amplifier should be maximized to minimize noise as suggested by Leach *et al.* [90]. The gain of the first stage should to be set to -1×10^8 V/A by using a feedback resistor (R_f) of 1×10^8 Ω. The bandwidth of this amplifier should be lowered by changing the feedback capacitor (C_f) to 100×10^{-12} F. This amplifier will be connected to a voltage gain stage of 11 V/V. The AC simulation of this proposed amplifier (Fig. 3.20) shows a bandwidth of about 16 Hz, which is about ten times of the bandwidth of the current TTK single-stage amplifier. The passage of a DNA peak (of a given size) usually takes about a second and typical sampling rates for DNA detection are about 10 Hz to obtain all available peak information. Hence, a bandwidth of 16 Hz is adequate for detection and lowering the bandwidth from the current multistage amplifier design reduces thermal noise and interference noise.

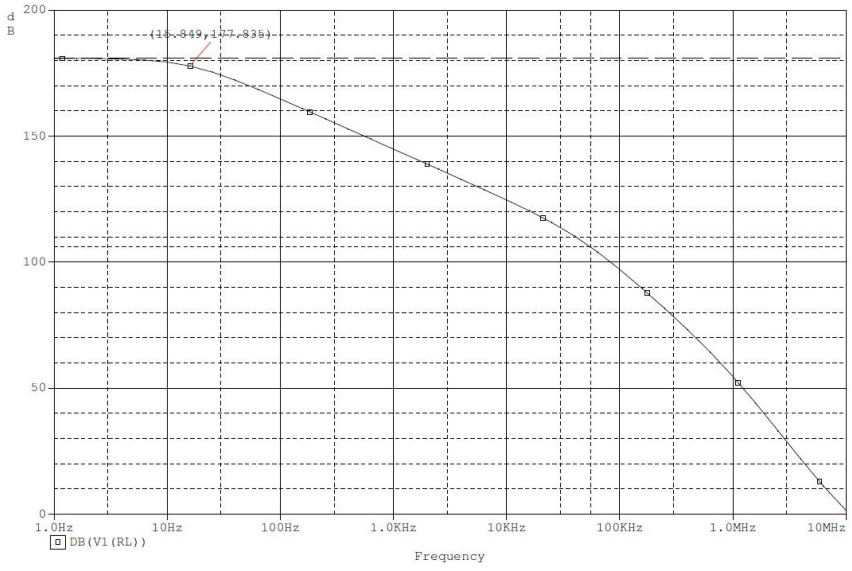


Figure 3.20: AC frequency simulation of the recommended amplifier. This amplifier has a bandwidth of 16 Hz and a current-to-voltage gain of -1.1×10^9 V/A (or 180.8 dB)

The voltage swing of the multistage amplifier can further be increased by adding an offset adjustment circuitry and connecting to the second stage of the amplifier as shown in Fig. 3.21. The reference voltage V_r can be adjusted by tuning the potentiometer. The overall amplifier's output voltage can be represented by Eq. (3.17),

$$V_{out} = I_f R_f \frac{R_2 + 1}{R_1} - \frac{R_2 V_r}{R_1}, \tag{3.17}$$

where I_f and R_f are the photodiode current and the feedback resistor from the first stage, respectively. By tuning V_r, the DC offset in V_{out} can be lowered effectively, thus increasing the output voltage swing of the amplifier.

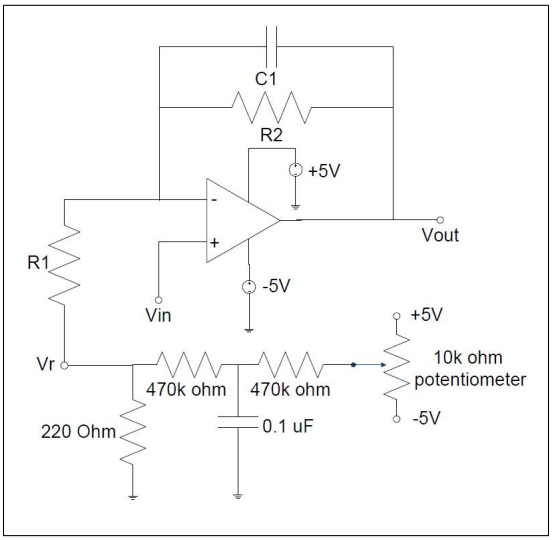


Figure 3.21: Recommended 2nd stage amplifier design to increase voltage swing

We also recommend exploring other op amps in the design of the multistage amplifier. One op amp to consider is the LMV792 op amp manufactured by National Semiconductor. We compared the specifications of the LMV792 and the OPA129 op amp currently used in Table 3.4. The open loop gains for the two op amps are very similar. However, the LMV792 op amp has a higher GBP (17 MHz) and lower input referred voltage noise at various frequencies than the OPA129. Higher GBP

Specifications	OPA129	LMV792
Voltage noise at 10 Hz (nV/$\sqrt{Hz}$)	85	20
Voltage noise at 100 Hz (nV/$\sqrt{Hz}$)	28	10
Voltage noise at 1 kHz (nV/$\sqrt{Hz}$)	17	5.8
Gain bandwidth product (MHz)	1	17
Typical input bias current (fA)	30	50
Over loop gain (dB)	94	92
Cost ($)	8.98	2.50

Table 3.4: Specifications comparison between the OPA129 and LMV792 op amps

allows for wider bandwidth applications and low input referred voltage noise enables a high SNR [85]. Since the current from the photodiode is in the range of nA, a low noise op amp is essential to amplifier design.

At the time when we were building and testing the multistage amplifiers, we were also working on the post-experiment signal processing aspect of this project. We were making more progress on the signal processing side and we focused our efforts on the signal processing instead. However, with the detailed design, analysis, and testing procedures, and recommendations we have provided, it would be easy for another person to build and implement the recommended multistage amplifier.

3.9 Conclusion

A multistage amplifier for the optical detection subsystem of the TTK was designed, built, and tested. The multistage amplifier provided a higher bandwidth (160 Hz) and voltage swing (4 V) compared to the single-stage amplifier currently used, while maintaining the same current to voltage gain of -1×10^9 V/A. However, detailed noise analysis and measurements of the single-stage and multistage amplifiers showed that the single-stage has lower noise compared to the multistage amplifier due to the increase of bandwidth. Recommendations were made for future improvements to the multistage amplifier design.

Chapter 4

Wavelet Denoising

4.1 Introduction

Now that we have looked at improving the TTK's amplifier circuits in Chapter 3 of this thesis, we move to software signal processing of capillary electrophoresis (CE) signals. In other words, we will investigate what can be done to improve CE signals post experiment and data collection. We describe the removal of high-frequency noise in CE signals in this chapter.

The use of a non-confocal optics has allowed for drastic cost and size reductions in the TTK compared to commercial devices that use confocal optics. This is a step toward the realization of a point-of-care (POC) medical diagnostic device. However, compared to the PMT-based detection method, the photodiode-based detection method is not as sensitive and is more prone to noise interference. As a result, a lower signal-to-noise ratio (SNR) and a higher limit of detection (LOD) are issues commonly observed in the captured signals. We reviewed various types of high-frequency noise observed in CE signals and methods to remove them in Chapter 2. We made a contribution in Chapter 3 to the electronics side and now we are focusing on post processing of recorded CE signals to extract relevant information. We will be performing our signal processing with MATLAB software. Our intent is to record the signal out of our electronics and perform post processing with MATLAB.

4.2 Theory

4.2.1 Discrete Wavelet Transform

Since CE signals are time dependent non-stationary signals, the wavelet transform (WT) approach is commonly used in removing high-frequency noise in CE signals because wavelets are localised in both time and frequency domain [46]. However, no one has researched on the use of wavelets for noise removal in signals collected by non-confocal CE instruments. For this reason, we will be incorporating WT into the TTK signal processing.

WT analysis is a hybrid between time and frequency domain analysis [91]. WT is similar to FT; but instead of representing the data in series of sinusoidal signals as in the FT, WT represents data with a superposition of scaled and translated wavelets. Fig. 4.1 shows examples of some commonly used wavelets for denoising [46].

WT decomposes a signal onto a set of wavelet orthogonal basis functions. In general, any function $f(t)$ can be represented by a superposition of wavelets defined by equation (4.1) [46]. $C_f(j,k)$ are wavelet coefficients defined by equation (4.2):

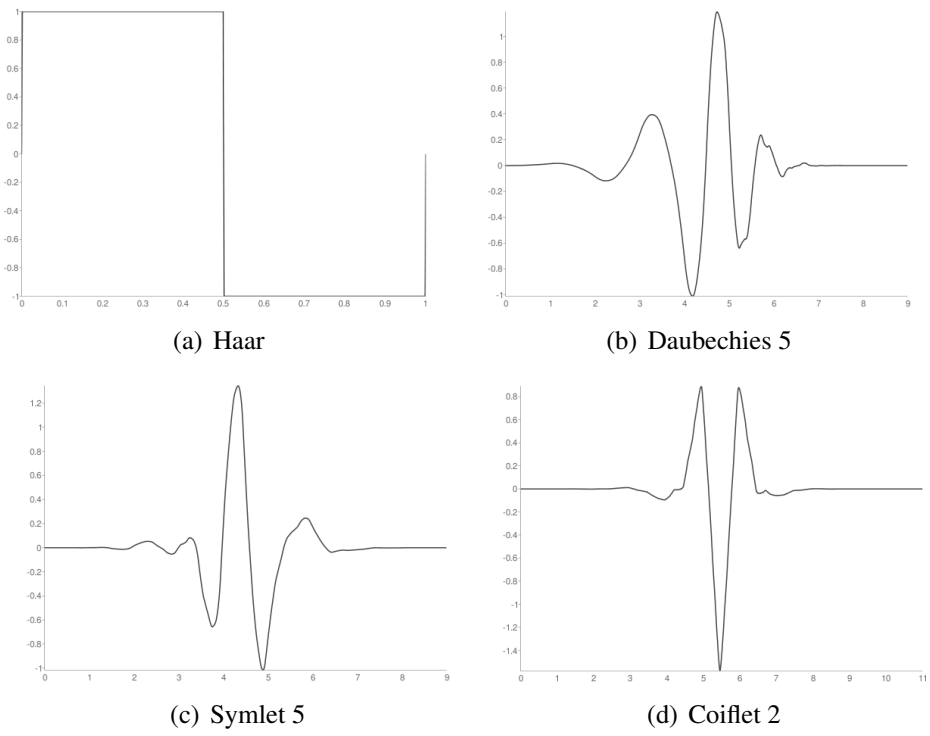
$$f(t) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} C_f(j,k)\psi_{j,k}(t) \tag{4.1}$$

$$C_f(j,k) = \int_{-\infty}^{\infty} f(t)\overline{\psi_{j,k}(t)}\,dt \tag{4.2}$$

where j and k are the scale and translation parameters for the wavelet, respectively. $\psi_{j,k}(t)$ is the wavelet function and is defined by the scaled and translated version of the mother wavelet $\psi(t)$ shown in equation (4.3).

$$\psi_{j,k}(t) = \frac{1}{\sqrt{|j|}}\psi(\frac{t-k}{j}) \tag{4.3}$$

Equation (4.4) shows the discrete representation of the wavelet. a_0 and b_0 are commonly set to 2 and 1, respectively.



(a) Haar (b) Daubechies 5

(c) Symlet 5 (d) Coiflet 2

Figure 4.1: Four different types of wavelets: (a) Haar, (b) Daubechies 5, (c) Symlet 5 and (d) Coiflet 2.

$$\psi_{j,k}(t) = a_0^{-j/2}\psi(a_0^{-j}t - kb_0) \tag{4.4}$$

To perform a discrete wavelet transform (DWT) , equation (4.2) is rewritten in matrix form as shown in equation (4.5) and we calculate the DWT coefficient ($\mathbf{w}$) as follows:

$$\mathbf{w} = \mathbf{W}f, \tag{4.5}$$

where $\mathbf{W}$ is an orthogonal matrix consisting of the wavelet basis functions. The signal can then be reconstructed by the inverse WT with equation (4.6), where $\mathbf{W}^T$ represents the transpose matrix of $\mathbf{W}$.

$$f = \mathbf{W}^T \mathbf{w} \tag{4.6}$$

4.2.2 Wavelet Denoising

Haar, Coiflet, Daubechies, Symlet, Bior, Rbior, and Dmey are some of the commonly used wavelets for noise removal [92]. Different wavelet families make tradeoffs between their compactness and smoothness. Ten levels of wavelet decomposition are available in MATLAB software and its Wavelet Toolbox for noise removal applications.

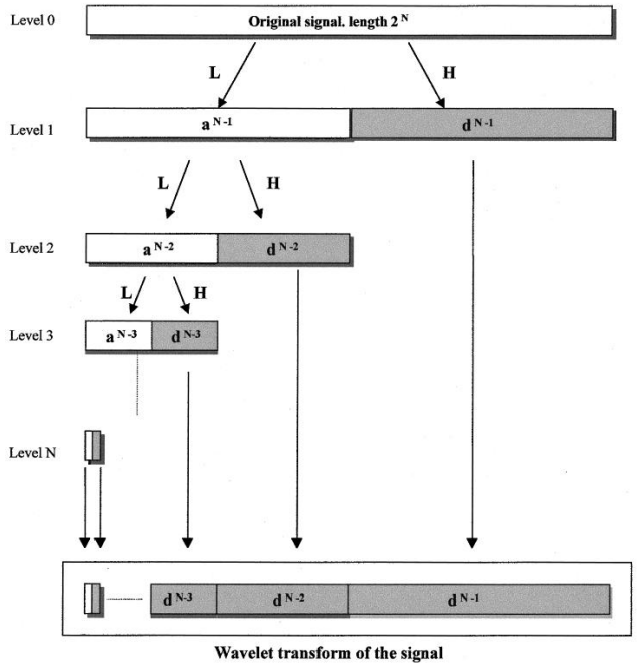


Figure 4.2: Fast decomposition of the DWT.

A fast algorithm to perform DWT was proposed by Mallat [55]. As shown in Fig. 4.2 [46], the signal (a) and noise (d) wavelet coefficients are obtained by decomposing the noisy signal by passing it through low-pass and high-pass filters. The signal coefficients (a) from the current decomposition level are used as the input for next decomposition level.

Because wavelets have excellent time localisation properties, signal peaks are decomposed into a small number of high-amplitude coefficients while noise is decomposed into many low-amplitude coefficients [46]. The removal of noise can be achieved by removing or reducing the coefficients that are smaller than a calculated threshold T. T is calculated by the threshold selection rule implemented in the MATLAB wavelet toolbox. Four common methods are available to calculate (T) for wavelet denoising [93]: fixed form, rigorous Stein's unbiased risk estimate (SURE), heuristic SURE, and minimax. The fixed form method calculates the threshold (T) for **w** using equation 4.7,

$$T = \sqrt{2 \times l_{og}(N)} \tag{4.7}$$

where N is the length of the signal. Using the rigorous SURE method, T is calculated by equation 4.8,

$$T = \sqrt{2log_e(Nl_{og2}(N))} \tag{4.8}$$

The heuristic SURE threshold is based on a combination of the rigorous SURE and the fixed method. For a signal with a high SNR, heuristic SURE uses a rigorous SURE threshold calculation method; if the SNR is low, the fixed form is used to calculate the threshold [92]. The minimax method calculates the threshold value by finding the minimax performance for mean square error [92].

There are two thresholding methods: soft and hard. In both thresholding methods, coefficients (w) smaller than the threshold (T) are set to zero. The difference between the two methods arises when the coefficients are larger than the threshold value. In the soft thresholding method, coefficients are reduced by the threshold value; in the hard thresholding method, the coefficient values are kept.

A basic WT-based denoising algorithm consists of the following steps [92]:

1. Perform DWT with equation (4.5) and obtain the coefficient vector, **w**.

2. Calculate threshold for **w**. Compare **w** to the calculated threshold, and suppress or remove elements in **w** based on the thresholding method used.

3. Reconstruct the denoised signal using equation (4.6).

4.3 Methodology

4.3.1 Overview

Our work and theory in this chapter closely resembles the theory and work of references [46,47,53,55–59,91–93]. Since there are over 4,000 wavelet parameter combinations (wavelet type, decomposition and threshold selection rule and thresholding method) that can be used to remove noise in the MATLAB tool box, we decided to use a brute force method to determine a set of reliable wavelet denoising parameters for various types of synthetic CE signals. A flowchart of our method is shown in Fig. 4.3. CE signals with various characteristics were synthesized by a signal generation function. Each synthetic signal was then subjected to wavelet denoising with all of the available wavelet denoising parameters. The wavelet parameters that resulted in the top 50 in each of the measured metrics (root mean square error, peak height error, peak shift error) were considered as candidates for the top wavelet denoising parameters. Because the noise generated was white noise and noise varied from signal to signal, this process was repeated for 30 iterations to improve reliability. At the end of 30 iteration cycles, the wavelet denoising parameters that placed in the top 50 most often were considered to be a reliable set of wavelet denoising parameters for the synthesized signal type.

4.3.2 Signal Synthesis

Synthetic signals are often used in the literature to test the effectiveness of signal processing algorithms. Not only can different types of synthetic data be rapidly generated to model experimental signals, comparison between the post-processing

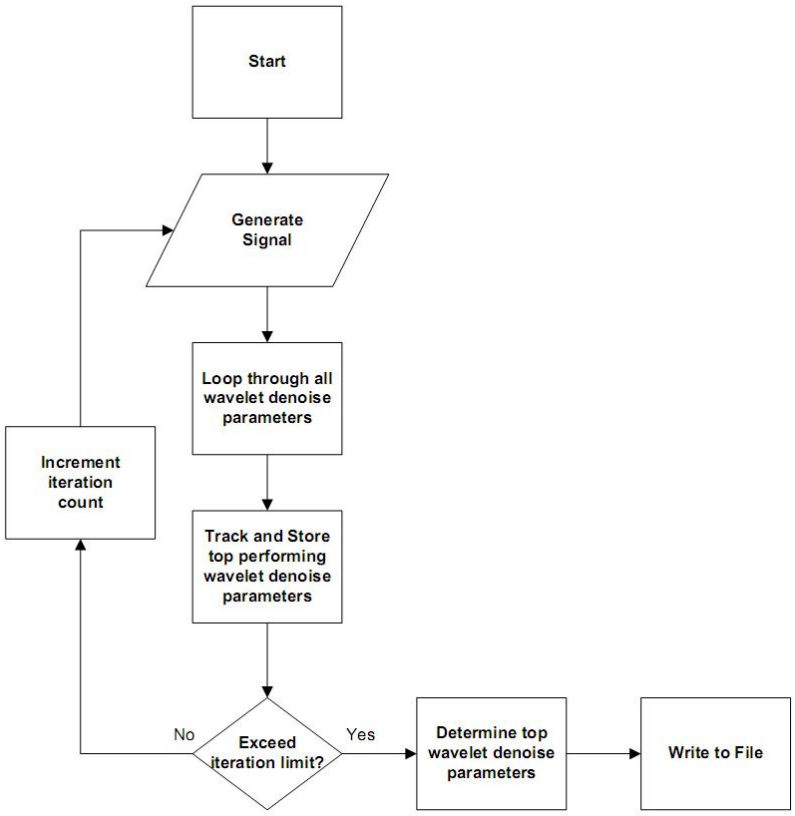


Figure 4.3: Wavelet denoising parameter flow chart.

signal and the original synthetic data (without the addition of artifacts) can be readily made. We synthesized several classes of CE signals in an attempt to mimic the different types of CE signals observed in a CE instrument. We divided the characteristics of CE signals into four types: baseline offset, peak shift, SNR, baseline variation, resolution, peak height variations. We wrote MATLAB scripts to synthesize CE signals. The scripts prompt for specific parameters such as SNR, resolution, baseline offset, and number of peaks and allowed for quick modifications and synthesis of various CE signals. The noiseless peak shape in our synthesized signal was modelled by curve fitting high SNR experimental CE signals with an empirically transformed Gaussian (ETG) function. The amount of noise in the synthetic signal was specified by the SNR parameter, which defines the ratio of the amplitude

of the maximum peak to the standard deviation of the noise. Baseline variation was added to the signal by superimposing a sinusoidal signal with the signal peaks. The amplitude of the baseline variation is dependent on the peak to baseline variation ratio (PBR). Resolution (R) of the peaks dictates the time separating peaks and is defined by equation 4.9,

$$R = \sqrt{2ln(2)}\frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \tag{4.9}$$

where μ and σ represent the peak location and the full width half maximum (FWHM) of the peaks, respectively. The height difference between peaks can be controlled by changing the peak degradation ratio (PDR), which defines the ratio between peak heights.

One of the challenges in synthesizing CE signals is accurately modelling the point spread function $h(t)$ for the TTK. An approximation of $h(t)$ can be obtained by curve fitting an ETG function to a high-SNR experimental single peak signal. The experimental signal we used to obtain $h(t)$ and all of the experimental CE signals shown in this thesis were collected by M.Sc. student Allison Bidulock using the TTK with AML's standard CE protocol as described in references [94] and [29]. 1 μL of 4% linear polyacrylamide (LPA) sieving matrix and a 3 μL 0.01x Tris TAPS-EDTA (TTE) buffer were used to fill the microfluidic chip channels. The sample used was a 1 μL of Cy-5 reverse primer (end-labelled DNA) with a diluted concentration of 0.749 ng/μL. CE was done by first injecting a voltage of 200 V (or equivalent to a electric field of 222 V/cm) for 80 seconds, then separated by a voltage of 600 V (67 V/cm) for 250 seconds. Detection was made 13 mm from the CE channel intersection.

ETG functions have been reported to model CE peaks by [95–97]; ETG functions are defined by equation (D.1),

$$h(t) = \frac{2He^{0.5}}{(1+\lambda_l e^{k_l*(t_l-t)})^{\alpha} + (1+\lambda_t e^{k_t*(t_t-t)})^{\beta} - 1}, \tag{4.10}$$

where H is the maximum peak height, t_t and t_l are the half width times for the leading and trailing edge, k, λ, α and β were used to adjust symmetrical properties of the peak. The ETG parameters used to generate $h(t)$ are summarized in Table 4.1. These parameters were manually tuned to fit the experimental signal peak's leading and trailing edge, as shown in Fig. 4.4. This $h(t)$ is used throughout this thesis as the point spread function (PSF) for the TTK.

We believe that the data synthesized are representative of real data because we modelled our PSF to fit an experimental signal's peak. We also added artifacts to mimic the signal conditions observed in the experimental CE signals we have collected using the TTK.

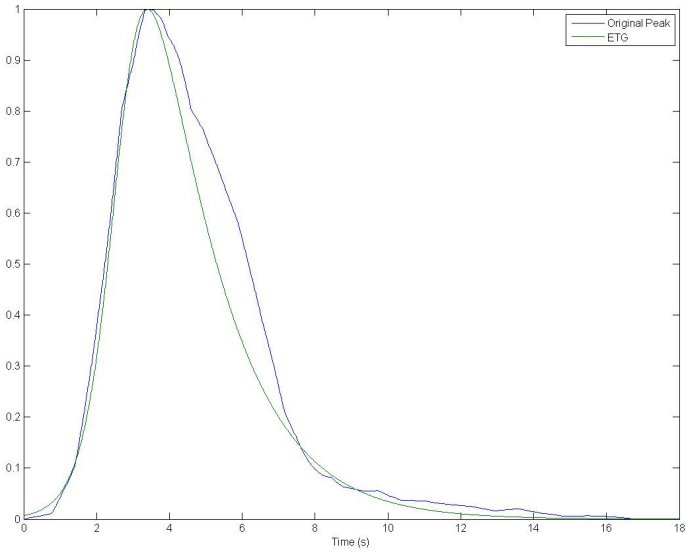


Figure 4.4: A normalized ETG function fitted to a CE run of a 1μL of Cy-5 end-labelled DNA with a concentration of 0.749 ng/μL. AML's standard CE protocol was used.

4.3.3 Wavelet Denoising Loop

The 55 wavelets, 10 decomposition levels, 4 threshold selection methods, and 2 thresholding methods available in the MATLAB wavelet toolbox is equivalent to

	Leading Edge Parameters	Trailing Edge Parameters
λ	0.8	1
k	2	2.4
t	3.2	0.6
α / β	1	1

Table 4.1: ETG parameters for modelling $h(t)$.

4,400 different combinations of wavelet denoising parameters. Instead of performing a full sweep of all the denoising settings, a subset of parameters can be used to find a set of reliable wavelet denoising parameters. We set the thresholding method to soft because a soft threshold provides smoother peaks in the post processing signals compared to hard threshold [58, 98]. Among the threshold selection methods available, the heuristic SURE method was selected because it automatically chooses between the fixed form threshold method and the rigorous SURE method based on SNR [92]. By setting the threshold selection rule and the thresholding method, the number of simulations in each iteration is reduced from 4400 to 550, which significantly reduces simulation time.

4.3.4 Wavelet Denoising Parameters Selection

Due to the randomness of the noise, we found the top wavelet denoising parameters varied in different simulations of the same signals with the same SNR. To increase reliability, thirty iterations of signal generation and wavelet denoising were performed for every signal type. Each wavelet denoising parameter combination used a counter to keep track of the number of times it yielded the top 50 metrics. The ten wavelet denoising parameters that yielded most frequently in the top 50 metrics were considered as suitable wavelet denoising parameters for the signal tested. A wavelet denoising parameter combination suitable for the most signal types was considered the most reliable wavelet denoising parameter for CE signals.

4.3.5 Evaluation Metrics

The root mean square error (RMSE), peak height error (PHE) and peak shift error (PSE) between the original signal and the post-processing signal were the metrics used to evaluate the performance of wavelet denoising parameters.

4.3.5.1 Root Mean Square Error

The RMSE was used to measure the error between the original noiseless signal and the denoised signal. The RMSE is defined by equation (4.11),

$$RMSE = \sqrt{\sum_{i=1}^{n} (x(i) - \hat{x}(i))^2}, \tag{4.11}$$

where n is the number of data points, x is the original signal and $\hat{x}$ is the denoised signal. The RMSE is useful in describing how well the post processing signal matches the original signal. The RMSE is a good method to evaluate denoising efficiency but it does not reveal information about peak preservation [99].

4.3.5.2 Peak Shift Error

Peak shift error (PSE) is a measure of peak preservation post signal denoising. The PSE is defined by equation (4.12),

$$PSE = \frac{\sum_{j=1}^{n} |p(j) - \hat{p}(j)|}{n} \tag{4.12}$$

where $\hat{p}$, p, and n represent the peak location post-processing, the true peak location, and the number of peaks, respectively. In cases where there are multiple peaks in the signal, the average of the peak shift error is used.

4.3.5.3 Peak Height Error

The peak height error (PHE) is another metric used to quantify peak preservation post signal processing and is defined by equation (4.13),

$$PHE = \frac{\sum_{j=1}^{n} \frac{|y(j) - \hat{y}(j)|}{y(j)}}{n} \tag{4.13}$$

where y, $\hat{y}$ and n represents the true and post-processing peak heights, and the number of peaks, respectively. In cases where there are multiple peaks in the signal, the average of the peak height ratio error is used.

4.4 Results

4.4.1 Wavelet Denoising

4.4.1.1 Base Signal

We synthesized a base signal with 5 peaks, a SNR of 5 V/V, a resolution of 3, and no baseline variation or offset, as shown in Fig. 4.5. Table 4.2 shows and sorts the top 10 wavelet denoising parameters by their occurrence count in the top 50 metrics. The Wavelet and Lvl columns represent the type of wavelet and the decomposition level of the wavelet denoising parameters.

We found that sym 8, sym 6, coif3, rbior6.8, and bior6.8 wavlets yielded RMSEs below 0.09, PSEs less than 0.19 seconds, and PHEs less than 2%. Note that a decomposition level of 8 is in 50% of the wavelets denoising settings for this signal with a SNR of 5 V/V.

Since 30 iterations were performed, the highest possible count for each wavelet parameter was 90 because three metrics were used to measure the performance. The top 10 wavelet denoising parameters appearance counts were calculated to be between 68 and 74 (76% to 82% of maximum count), which indicates that they were highly reliable in terms of noise removal efficiency and peak preservation. After finding some reliable wavelet denoising parameters for the base signal, variants of the base signal were used to study the effect of signal changes on the top wavelet denoising parameters.

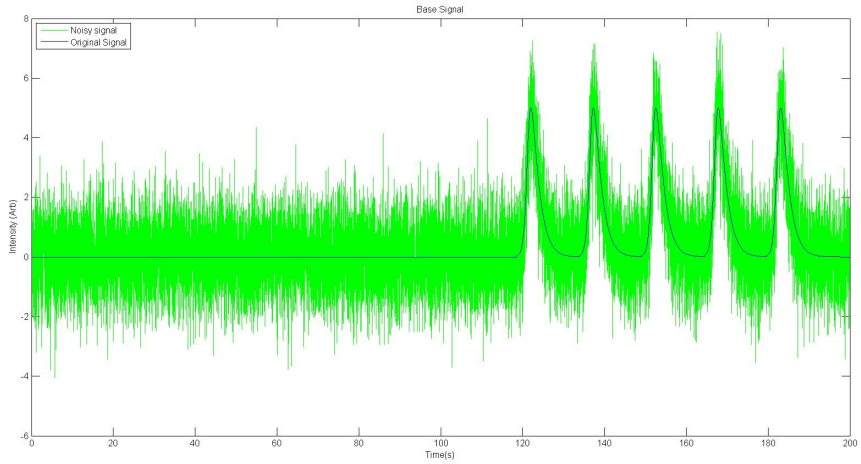


Figure 4.5: Example of the synthesised waveform of a base CE signal: 5 peaks, SNR of 5 V/V, resolution of 3, and no baseline offset or variations.

Rank	Wavelet	Lvl	Count	RMSE	PSE	PHE
1	sym8	8	74	0.0877	0.1768	0.0173
2	sym6	8	72	0.0897	0.1645	0.0171
3	bior6.8	8	72	0.0864	0.1599	0.0169
4	coif3	9	71	0.0869	0.1816	0.0172
5	sym6	9	71	0.0876	0.1673	0.0169
6	coif3	8	70	0.0889	0.1743	0.0173
7	sym6	10	70	0.0871	0.166	0.0171
8	sym8	9	70	0.0857	0.1844	0.0165
9	rbior6.8	8	70	0.0879	0.1592	0.0175
10	rbior6.8	9	68	0.0861	0.1713	0.0183

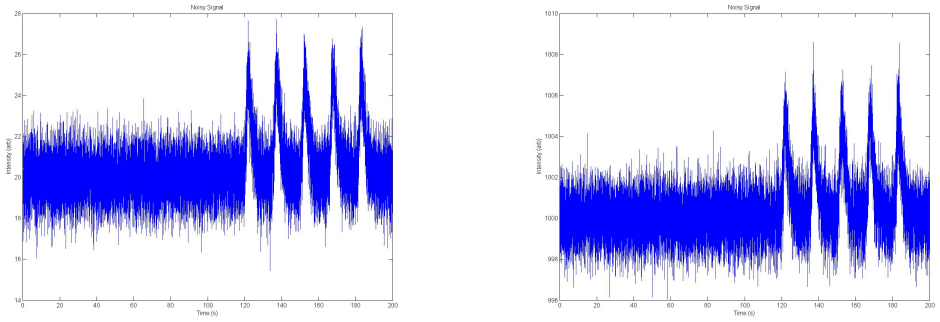
Table 4.2: Top 10 wavelet denoising parameters for the base signal

4.4.1.2 Baseline Offset

Baseline offset is a common problem in the analysis of CE signals due to autofluorescence of microfluidic chip walls. To mimic this effect, baseline offsets of 20 V and 1000 V were added to the base signal as shown in Fig. 4.6(a) and Fig. 4.6(b), respectively. The top 10 wavelet denoising parameters for the base signal with an offset of 20 V is shown in Table 4.3. This table illustrates that the top 10 wavelet

denoising parameters were the same top 10 wavelet denoising parameters found in the base signal shown in Table 4.2. The RMSEs were calculated to be below 0.09, the PSE were less than 0.18 seconds, and the PHE were no greater than 0.35%.

For the signal with an offset of 1000 V, a majority (9/10) of the top 10 wavelet denoising parameters were the same as those of the base signal, as shown in Table 4.4. The only difference was that the rbior2.8 wavelet was in the top 10 list for a base signal with an offset of 1000 V in the baseline. From these simulations, we conclude that the baseline offset in CE signals has very little effect on wavelet denoising parameters.



(a) Baseline offset of 20 V

(b) Baseline offset of 1000 V

Figure 4.6: Synthetic base signal with a baseline offset.

4.4.1.3 Peak Shift

CE peak locations depend on experimental conditions such as injection voltage, separation voltage, and the sample tested. As a result, peak locations vary from experiment to experiment and it is important determine the effects of peak shifts on wavelet denoising parameters. To investigate the effects of peak shift, we compared the top wavelet denoising parameters between the base signal and the base signal shifted by 50 seconds (Fig. 4.7). In the top wavelet denoising parameters for the shifted signal shown in Table 4.5, all except the bior5.5 wavelet are the same as the

Rank	Wavelet	Lvl	Count	RMSE	PSE	PHE
1	sym8	8	74	0.0875	0.1767	0.0032
2	bior6.8	8	72	0.0861	0.1588	0.0031
3	sym6	8	71	0.0898	0.1652	0.0031
4	bior6.8	9	70	0.0835	0.1643	0.0032
5	sym6	9	69	0.0875	0.1637	0.003
6	coif3	8	68	0.0886	0.1728	0.0034
7	sym6	10	68	0.0874	0.1642	0.003
8	bior6.8	10	67	0.083	0.1635	0.0034
9	rbior6.8	8	67	0.0879	0.1576	0.0033
10	coif3	10	66	0.0857	0.1779	0.0032

Table 4.3: Top 10 wavelet denoising parameters for the base signal with a baseline offset of 20 V.

Rank	Wavelet	Lvl	Count	RMSE	PSE	PHE
1	rbior6.8	8	77	0.0887	0.1635	0.0001
2	coif3	8	76	0.0899	0.1802	0.0001
3	sym6	8	76	0.0906	0.1709	0.0001
4	sym6	9	75	0.0888	0.1723	0.0001
5	sym8	8	75	0.089	0.1844	0.0001
6	rbior2.8	8	74	0.0896	0.1697	0.0001
7	rbior6.8	9	72	0.0865	0.1741	0.0001
8	sym6	10	71	0.088	0.171	0.0001
9	bior6.8	8	71	0.0877	0.164	0.0001
10	bior6.8	9	71	0.0856	0.1707	0.0001

Table 4.4: Top 10 wavelet denoising parameters for the base signal with a baseline offset of 1000 V.

wavelet parameters seen in the base signal. The top wavelet denoising parameters yielded a very low RMSE of less than 0.09, a PSE less than 0.24 seconds, and PHE less than 2%. Shifting the peaks in the base signal did not significantly change the top wavelet denoising parameters for the base signal.

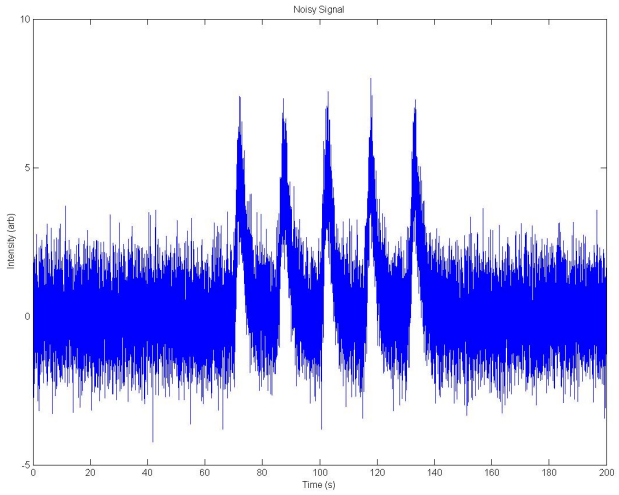


Figure 4.7: Base signal shifted by 50 seconds.

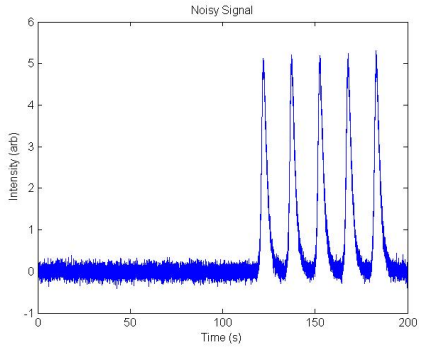
Wave	Wavelet	Wave	Wave	Wave	Wave	Wave
1	sym6	8	70	0.088	0.2319	0.0155
2	bior5.5	8	67	0.087	0.2197	0.0148
3	bior6.8	8	67	0.0852	0.2287	0.0149
4	bior6.8	9	64	0.0831	0.2307	0.0154
5	rbior6.8	8	64	0.0871	0.2241	0.0134
6	bior6.8	10	62	0.083	0.2292	0.0167
7	rbior6.8	9	62	0.0853	0.223	0.0149
8	sym6	9	61	0.0864	0.2281	0.0164
9	coif3	9	60	0.0863	0.2401	0.0158
10	bior5.5	9	60	0.0861	0.2183	0.0166

Table 4.5: Top 10 wavelet denoising parameters for the base signal shifted by 50 seconds.

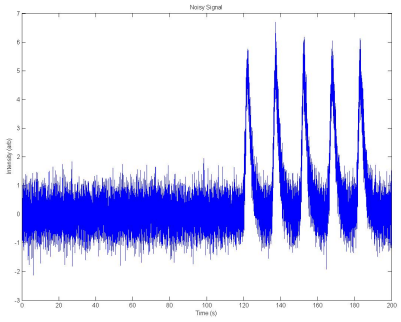
4.4.1.4 SNR Level

It is important that the wavelet denoising not only remove noise in the signal, but also preserve signal information. As shown in Fig. 4.8, CE signals with various SNR values were simulated to study wavelet denoising parameters for noisy CE signals. We found that for signals with low SNR, a high decomposition level is re-

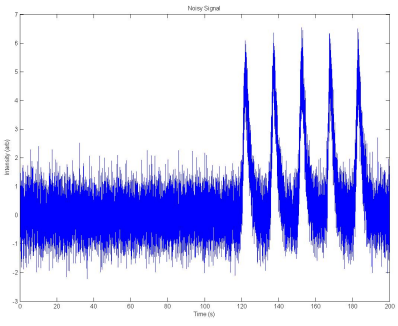
quired to remove noise. This is because WT with high decomposition breaks down the signal into smaller coefficients. The relationship between SNR and decomposition level is illustrated by the top 10 wavelet denoising parameters for a base signal with various levels of noise as shown in Tables 4.6 – 4.9.



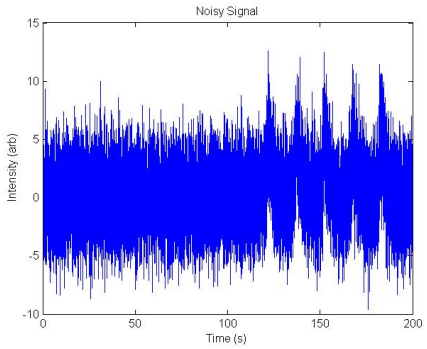
(a) Base signal with SNR of 50 V/V



(b) Base signal with SNR of 10 V/V



(c) Base signal with SNR of 8 V/V



(d) Base signal with SNR of 2 V/V

Figure 4.8: Base signal with various SNR levels.

The top wavelet denoising parameters for the base signal with a high SNR of 50 V/V (Fig. 4.8(a)) are shown in Table 4.6. We noticed that decomposition level 5 is in 9 out of the top 10 wavelet denoising parameters. For the base signal with a medium SNR of 10 V/V shown in Fig. 4.8(b), the top wavelet denoising parameters shown in Table 4.7 illustrates that level 6 decomposition provided excellent denoising capabilities as demonstrated by the low RMSE (< 0.07), PSE (< 0.12 seconds), and PHE ($< 1\%$). As the SNR of the signal decreased to 8 V/V (Fig. 4.8(c)), levels

Rank	Wavelet	Lvl	Count	RMSE	PSE	PHE
1	dmey	6	87	0.0127	0.0237	0.0021
2	dmey	5	86	0.0176	0.0507	0.0023
3	sym8	5	85	0.0177	0.0482	0.0025
4	sym6	5	84	0.0177	0.0494	0.0026
5	rbior6.8	5	84	0.0177	0.0503	0.0025
6	bior6.8	5	83	0.0176	0.0498	0.0025
7	coif4	5	82	0.0176	0.0449	0.0024
8	coif5	5	81	0.0176	0.0512	0.0022
9	db8	5	81	0.0176	0.0536	0.0022
10	sym7	5	81	0.0176	0.0506	0.0023

Table 4.6: Top 10 wavelet denoising parameters for the base signal with a SNR of 50 V/V.

Rank	Wavelet	Lvl	Count	RMSE	PSE	PHE
1	sym8	6	76	0.062	0.0987	0.0097
2	bior6.8	6	76	0.062	0.0996	0.0099
3	dmey	6	76	0.0614	0.0953	0.0097
4	db9	6	75	0.0618	0.1043	0.0093
5	coif5	6	73	0.0617	0.0914	0.009
6	db8	6	73	0.0615	0.1131	0.0089
7	db10	6	73	0.0617	0.1023	0.0096
8	sym6	6	73	0.0619	0.104	0.0098
9	coif4	6	72	0.0617	0.0938	0.0097
10	bior5.5	6	72	0.0616	0.1134	0.0095

Table 4.7: Top 10 wavelet denoising parameters for the base signal with a SNR of 10 V/V.

7 and 8 decomposition were required to provide adequate denoising (Table 4.8).

For the base signal (SNR 5 V/V) shown in Fig. 4.5, we found that levels 8 or 9 decomposition provided very good noise removal and peak preservation properties. As shown in the top 10 wavelet denoising parameters for the base signal (Table 4.2), the RMSE is below 0.09, the PSE is less than 0.19 seconds, and a PHE is less than 2%. As the SNR of the signal decreased to 2 V/V, as shown in Fig 4.8(d), we found that levels 9 and 10 wavelet decompositions were required (Table 4.9). With

Rank	Wavelet	Lvl	Count	RMSE	PSE	PHE
1	bior6.8	7	69	0.0628	0.1504	0.0145
2	rbior6.8	8	68	0.0643	0.1479	0.0138
3	sym6	8	67	0.068	0.1557	0.0127
4	sym8	8	66	0.0642	0.1703	0.0122
5	coif3	8	64	0.0666	0.1672	0.013
6	sym8	7	64	0.0637	0.1669	0.0127
7	bior6.8	8	61	0.0658	0.1505	0.0135
8	bior5.5	7	60	0.0634	0.1412	0.0176
9	rbior2.8	8	60	0.0652	0.1541	0.0147
10	sym6	9	59	0.0684	0.1567	0.0129

Table 4.8: Top 10 wavelet denoising parameters for the base signal with a SNR of 8 V/V.

Rank	Wavelet	Lvl	Count	RMSE	PSE	PHE
1	coif5	10	40	2.2089	22.2203	1.1972
2	db10	10	37	2.2136	9.5295	1.1892
3	sym7	10	37	2.2115	21.0887	1.1865
4	db8	10	34	2.2074	0.9643	1.2102
5	db10	9	34	2.2116	9.5295	1.2012
6	coif5	9	33	2.209	22.2203	1.2058
7	db10	8	33	2.2128	9.5295	1.205
8	sym7	9	33	2.2116	21.0887	1.183
9	dmey	10	33	2.2049	NaN	1.179
10	db8	9	31	2.2038	0.9643	1.2013

Table 4.9: Top 10 wavelet denoising parameters for the base signal with a SNR of 2 V/V.

a combination of the low SNR and high decomposition level, the RMSE rose above 2, peaks shifted more than 20 seconds in some cases, and peak height changed by more than 110%.

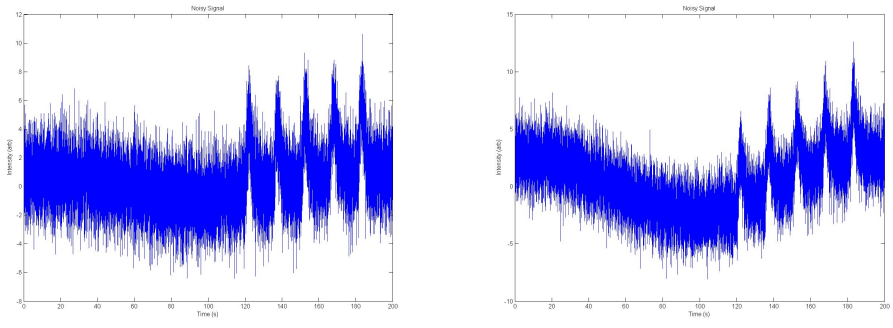
From these examples, we concluded that for signals with high SNR ($>$ 10 V/V), decomposition levels 5 and 6 are enough to remove the noise in CE signals. For signals with medium SNR (5 – 10 V/V), decomposition levels of 7 or 8 could be used to denoise signals. For signals with low SNR ($<$ 5 V/V), decomposition levels

of 8 or 9 were required. For signals with SNR below 2 V/V, we found that wavelet denoising did not preserve the signals.

4.4.1.5 Baseline Variations

Due to unstable light source and autofluorescence of microfluidic chip walls, baseline variation is observed in the CE data collected with the TTK [29]. To investigate the effect of baseline variation on wavelet denoising parameters, we analysed the base signal with peak-to-baseline variation ratio (PBR) of 3 and 1 as shown in Fig. 4.9(a) and Fig. 4.9(b), respectively.

As shown in Table 4.10 and Table 4.11, the top wavelet denoising parameters for baseline variations of PBR of 3 and 1 are similar to those of the base signal. In terms of the measured metrics, for the base signal with a PBR of 3 (Table 4.10) , the RMSEs are below 0.09; the PSEs are less than 0.2 seconds; and the PHE is less than 2%. In baseline variations with a PBR of 1, the maximum PSE slightly exceeds 0.2 seconds and the maximum PHE is slightly above 2% and with RMSEs below 0.09 (Table 4.11). The top wavelet denoising parameters did not change with baseline variation; however, the peak preservation metrics varied slightly for the base signal with a PBR of 1.



(a) Peak-to-baseline variation ratio of 3 (b) Peak-to-baseline variation ratio of 1

Figure 4.9: Base signals with baseline variation.

Rank	Wavelet	Lvl	Count	RMSE	PSE	PHE
1	sym8	8	79	0.0871	0.1844	0.0178
2	coif3	8	77	0.0885	0.1744	0.0169
3	sym8	9	76	0.085	0.1906	0.0179
4	rbior6.8	8	76	0.0876	0.1592	0.017
5	coif3	9	75	0.086	0.1804	0.0175
6	sym6	8	75	0.0894	0.1639	0.016
7	sym6	9	74	0.0873	0.1659	0.0166
8	bior6.8	8	74	0.0859	0.1651	0.017
9	coif3	10	73	0.0853	0.1803	0.0177
10	rbior6.8	9	73	0.0856	0.1702	0.0171

Table 4.10: Top 10 wavelet denoising parameters for the base signal with a PBR of 3.

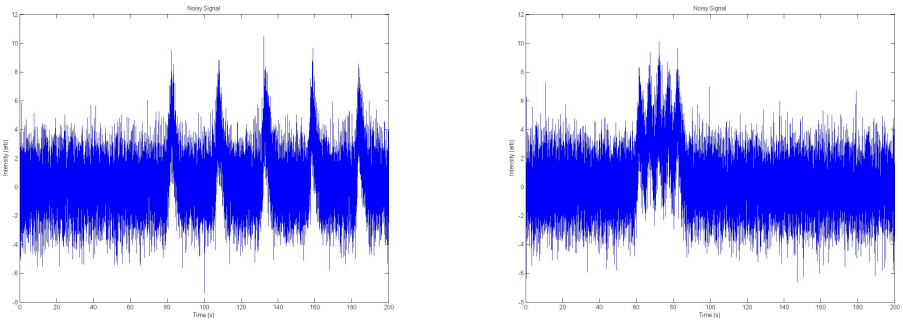
Rank	Wavelet	Lvl	Count	RMSE	PSE	PHE
1	sym8	8	86	0.089	0.198	0.0193
2	coif3	8	85	0.0904	0.1822	0.0197
3	coif3	9	85	0.0878	0.188	0.0201
4	sym8	9	84	0.0867	0.2024	0.0192
5	sym6	8	83	0.0912	0.1704	0.0197
6	sym6	9	82	0.0891	0.173	0.0199
7	coif3	10	80	0.087	0.1874	0.0199
8	bior6.8	8	80	0.0876	0.171	0.0195
9	rbior6.8	8	79	0.0893	0.1606	0.0215
10	rbior6.8	9	77	0.0867	0.1691	0.0205

Table 4.11: Top 10 wavelet denoising parameters for the base signal with a PBR of 1.

4.4.1.6 Resolution

Peak resolution is defined as the ratio of the distance between two peaks and the average peak width as shown in equation 4.9 [37]. Because CE signal resolution decreases as the size of the microfluidic chip decreases, it is important to investigate the effects of high and low signal resolution on the top wavelet denoising parameters. For a base signal with high resolution of 5, shown in Fig. 4.10(a),

we found that db6, db10, and rbior6.8, and bior6.8 wavelets provided the lowest RMSE, PSE and PHE as shown in Table 4.12. All of the top wavelet denoising parameters yielded RMSEs below 0.1, PSEs less than 0.26 seconds, and PHEs less than 2%. For a base signal with low resolution of 1, as shown in Fig. 4.10(b), we found that coif4, coif5, db7, and sym8 wavelets were the most suitable for wavelet denoising providing RMSEs below 0.081, PSEs less than 0.16 seconds, and PHE of no more than 1.7% (Table 4.13). The top wavelet denoising parameters strongly depend on the resolution of signals.



(a) High resolution of 5

(b) Low resolution of 1

Figure 4.10: Various resolution signals types

4.4.1.7 Peak Height Variation

Peak height varies in different CE signals and it is important to identify the relationship between peak height variations and the top wavelet denoising parameters. We used the peak degradation ratio (PDR) to change the peak height variation for synthetic CE signals. Fig. 4.11 shows the base signal with a PDR of 1.7. In order to prevent the 4th and 5th peaks from dipping below the noise threshold, we increased the SNR of the signal to 10 V/V.

The top 10 wavelet denoising parameters for the base signal with a PDR of 1.7 are shown in Table 4.14. The top wavelets are db7, sym6, sym8, coif 3-5, bior5.5,

Rank	Wavelet	Lvl	Count	RMSE	PSE	PHE
1	db10	8	60	0.0946	0.2479	0.0153
2	db10	10	58	0.0938	0.2487	0.0195
3	db6	8	56	0.0955	0.2782	0.0172
4	db10	9	55	0.0939	0.2449	0.0172
5	bior6.8	8	55	0.0966	0.231	0.0159
6	db6	9	53	0.0948	0.277	0.016
7	db10	7	53	0.0982	0.2587	0.0148
8	rbior6.8	8	53	0.0973	0.252	0.0165
9	db6	10	52	0.0948	0.2764	0.015
10	bior6.8	9	51	0.0961	0.2331	0.0171

Table 4.12: Top 10 wavelet denoising parameters for the base signal with a resolution of 5.

Rank	Wavelet	Lvl	Count	RMSE	PSE	PHE
1	coif5	8	86	0.0772	0.0773	0.0152
2	coif4	8	85	0.0757	0.1184	0.0148
3	db7	8	84	0.0808	0.1573	0.0155
4	sym8	8	83	0.0763	0.1798	0.0155
5	coif4	9	76	0.0702	0.1123	0.0153
6	coif5	9	75	0.0718	0.08	0.0148
7	coif4	10	73	0.0688	0.1122	0.0149
8	coif5	10	72	0.0709	0.0804	0.0144
9	db7	9	72	0.0755	0.1496	0.0159
10	db7	10	71	0.0742	0.1491	0.0162

Table 4.13: Top 10 wavelet denoising parameters for the base signal with a resolution of 1.

bior 6.8, rbior2.8, and rbior 3.5. With these wavelets and a decomposition level of 7, the RMSEs were below 0.05, the PSEs were less than 0.3 seconds, and PHE were less than 3.1%.

4.4.1.8 Top Wavelet Denoising Parameters

The most reliable wavelets for CE signals with various characteristics is summarized in Table 4.15. The left most column shows the various signal conditions

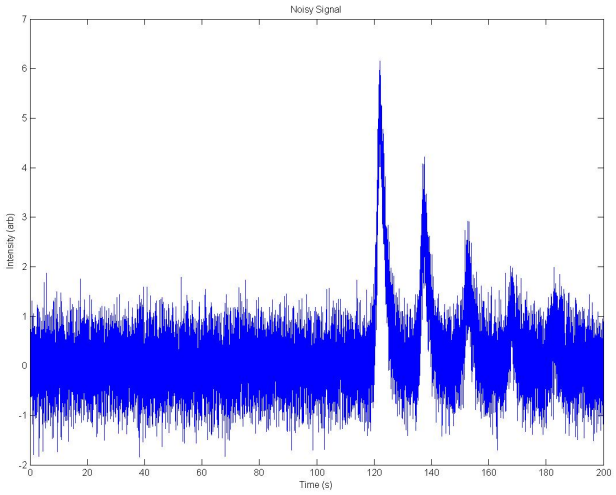


Figure 4.11: Base signal with PDR of 1.7.

Rank	Wavelet	Lvl	Count	RMSE	PSE	PHE
1	db7	7	83	0.0465	0.2247	0.0275
2	sym8	7	83	0.0451	0.1923	0.028
3	coif3	7	82	0.0454	0.1909	0.0289
4	sym6	7	82	0.0457	0.1808	0.0291
5	bior5.5	7	82	0.0467	0.1716	0.0302
6	bior6.8	7	82	0.0454	0.1767	0.0285
7	rbior6.8	7	82	0.0462	0.1786	0.0288
8	coif4	7	81	0.0446	0.2472	0.0276
9	rbior2.8	7	81	0.0466	0.1867	0.0287
10	coif5	7	76	0.0447	0.2851	0.0283

Table 4.14: Top 10 wavelet denoising parameters for the base signal with a PDR of 1.7.

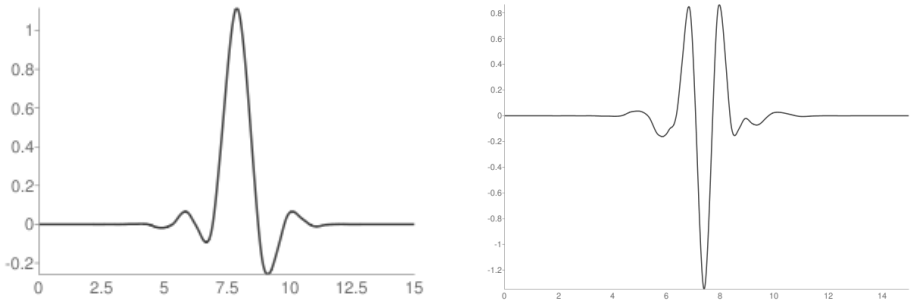
tested and the top row describes the severity of the conditions (as described in the previous sections). The body of Table 4.15 shows the wavelets that appeared in the top 10 for each simulation. Although there was no single wavelet that provided the best noise removal and peak preservation performances for all signals, the sym8 wavelet was consistently placed in the top wavelets. The sym8 wavelet placed in the top metrics for all signal types except for the signal with high resolution or low

	High	**Medium**	**Low**
Baseline Offset	sym6, **sym8**, coif3, bior6.8, rbior2.8, rbior6.8	sym6, **sym8**, coif3, bior6.8, rbior6.8	sym6, **sym8**, coif3, bior6.8, rbior6.8
Peak Shift	sym6, **sym8**, coif4, bior5.5, bior6.8, rbior6.8	-	sym6, **sym8**, coif3, bior6.8, rbior6.8
SNR	dmey, **sym6-8**, rbior 6.8, bior 6.8, coif4 - 5, db7	dmey, sym6, **sym8**, coif4-5, bior6.8, bior3.9, db8-10	dmey, sym7,coif5, db8, db10
Baseline Variation	sym6, **sym8**, coif3, rbior6.8, bior6.8	sym6, **sym8**, coif3, bior6.8, rbior6.8	sym6, **sym8**, coif3, bior6.8, rbior6.8
Resolution	db6, db10, bior6.8, rbior6.8	sym6, **sym8**, coif3, bior6.8, rbior6.8	coif4 - 5, **sym8**, db7
Peak Height Variation	db7, **sym8**, rbior2.8, rbior6.8	-	sym6, **sym8**, coif3, bior6.8, rbior6.8

Table 4.15: Top wavelets for synthesized signals with various signal conditions.

SNR. For its ability to remove noise for various levels of SNR while preserving peak information, the sym8 wavelet was the most reliable wavelet for the signals we synthesized and tested. We believe that the reason sym8 wavelets works well in wavelet denoising for various CE signals is because the shape of the wavelet scaling function (Fig. 4.12(a)) and the near symmetric property of the wavelet basis function (Fig. 4.12(b)) match with CE peaks. Both these functions are also smooth with no discontinuities or sharp edges, like CE signals.

We tested various SNR levels of the base signal to determine the most reliable decomposition level. Table 4.16 shows the most reliable wavelet decomposition levels against SNR. In general, the higher the SNR, the lower the decomposition level required. For a signal with SNR less than 5 V/V, denoising using wavelet transform causes significant PSE and PHE because a level 9 or higher decomposi-



(a) Sym8 scaling function (b) Sym8 wavelet basis function

Figure 4.12: Sym8 wavelet and scaling functions

SNR (V/V)	Decomposition Level Required
>10	5, 6
7 to 10	7, 8
5 to 6	8, 9
<5	9, 10

Table 4.16: Wavelet decomposition level requirements for various SNR

tion is required. We believe that eight levels of decomposition provided the equivalent of a low-pass filter with a cut-off frequency that is compatible for removal of high-frequency noise in CE signals with low-frequency peaks.

In conclusion, the sym8 wavelet provided the most reliable wavelet denoising metrics in terms of peak preservation and noise removal capabilities for all types of signals except for the signal with high SNR and low resolution. Decomposition level 8 provided the most versatile and reliable denoising for low SNR signals. Although level 9 and 10 decomposition can remove more noise for signals with very low SNR, they cause significant peak shifts and peak height errors in the post-processing signal. It is interesting to note that for one-dimensional discrete signals, such as CE signals, the decomposition level is analogous to low-pass filtering (LPF): increasing the decomposition level lowers the cut-off frequency of the LPF.

4.4.1.9 Noise Removal of Synthetic Signals with Wavelet Transform

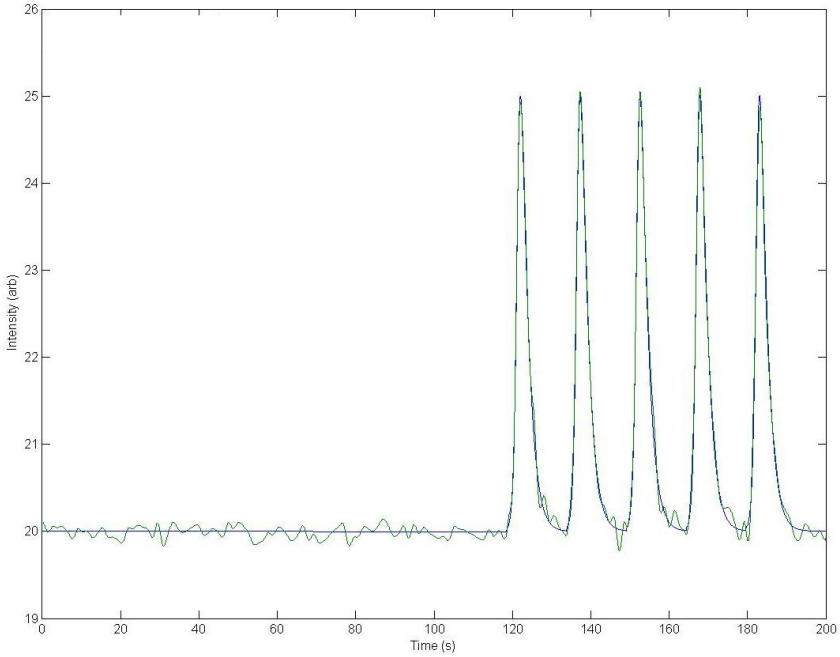
Figs. 4.13 – 4.18 illustrate the original noiseless and the denoised versions of the synthetic signals tested in Section 4.4.1. In all of these figures, the noiseless signal is shown by the blue line and the denoised signal is shown by the green line. By visual inspection, these figures showed that the sym8 wavelet with level 8 decomposition could be used to remove noise in various CE signals.

4.4.2 Comparison of Denoising Methods

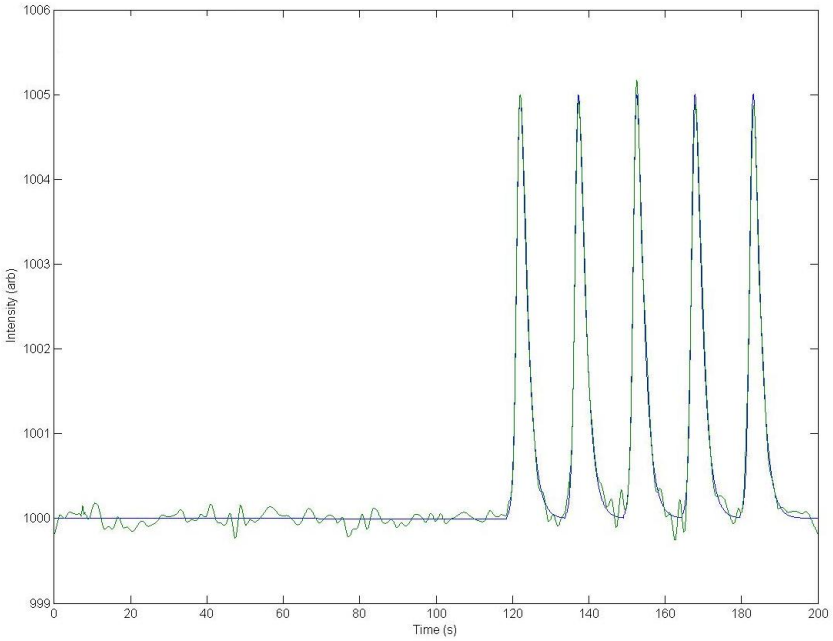
We used the synthetic base signal shown in Fig. 4.5 to compare the performance of various noise removal methods such as LPF, SG smoothing, MA. The base signal was first denoised by a sym 8 wavelet with level 8 decomposition. The wavelet denoised signal is shown in Fig.4.20(a) and the removed noise is shown in Fig. 4.20(b).

We performed FT on the base signal and its frequency contribution as shown in Fig. 4.19. Since we observed that most of the signal power is in the very low frequency band, we designed a digital finite impulse response (FIR) equripple LPF with cut-off and stop band frequencies at 1.2 Hz and 1.5 Hz using MATLAB's filter analysis toolbox. This filter is the same filter used in a recent publication from the AML [29]. Fig. 4.20(c) shows the signal denoised with the LPF method. By visual inspection, the LPF denoised signal is noisier than the wavelet denoised signal. Comparing the removed components shown in Fig. 4.20(b) and Fig. 4.20(d), we observed that the LPF removed peak information in the peak regions, whereas in wavelet denoising, very little or no peak information was removed.

To determine the best MA smoothing window and SG polynomial degree and smoothing window, we performed parametric sweeps and found the smoothing windows that yielded the lowest RMSE between the original noiseless signal and the denoised signal. For SG smoothing, we set the polynomial to 10 and found that a 4.77 seconds smoothing window provided the lowest RMSE. For MA, we found



(a) Baseline offset of 20 V



(b) Baseline offset of 1000 V

Figure 4.13: Denoised signal with baseline offsets.

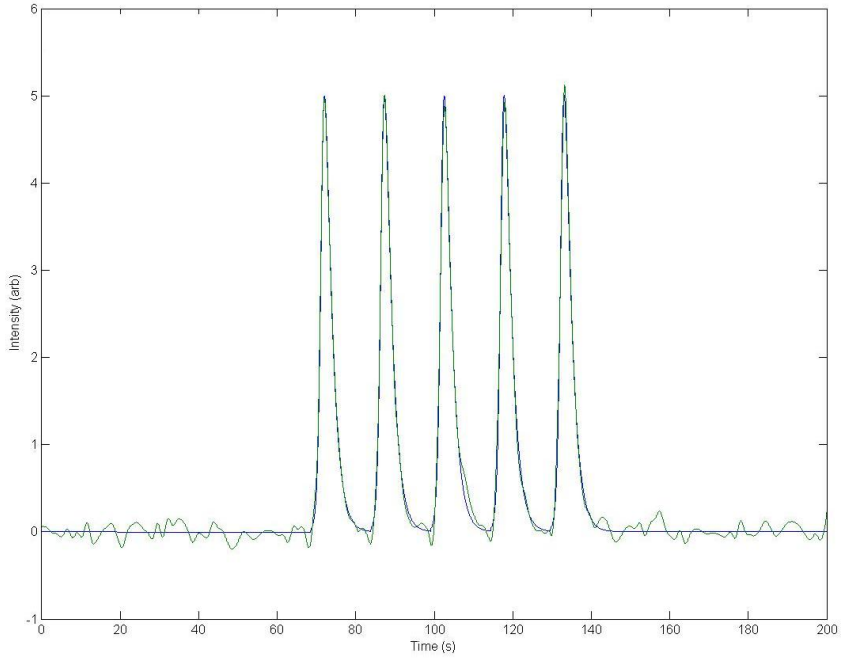


Figure 4.14: Denoised signal with peak shift.

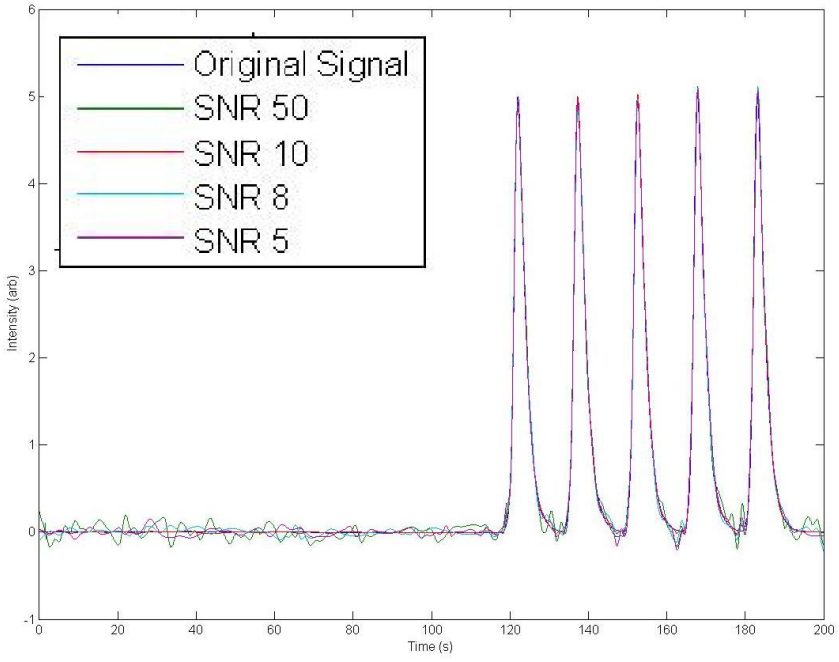
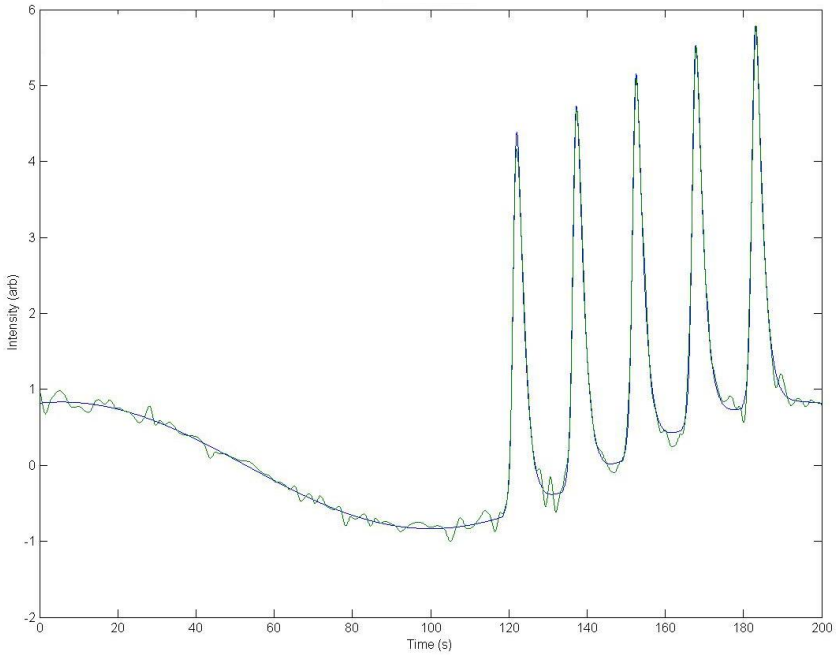
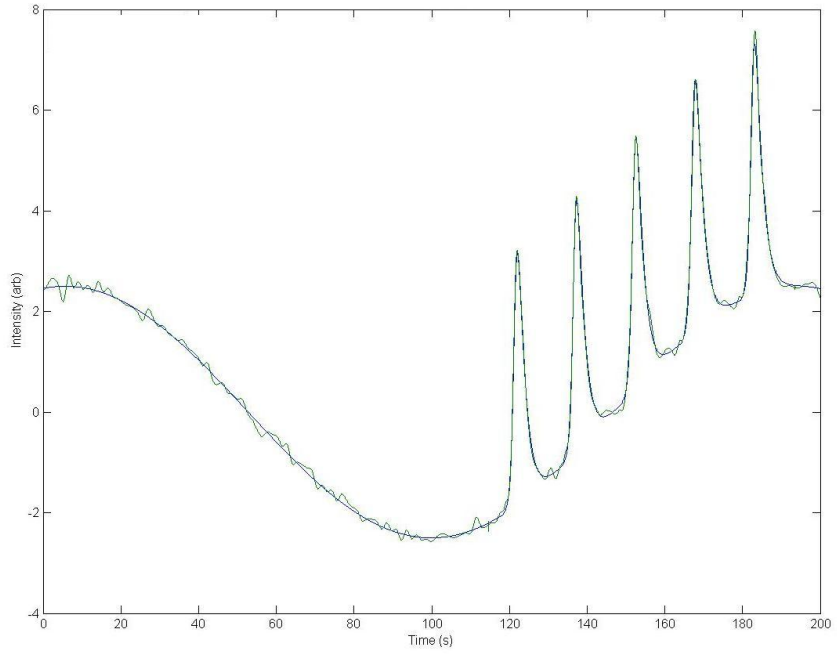


Figure 4.15: Denoised signal with various SNR levels.

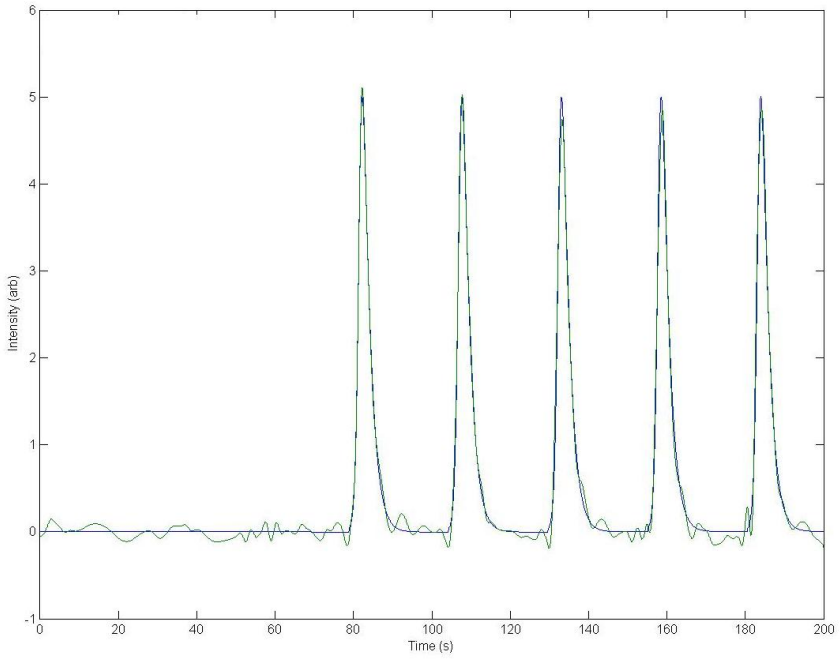


(a) PBR of 3

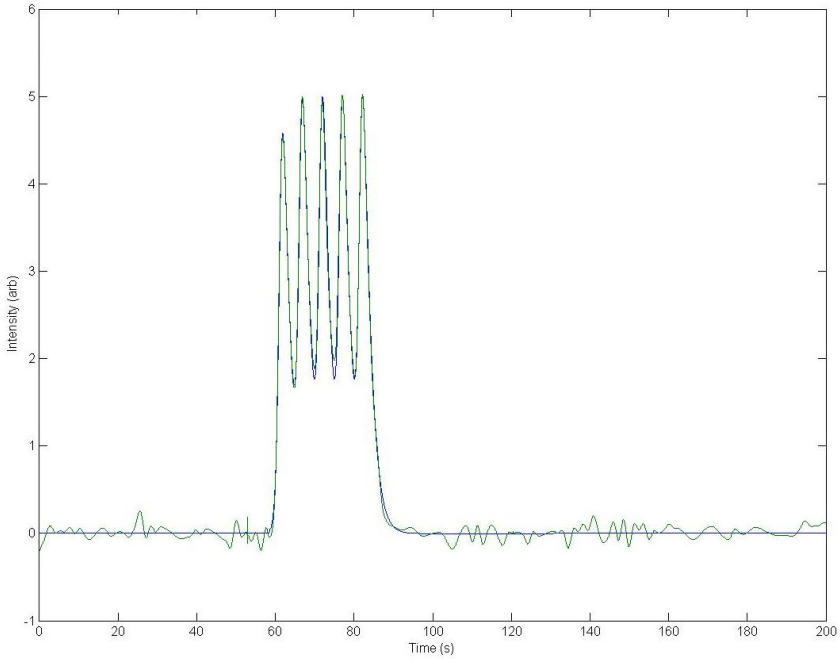


(b) PBR of 1

Figure 4.16: Denoised signal with baseline variations.



(a) Resolution of 5



(b) Resolution of 1

Figure 4.17: Denoised signal with different resolutions.

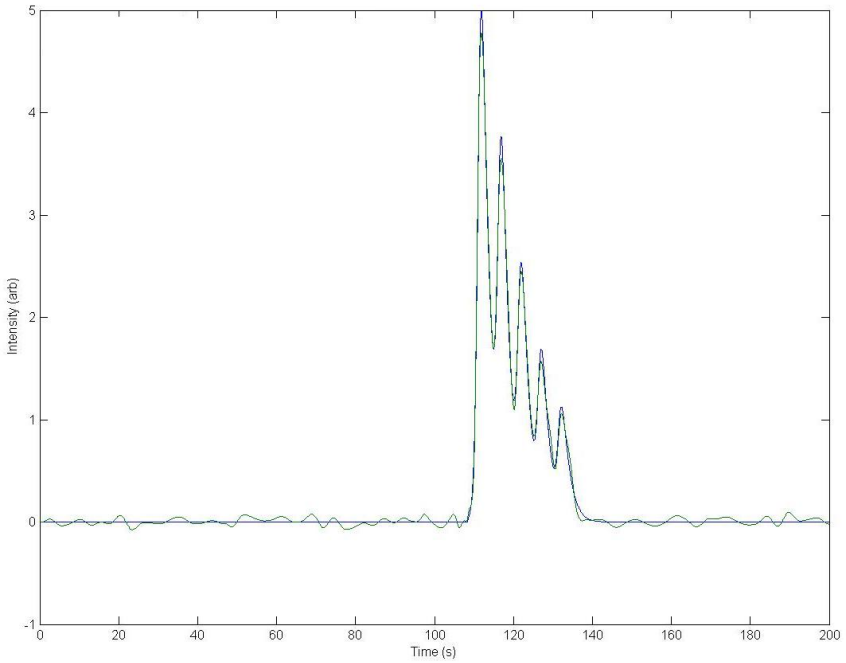


Figure 4.18: Denoised signal with peak height differences.

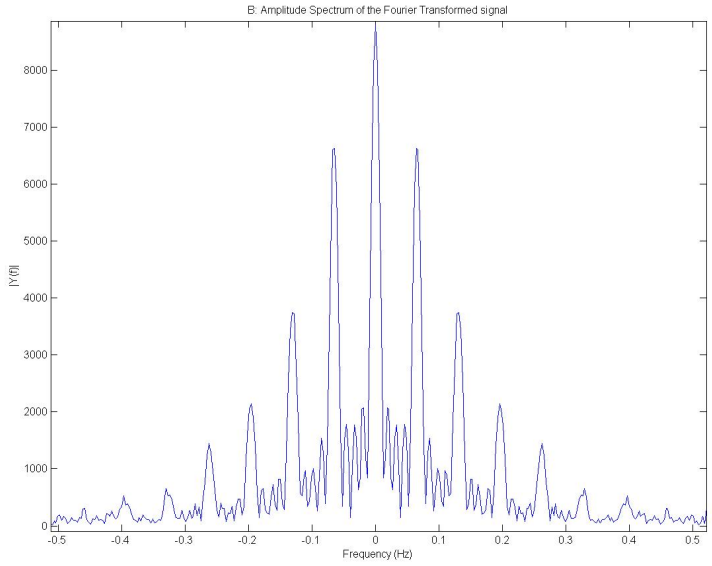
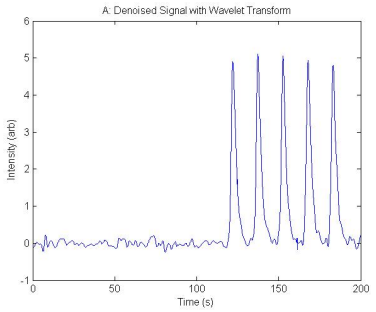
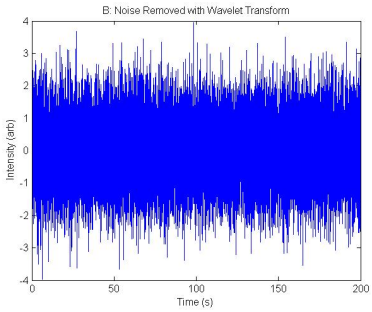


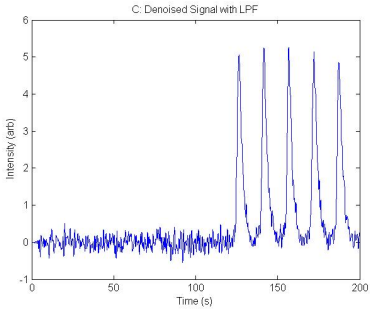
Figure 4.19: Frequency response of the base signal



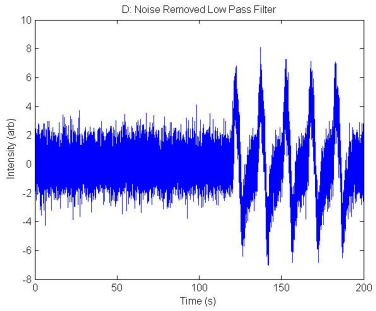
(a) Denoised signal with WT



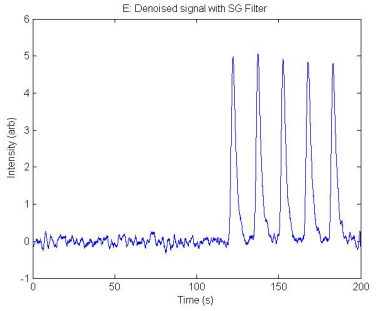
(b) Noise removed with WT



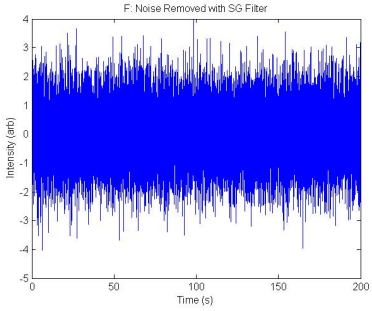
(c) Denoised signal with LPF



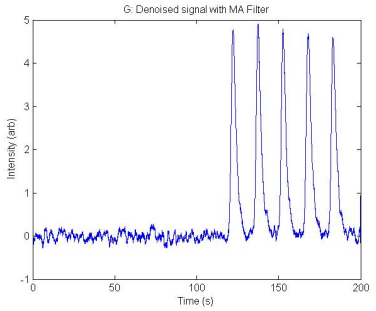
(d) Noise removed with LPF



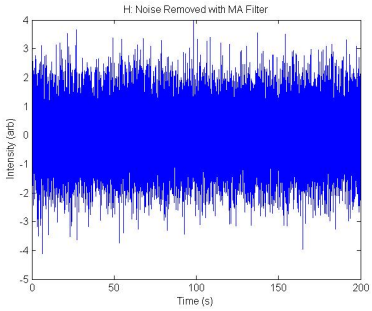
(e) Denoised signal with SG smoothing



(f) Noise removed with SG smoothing



(g) Denoised signal with MA



(h) Noise removed with MA

Figure 4.20: Noise removal and peak preservation: Comparison of various noise removal methods.

	WD	LPF	SG	MA
RMSE	0.0825	1.4983	0.0855	0.0994
Noise	0.0512	0.1421	0.0758	0.0742

Table 4.17: RMSE and noise comparison between different denoising methods

that a 1.23 second moving average window resulted in the lowest RMSE. Both SG and MA smoothing did not remove peak components, as shown in Fig. 4.20(f) and Fig. 4.20(h); however, the SG (Fig. 4.20(e)) and MA (Fig. 4.20(g)) denoised signals are noisier than the wavelet denoised signal (Fig. 4.20(a)).

Table 4.17 summarizes the baseline noise and the RMSE (between the noiseless and the denoised signals) for the denoising methods discussed in this section. This table shows that wavelet denoising (WD) achieved the lowest RMSE and the lowest baseline noise among the denoising methods tested. Fourier filtering produced the worst results because we did not adjust the cut-off frequency of the LPF for the signal tested. Even though MA and SG can achieve RMSEs and baseline noise similar to wavelet denoising, the parameters used for SG and MA were specifically tuned and optimized for the signal tested. During a CE run, the user cannot tune the signal processing parameter to obtain the best result.

Table 4.18 shows the post-processing peak locations for the denoising methods described. We found that the average peak shift using wavelet denoising was 0.062 seconds, which was less than the peak shift of the LPF (4.18 seconds), SG (0.1 seconds), and MA (0.12 seconds).

Table 4.19 shows the PHE for the various denoising methods described. Although the PHE for wavelet denoising (3.00%) is slightly higher than the PHE for the LPF (2.31%) and SG (2.88%), we found that wavelet denoising provided a much better post signal-processing SNR, as shown in Table 4.20.

The denoising capability of LPF, SG, and MA can be increased by lowering the cut-off frequency or widen the smoothing window, but these manipulations can distort the peak shape [47]. Wavelet denoising is an alternate method to remove

Peak #	True Peak Location (s)	WD (s)	LPF (s)	SG (s)	MA (s)
1	122.04	126.01	126.43	122.13	122.1
2	137.3	137.3	141.44	137.43	137.35
3	152.56	152.59	157.05	152.65	152.66
4	167.82	167.93	171.78	167.94	167.99
5	183.08	183.22	186.99	183.15	183.29
PSE (s)		0.062	4.178	0.1	0.118

Table 4.18: Peak shift error for different denoising methods.

Peak #	True Peak Height	WD	LPF	SG	MA
1	5	4.63	4.87	4.74	4.55
2	5	5.00	5.26	5.03	4.87
3	5	4.83	4.99	4.84	4.73
4	5	4.92	4.87	4.87	4.71
5	5	4.87	4.95	4.86	4.72
PHE (%)		3.00	2.31	2.88	5.68

Table 4.19: Peak height error comparison for different denoising methods.

noise without distorting the signal shape.

4.4.3 Experimental Signals

We used the sym8 wavelet with level 8 decomposition wavelet denoising to remove noise in various experimental CE runs collected with the TTK. All of the experimental data collected with the TTK in this thesis were collected by M.Sc student

Peak #	WD	LPF	SG	MA
1	69.6	37.0	57.0	54.6
2	75.1	40.0	60.5	58.5
3	72.5	37.9	58.1	56.8
4	73.9	37.0	58.6	56.5
5	73.1	37.6	58.4	56.7
Average SNR (V/V)	72.8	37.9	58.5	56.6

Table 4.20: Post processing SNR comparison for different denoising methods.

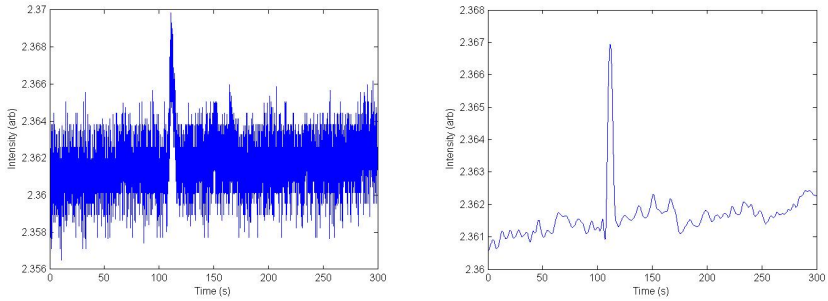
Allison Bidulock using our standard CE protocol as described in Section 4.3.2 and references [94] and [29]. We showed four examples of pre-processing and post-processing signals in Fig. 4.21. Fig. 4.21(a) and Fig. 4.21(c) show two CE runs of Cy-5 primer (end-labelled DNA) with a concentration of 0.749 ng/uL. Wavelet denoising removed the noise in these single peak signals with low SNR and baseline variation.

Fig. 4.21(e) shows the raw data of a CE run of a 0.5 μL BK virus PCR product from a thermo-cycler. The first peak is the primer peak and has a length of 25 or 26 base pairs (bp). The second peak is the product peak at 299 bp. We believe the third peak is an unspecified peak due to contamination or artifacts from the unoptimised PCR recipe. This signal has high resolution with no overlapping peaks and wavelet denoising was able to remove the high-frequency noise in this signal. Wavelet denoising also removed the noise in a low resolution CE signal. The sample tested was a 1μL DNA red ladder (11 peaks) with lengths from 50 to 550 bp, each peak separated by 50 bp.

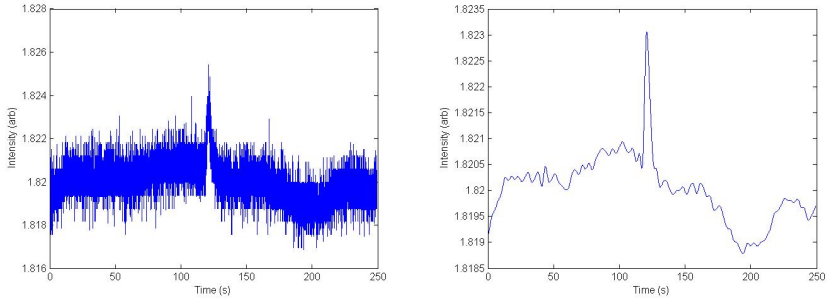
In these examples, we demonstrated that wavelet denoising can remove the high-frequency noise in experimental CE signals with various levels of baseline variations, SNR, and resolutions. However, wavelet denoising was not able to remove the low frequency baseline variations or improve the resolution of the signal. We will explore alternative methods to resolve these issues in Chapter 5 and 6 of this thesis.

4.5 Conclusion

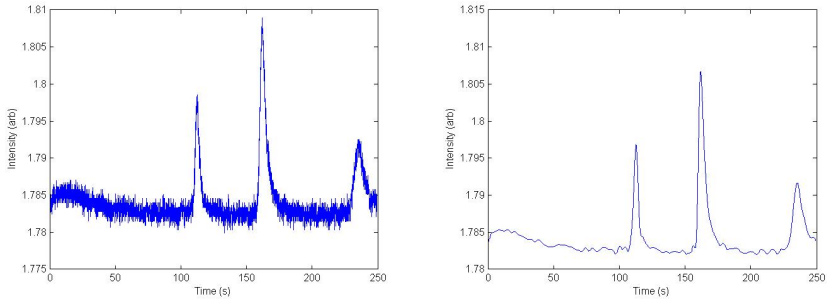
A method to determine a set of reliable wavelet denoising parameters for CE signals is presented in this chapter. Among the wavelet denoising parameters tested, we found the Symlet 8 wavelet with level 8 decomposition provided the most reliable denoising for the synthetic CE signals we have examined. Wavelet denoising with these parameters removed noise in both synthetic and experimental CE



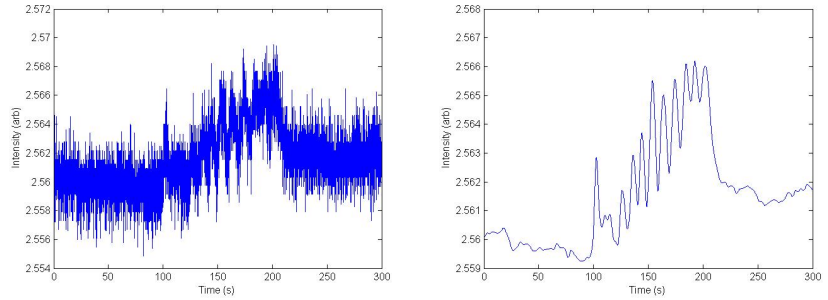
(a) Low SNR signal (b) Low SNR denoised signal



(c) Baseline varying signal (d) Baseline varying denoised signal



(e) High resolution signal (f) High resolution denoised signal



(g) Low resolution signal (h) Low resolution denoised signal

Figure 4.21: Experimental CE signals denoised with sym8 wavelet and a level 8 decomposition.

signals with various levels of SNR, baseline variation, baseline offset, resolution, and peak shifts. Without additional parameter changes, wavelet denoising obtained lower RMSE, PSE, and PHE between the true signal and the denoised signal compared to traditional methods such as FT, SG, and MA. We were able to remove the high-frequency (such as shot and thermal) noise in various CE signals with wavelet denoising.

Chapter 5

Baseline Variation Removal

5.1 Introduction

This chapter focuses on the baseline variation removal aspect of our signal processing algorithm for capillary electrophoresis (CE) signals. Because the TTK uses a non-confocal detection approach, the photo-detector collects more scattered excitation light compared to confocal methods. This type of system is more sensitive to laser intensity fluctuations which causes a baseline variation in electropherograms and makes it difficult to distinguish from fluorescent signal peaks [29]. This type of noise usually has low frequency compared to shot and thermal noise discussed in chapter 3 and 4. Slow varying baseline variation with a high amplitude resembles the shape of DNA peaks. The central challenge is to automatically identify and to remove the fluctuations in the baseline that may resemble a passing DNA peak. Using wavelet transform (WT), we were not able to remove the low frequency baseline variations in CE signals. The priority for the WT was to remove the high-frequency Gaussian-like white noise. The reason we are not using wavelet transform to remove baseline variations in CE signals is because other people in literature had already studied this problem and found that WT can only remove low baseline variations or linear baseline variations [46, 47, 66]. Thus, research of baseline variation removal using WT was not worth the effort. We are exploring alternate techniques that are applicable for the removal of baseline signal variations in the TTK.

5.2 Baseline Variation Removal Algorithms

5.2.1 Iterative Polynomial Baseline Fit

Gan *et al.* showed that baseline variations in electropherograms can be removed by an iterative polynomial curve fitting (IPBF) technique [60]. The idea is that since any observed CE signal $y(t)$ can be represented by equation (5.1):

$$y = b + s + n + \varepsilon \tag{5.1}$$

where b, s, n, and ε represent baseline variations, signal peak, noise and measurement errors, respectively. If n and ε are small and s is removed from equation (5.1), then b can be estimated using the measured data, y. The idea of IPBF is to fit a low order polynomial to y_{k-1} to obtain an estimate of b_k, where k is the iteration level. Peak removal is done by setting y_k to b_k if y_k is greater than b_k in peak regions. This process is repeated until the difference between the current and the previous baseline estimates ρ is smaller than a pre-specified error factor ρ_{thr} or it reaches a maximum iteration count k_{max}. The initial estimate of the baseline b_1 is obtained by fitting a low-order polynomial to y. The pseudocode for the IBPF is outlined in Appendix B. However, the IPBF requires the start and end coordinates of each peak region. If these coordinates are not properly identified, then the IPBF baseline estimate will not converge to the actual baseline variation. Since IPBF is an iterative-based method, it cannot perform real time processing. With the baseline variation removed from CE signals, signals peaks can be more easily identified.

5.2.2 Peak Region Detection

Finding the peak region coordinates is crucial for automation of baseline variation removal because the IPBF requires the start and end coordinates of the peak regions. The first derivative of the signal measures the rate of change of the signal and can be used to calculate the peak regions. In an ideal world, y' alone can be used to

find peak regions [69]. But in reality, using the first derivative threshold alone to determine the peak region often leads to false peak region detection due to noise in the signal. To address this problem, we have implemented a peak region detection algorithm for CE signals by incorporating error checking mechanisms.

Our peak region detection algorithm is based on the threshold crossings in the first derivative. Error checking mechanisms were employed to help distinguish between signal peaks and noise. In our algorithm, a peak region in y is defined by the region where y' exceeds the region threshold and has at least one sign change. The peak region must be wider than the minimum peak region width. If two peak regions are separated by less than a prespecified maximum peak region separation time, then all the regions in between these two regions were also considered as peak regions. We describe our peak region detection algorithm in detail in the following steps:

1. Remove noise in the signal with WT denoising.

2. Calculate y', the first derivative of the signal y.

3. Calculate the zone threshold ZT for y', defined by equation (5.2),

$$ZT = ZTM \times Median(\hat{y}'), \tag{5.2}$$

   where ZTM is the zone threshold multiplier and $\hat{y}'$ is a three point moving average of y'. ZT is used to form peak region start and end coordinates. This threshold calculation method was used in a peak location detection algorithm [52]. We applied the same threshold calculation method for our peak region detection algorithm.

4. Find all the coordinates where y' crosses ZT and $-ZT$. Store and sort these coordinates into an index array.

5. Divide y' into $L-1$ zones using the index array, where L is the length of the index array. Each point in the index array forms a start or end coordinate of a zone, with the exceptions of the first and last points.

6. Compare the magnitude of each formed zone with the region threshold (RT), defined by equation (5.3),

$$RT = RTM \times Median(\hat{y}'), \tag{5.3}$$

where RTM is the region threshold multiplier. The zones with magnitude greater than RT are considered candidates for peak region zones. RT is used to differentiate between peaks and noise.

7. Compare the length of each peak region zone candidate to the minimum zone width. If the width of a candidate peak region is smaller than the minimum zone width, then the peak region candidate is removed from the peak region zone candidate list. This verification is used to remove zones with short duration that exceed RT. If a zone exceeds RT less than the minimum zone width, then that zone is considered to be noise and is removed from the candidate list.

8. For the remaining peak region zone candidates, check for peak region candidates separated by less than the maximum zone separation time. If any are found, change all the zones between the peak region candidates to peak region candidates as well. This condition is used to compensate for transitional zones formed between ZT and $-ZT$ and overlapping peaks.

9. Store the remaining peak region candidates into peak region start and finish arrays using equation (5.4) and equation (5.5),

$$start[i] = index[i] + 1, \tag{5.4}$$

and

$$finish[i] = index[i+1], \tag{5.5}$$

where i represents the peak region candidate number.

10. Check for peak regions with start and finish coordinates that are offset by one data point. Peak start and finish coordinates that are offset by one data point belong to the same peak region. Combine these peak regions into one peak region by joining the first peak region start coordinate and the last peak region end coordinate, and discard all the coordinates in between.

11. Compare the length of the peak region to the minimum region width. Discard the peak regions that do not satisfy this condition.

12. For a signal peak, the derivative of the peak signal will have a magnitude that change signs at least once [52]. Verify that each peak region has at least one sign change. If a peak region does not satisfy this condition, remove the peak region.

13. Compare each peak region's mean value to the mean of the entire signal. To verify that the peak regions have higher intensity than the overall signal, discard all peak regions with means smaller than the mean of the entire length of the signal. This step is used to remove noise.

14. If no peak regions are identified, treat the entire length of the signal as a peak region. Otherwise, the remaining peak regions are considered to be peak regions for a signal.

5.2.3 Baseline Variation Algorithm Parameters

We started our analysis by testing our peak region detection algorithm and the IPBF algorithm on a synthetic noisy multiple peak signal with added sinusoidal baseline and noise, as shown in Fig. 5.1. This signal was synthesized by the method described in Section 4.3.2 of this thesis. The starting baseline variation removal parameters for a signal are summarized in Table 5.1. Gan *et al.* advised that the polynomial order should be low and the error factor threshold ρ_{thr} should be set to 1×10^{-3} [60]. We set ρ_{thr} to 1×10^{-7} and set k_{max} to a large number to ensure convergence of the IPBF. The RTM was set to 3, the same value as another peak location detection algorithm [52]. The ZTM was set to 2 to form zones in the first derivative. RTM and ZTM are used to form regions in y' to determine the peak regions.

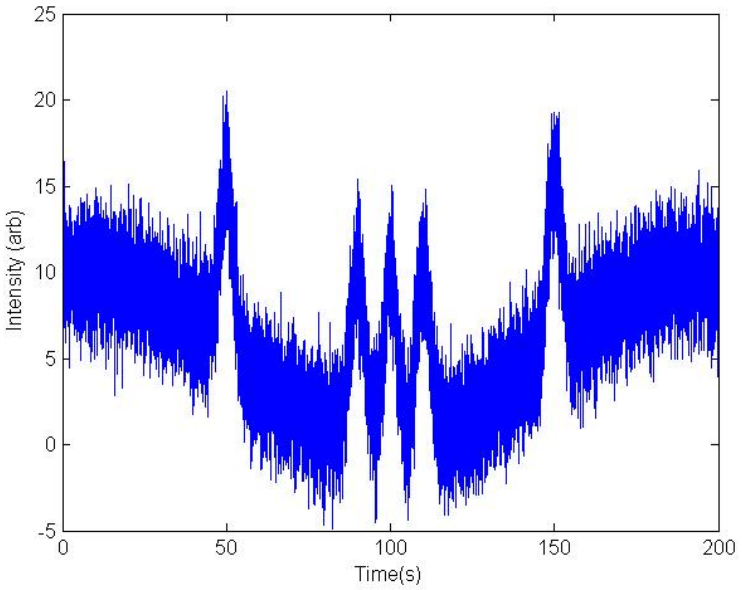


Figure 5.1: A synthetic multiple peak signal with baseline variation and noise (SNR = 3 V/V).

The first baseline removal parameter we tested was ρ_{thr}. ρ_{thr} was used to test the convergence between the current and the previous iteration of the baseline es-

Parameter	n	k_{max}	ρ_{thr}	RTM	ZTM
Value	10	200	1×10^{-7}	3	2

Table 5.1: IPBF, peak region detection starting and synthetic signal parameters.

ρ_{thr}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
RMSE	0.361	0.0704	0.0507	0.0493	0.0492	0.0491	0.0491

Table 5.2: Parametric sweep of error factor threshold ρ_{thr}.

timate. When ρ was smaller than ρ_{thr}, we considered that the IPBF had found an adequate baseline estimation. To explore the effect of ρ_{thr}, we swept various values of ρ_{thr} while keeping all other parameters in Table 5.1 constant. Table 5.2 shows the RMSE between the true baseline and the estimated baseline for various ρ_{thr} values. The RMSE quantifies how well the baseline variation estimate matches the true baseline variation. We found that the RMSE was initially high (0.361) for ρ_{thr} set at 1×10^{-2}. As we lowered ρ_{thr} toward 1×10^{-7}, the RMSE converged to 0.0491. As ρ_{thr} decreased, the RMSE between the true baseline and the baseline estimation also decreased, until a limit was reached.

A maximum iteration count k_{max} was used to prevent the IPBF from iterating beyond the point where iterating further did not lower ρ below ρ_{thr}. We swept k_{max} and measured the RMSE between the true and estimated baselines. As shown in Table 5.3, we found that the RMSE was high (0.307) when we set k_{max} to 10. As we increased k_{max} to 30, the RMSE decreased to 0.0838. For k_{max} larger than 50, the RMSE converged to 0.0789. Thus, increasing k_{max} lowered the RMSE, until a limit was reached.

k_{max}	10	20	30	40	50	60	70	80
RMSE	0.307	0.112	0.0838	0.0796	0.0790	0.0789	0.0788	0.0788

Table 5.3: Parametric sweep of maximum iteration count k_{max}.

n	5	10	15	20	25	30	35	40
RMSE	0.132	0.108	0.142	0.2046	0.371	0.404	0.366	0.353

Table 5.4: Parametric sweep of polynomial order degree n.

RTM	1	2	3	4	5	6	7	8	9
RMSE	0.23	0.11	0.092	0.092	0.093	0.099	0.098	0.40	0.40

Table 5.5: Parametric sweep of the region threshold multiplier RTM.

We performed a parametric sweep on the polynomial order degree (n) and measured the RMSE between the estimated and the true baseline variation. As shown in Table 5.4, we found that the RMSE was lowest for n is 10. For large n, the estimated baseline diverged near the end of the signal. If n was too small, the polynomial could not estimate the baseline variation accurately.

The results from sweeping the region threshold multiplier RTM are shown in Table 5.5. If the RTM was small (1 or 2), many false peak regions were detected due to noise. When the RTM was too high (greater than 7), our peak region detection algorithm found only peak regions with high amplitudes. For $RTMs$ between 3 and 7, the RMSEs were below 0.1. Thus a balance was achieved between high and low $RTMs$ to ensure the success of correct peak region detection.

We also swept the zone threshold multiplier ZTM and measured the RMSE between the true and the estimated baseline variations, as shown in Table 5.6. The effect of varying the ZTM was very similar to the effect of varying the RTM: if the ZTM was very low (0), the correct peak regions were not detected. If the ZTM was set too high (greater than 3.5), our peak region detection algorithm considered only a portion of the peak region as peak region.

ZTM	0	0.5	1	1.5	2.5	3.0	3.5	4
RMSE	0.244	0.107	0.0699	0.0719	0.0733	0.0756	0.313	0.788

Table 5.6: Parametric sweep of zone threshold multiplier ZTM.

Parameter	n	k_{max}	ρ_{thr}	RTM	ZTM
Value	10	60	1×10^{-7}	3	2

Table 5.7: Suitable parameters for IPBF and peak region detection algorithms

5.3 Baseline Variation Removal Test Results

5.3.1 Synthetic Signals

A suitable set of baseline variation removal algorithm parameters derived from Section 5.2.3 are shown in Table 5.7. With these parameters, we tested our baseline removal algorithm on a synthetic multiple peak signal, as shown in Fig. 5.1. As indicated by the vertical lines in its first derivative shown in Fig. 5.2, the peak regions were accurately detected. RT and ZT are shown by horizontal dotted lines. Despite the noise spikes that exceeded RT, our peak region detection algorithm did not detect the noise spikes as peak regions. With the correct peak regions detected, the IPBF removed the baseline variations as shown in Fig. 5.3. Fig. 5.4 shows the true synthetic baseline and the estimated baseline. Throughout the length of the signal, the baseline estimate matches the true baseline very closely despite the presence of noise and the overlapping peaks in the noisy signal.

5.3.2 Median Subtraction Comparison

We compared our baseline variation removal algorithm with median subtraction (MS), a method currently used in the TTK to remove baseline variations and offset. Using the signal shown in Fig. 5.1, we performed a parametric sweep of various median subtraction widths and measured the RMSE between true and estimated baseline. The results are shown in Table 5.8. We found that a 23-second median

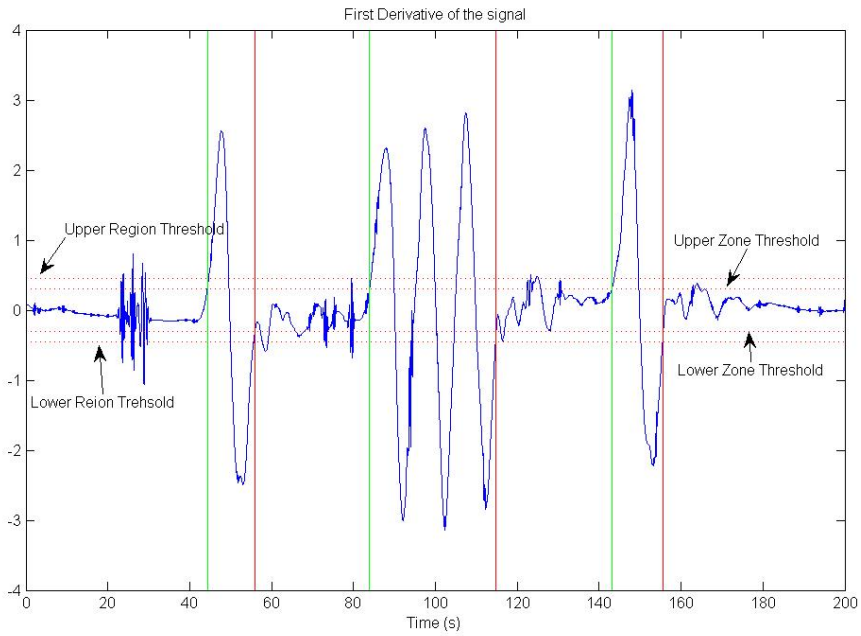


Figure 5.2: The first derivative of the multi-peak synthetic signal with baseline variation. The start and end coordinates of the peak region are indicated by green and red vertical lines, respectively.

subtraction filter yielded the lowest RMSE at 1.42, which was larger than the IPBF's RMSE. Fig. 5.5 shows the synthetic signal with its baseline variation removed by this median subtraction filter. With median subtraction, the post-processing signal dipped into the negative fluorescent range which is a non-physical signal. MS also caused the post-processing signal to dip in the leading edge of the leading peak and the trailing edge of the trailing peak. MS failed to estimate the true baseline because the peaks were overlapped [66]. This is because for signals with overlapping peaks, if the median width is too large, the median values are calculated from multiple peaks, which corrupts the peak shapes. But if the median width is too small, then MS cannot remove high baseline variations.

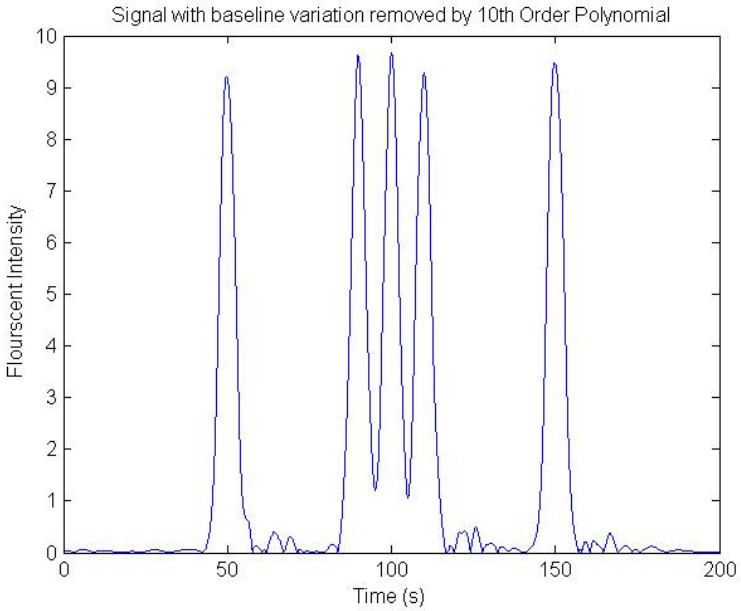


Figure 5.3: Synthetic signal with baseline variation removed by IPBF and peak region detection.

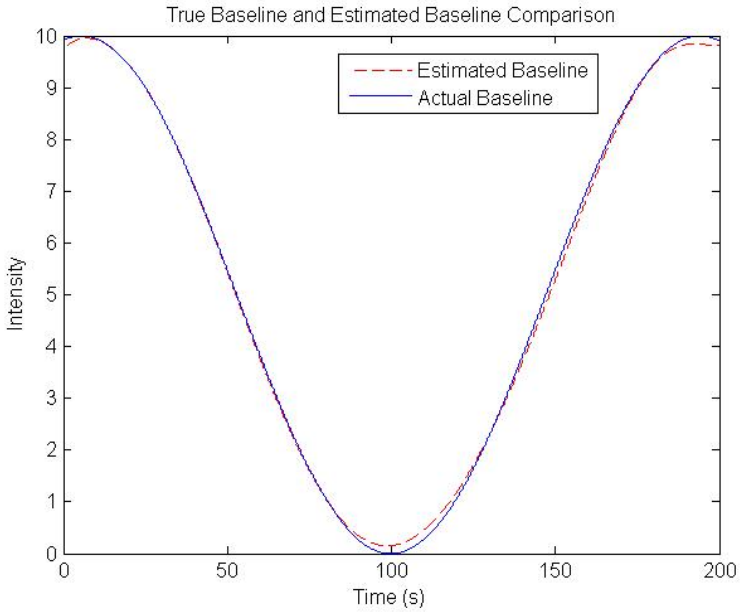


Figure 5.4: Comparison of true baseline and estimated baselines.

Seconds	1	5	10	15	20	23	27	30	35	40
RMSE	3.04	2.40	1.81	1.65	1.46	1.42	1.52	1.63	1.93	2.29

Table 5.8: Parametric sweep of median subtraction length.

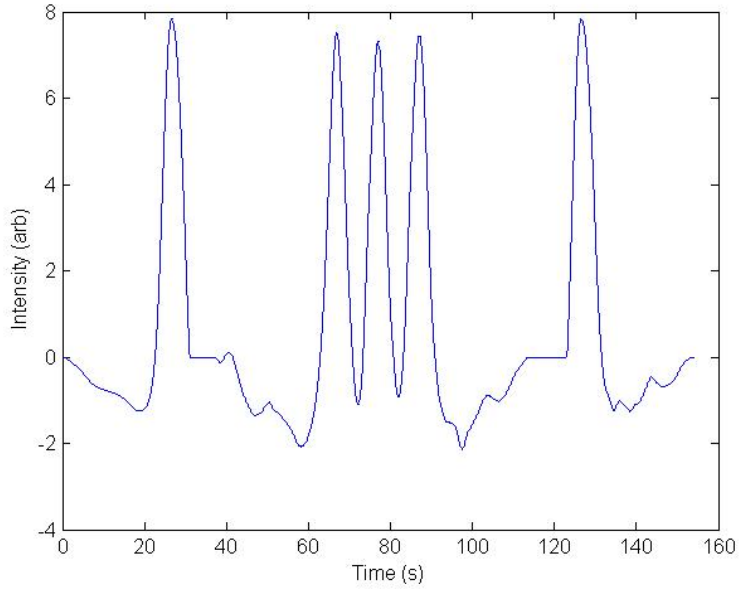


Figure 5.5: Synthetic signal with baseline variation removed by a 23-second median subtraction filter.

5.3.3 Baseline Variation and Noise Limits

We tested the limits of our baseline variation removal algorithm by calculating the RMSE between the true and estimated baselines for signals with various peak height to baseline variation ratio (PBR) and SNR. In our synthetic data, PBR was defined as the ratio between the maximum peak height and the amplitude of the sinusoidal baseline variation. As shown in Table 5.9, we found that the RMSE remained relatively constant and below 0.1 for PBRs from 20 down to 0.75. The slight variation in RMSE was due to noise variations. At a PBR of 0.5, the RMSE increased to 0.153. As the PBR decreased to 0.4, the RMSE increased drastically and the IPBF began to fail to estimate baseline variations. At PBRs of 0.3 and lower, the IPBF

PBR	20	10	1	0.75	0.5	0.4	0.3	0.2
RMSE	0.058	0.047	0.073	0.089	0.153	0.642	1.77	1.94

Table 5.9: RMSE between true and baseline variation removed signals with various levels of PBR.

SNR	50	10	2	1.5	1.4
RMSE	0.0753	0.0665	0.0746	0.0773	0.101
SNR	1.3	1.2	1.1	1	0.75
RMSE	0.407	0.546	0.945	1.46	2.04

Table 5.10: RMSE between the true and the estimated signal for various levels of SNR.

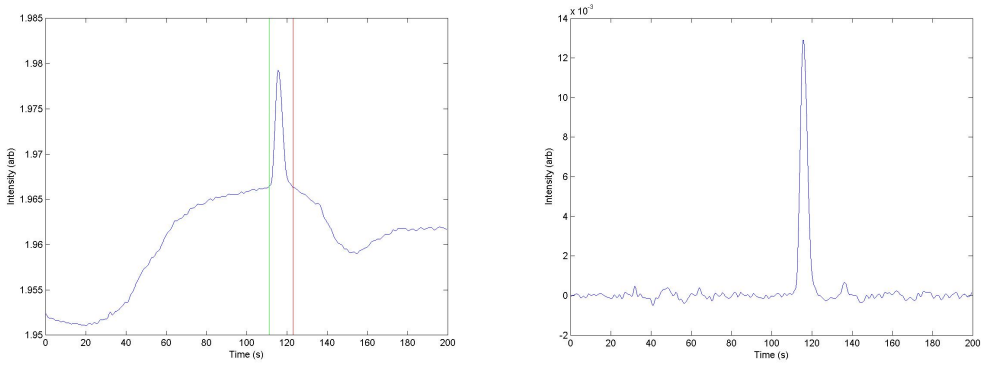
failed to match the baseline variation at all as indicated by RMSEs of 1.77 and 1.94 for PBRs of 0.3 and 0.2, respectively. The IPBF failed to remove the baseline variations at low values of PBR because our peak region detection algorithm could not determine a threshold value that could be used to distinguish between the baseline and the peak regions.

We also tested our baseline variation removal algorithm on signals with various SNRs. As shown in Table 5.10, the IPBF removed the baseline accurately with SNRs down to 1.5, as evidenced by the RMSE below 0.08. At a SNR of 1.4, the RMSE rose to 0.101 and the trend continued with RMSE values of 0.407 and 2.04 at SNRs of 1.3 and 0.75, respectively. At a SNR level of 0.5, our peak region detection algorithm failed to find any peak regions as wavelet denoising did not remove an adequate amount of noise and caused the peak region algorithm to fail to detect peak regions.

5.3.4 Experimental Signals

Our baseline removal algorithm was tested using three experimental CE signals collected with the TTK performed by Allison Bidulock: a single peak (Fig. 5.6), three

separated peaks (Fig. 5.7), and multiple overlapping peaks (Fig. 5.8). The data was collected using AML's standard CE protocol as described in Section 4.3.2 and references [29, 94]. In the sub-figures shown in Figs. 5.6 – 5.8, (a) shows the denoised signal and (b) shows the signal after the baseline variation has been removed. In the Cy-5 end-labelled DNA with a concentration of 0.749 ng/μL single peak sample shown in Fig. 5.6(a), despite the high baseline variation, the peak region is correctly identified and the IPBF accurately estimated and removed the baseline variation. In the 0.5μL thermo-cycler PCR product of BK virus sample with lengths of 26 and 299 bp as shown in Fig. 5.7(a), our baseline removal algorithm detected the proper peak regions and removed the baseline variations. Where the peaks were severely overlapped in the 1μL of DNA ladder (11 peaks, from 50 - 550 bp) sample, as shown in Fig. 5.8(a), our peak region detection identified the overlapping peaks as a single peak region algorithm and the IBPF removed the baseline variations.

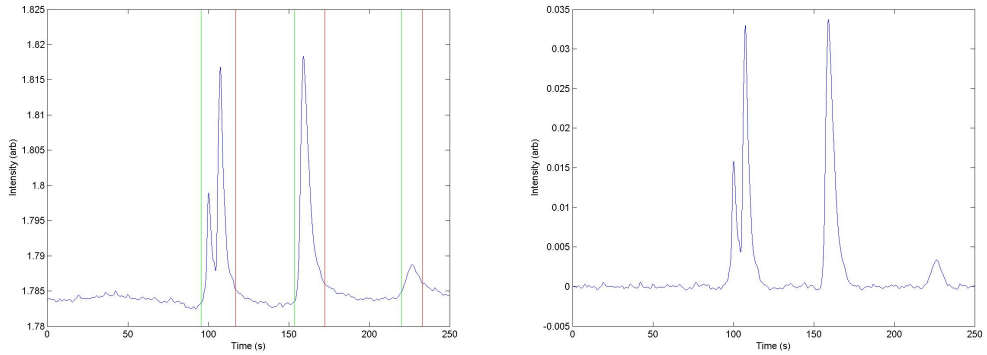


(a) Denoised signal (b) Baseline removed signal

Figure 5.6: The peak region algorithm correctly identified the peak region and the IPBF removed the baseline variations in a single peak experimental CE signal.

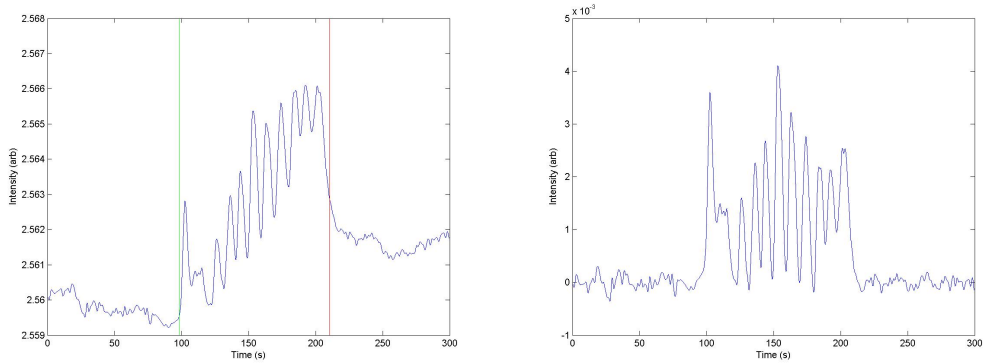
5.4 Conclusion

To address the baseline variation observed in the signals collected with the TTK, we implemented a baseline variation removal algorithm which consists of the IPBF



(a) Denoised Signal (b) Baseline Removed Signal

Figure 5.7: Our peak region detection algorithm correctly identified the peak region and the IPBF removed the baseline variations in an experimental CE signal with 3 separate peak regions.



(a) Denoised Signal (b) Baseline Removed Signal

Figure 5.8: Our peak region detection algorithm correctly identified the peak region and the IPBF removed the baseline variations for an experimental CE signal with overlapping peaks.

and our peak region detection algorithms. The combined algorithm removed baseline variations in synthetic signals with a SNR and a PBR as low as 1.4 V/V and 0.75 V/V, respectively. We also demonstrated that this baseline variation removal algorithm removed the baseline variations in CE signals with single, multiple, and overlapped peaks regions.

Chapter 6

Overlapping Peak Separation

6.1 Introduction

We improved the amplifier of the detection subsystem of the TTK in Chapter 3, and then removed the high-frequency noise with the wavelet transform and the low frequency baseline variations in Chapter 4 and Chapter 5. Now we will focus on separating overlapping peaks for capillary electrophoresis (CE) signals.

Under ideal conditions, the output from a laser-induced fluorescent (LIF) CE detection system is a train of impulses [80, 100]. However, unwanted convolution with an unknown point spread function (PSF) is inevitable as a result of the imperfections in the optical, fluidic and electronic systems, which causes peaks to widen and, in some cases, to overlap. Peak resolution is defined by the ratio of the distance between two peaks and the average peak width [37]. Overlapping peaks (or low resolution) hinder data analysis because it is difficult to identify signal peaks in low resolution signals. CE peak resolution can be increased by lengthening the microfluidic chip channel or by increasing separation times. However, these methods often result in bulkier systems or longer diagnoses.

Besides modifying the CE system or protocol to overlapping separate peaks, post signal-processing of CE signals can also separate overlapping peaks. The most common peak separation technique is curve fitting [13], but this can only separate overlapping peaks if noise is low and peak resolution is high. Initial estimates of

the number of peaks, peak shape, and peak width are required for the curve fitting method [70, 71]. Fourier self-deconvolution (FSD) with a Weiner smoothing filter has been shown to separate overlapping peaks in [72–74]; however, this method also requires a high SNR because artifacts are easily created [75]. For these reasons, curve fitting and FSD are not suitable for automatic deconvolution of CE signals. Olazabal *at el.* showed a method where signals were deconvoluted using wavelet transform. But in their results, although peaks are effectively separated, the deconvolved signal shifted and the peak shape are distorted from the original signal [13, 76].

Numeric deconvolution methods, such as Jansson's deconvolution have been reported to separate overlapping peaks in chromatographic signals [77–81]. Jansson's deconvolution can increase the resolution of chromatographic peaks with low SNRs. To our knowledge, no one has used Jansson's deconvolution to separate overlapping peaks in CE signals. The main benefit of Jansson's deconvolution is that it does not require knowledge of the signal's noise characteristics. The only parameters required are the impulse response of the system and the amplitude bound of the signal [82]. To increase the usability and automation of Jansson's deconvolution in a CE signal processing algorithm, we incorporated the use of normalization, peak detection, and deconvolution factor.

6.2 Peak Separation Methods

6.2.1 Peak Separation by Fourier Transform

In a typical electropherogram, the signal observed $y(t)$ can be modelled by equation (6.1),

$$y(t) = x(t) * h(t) + n(t) \tag{6.1}$$

where $h(t)$, $x(t)$, $n(t)$ represent the PSF, the true signal, and the noise in the sys-

tem, respectively. A method to retrieve $x(t)$ from $y(t)$ uses the forward and inverse Fourier transform (FT). Equation (6.2) represents the FT of equation (6.1):

$$\mathbf{Y} = \mathbf{XH} + \mathbf{N}, \tag{6.2}$$

where **Y**, **X**, **H**, **N** are the FT of $y(t)$, $x(t)$, $h(t)$, and $n(t)$, respectively. If the noise is small, then by rearranging equation (6.2) and taking the inverse FT, an estimate of the original signal $\hat{x(t)}$, can be retrieved by equation (6.3),

$$\hat{x(t)} = \mathbf{F}^{-1}\left(\frac{\mathbf{Y}}{\mathbf{H}}\right). \tag{6.3}$$

It is possible to use the FT method to obtain an estimate of $x(t)$, but since **H** is band limited and has values close to zero in some frequencies, dividing by **H** can cause the signal to diverge [79]. Variations of the FT by incorporating other methods, such as a Wiener filter, have been been reported to deconvolve signals [72–74]. Although effective, these method require require the prior knowledge and precise modelling of the signal and noise.

6.2.2 Jansson's Deconvolution

6.2.2.1 Theory

Jansson's deconvolution algorithm used for deconvolving signals, proposed by Crilly [79], is outlined in the following steps:

1. Set the initial estimate of x^k to y. In the first iteration, the iteration count k is set to 0.

2. Calculate the relaxation factor $r(x^k)$ defined by

$$r(x^k) = b(1 - \frac{2}{c}\left|x^k - \frac{c}{2}\right|), \tag{6.4}$$

where c is the maximum height of the signal and b is a noise dependent relaxation constant.

3. Estimate the original signal using:

$$x^{k+1} = x^k - r(x^k)(ah * x^k - y), \tag{6.5}$$

where x^k and x^{k+1} represent the current iteration and the next iteration of the estimated signal, respectively. The constant a is used to rescale $h(t)$ for proper convergence. $h(t)$ was obtained by curve fitting an ETG function to the signal of an experimental CE run of a 0.749 ng/μL of Cy-5 end-labelled DNA sample (or primer product), as described in Section 4.3.2 of this thesis. A procedure to obtain $h(t)$ is outlined in Appendix D.

4. Set $x^{k+1}(i) = \text{MAX}(x^{k+1}(i), 0)$ for all data points i.

5. Set $x^k = x^{k+1}$.

6. Return to step 2 if the iteration level k is less than the specified iteration level.

The idea behind Jansson's deconvolution is that if we substitute k for $k-1$ in equation (6.5) and take the FT of it, it becomes:

$$\mathbf{X}^k = \mathbf{X}^{k-1} - \mathbf{R}(\mathbf{X}^{k-1}) * (a\mathbf{H}\mathbf{X}^{k-1} - \mathbf{Y}). \tag{6.6}$$

The convolution performed in equation (6.6) extends the bandwidth of $\mathbf{X}^k$ because $\mathbf{R}$ is a function of $\mathbf{X}$, hence recovering the frequency component lost from equation (6.1). Furthermore, according to Crilly, equation 6.5 converges to the inverse filter estimate as shown in equation 6.3, if k is large and $\mathbf{H}(f) > 0$ [77].

6.2.2.2 Jansson's Deconvolution for CE Signals

The parameters for Jansson's deconvolution proposed by Crilly's papers are intended for chromatography signals [77–79]. Chromatography signals are generally wider and more overlapped than CE signals; therefore, Jansson's deconvolution

parameters must be modified. If the same parameters were used, Jansson's deconvolution will be very sensitive to noise and artifacts will be created easily.

The relaxation constant b and the maximum iteration count k_{max} are the parameters which are used to adjust the noise tolerance in Jansson's deconvolution. To determine a relaxation constant b suitable for CE signals, we performed a parametric sweep of b on a 11-peak DNA ladder (50 - 550 bp) CE signal collected with the TTK as shown in Fig. 6.1. This signal was obtained the using the standard CE protocol as described in Section 4.3.2. This signal was processed by wavelet denoising (in Chapter 4) and IPBF (in Chapter 5). Although we were able to remove the high-frequency noise and the low-frequency baseline variations in this signal, the post-processing peaks are still overlapped.

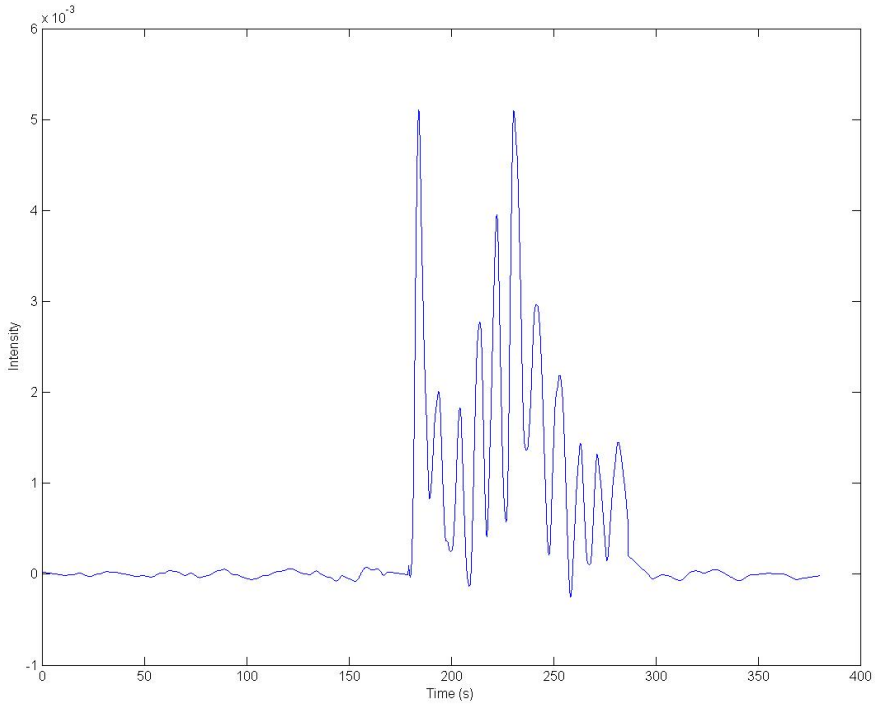


Figure 6.1: Experimental CE run with 11 overlapping peaks (DNA Ladder).

The deconvolved signals of Fig. 6.1 using Jansson's deconvolution with b set

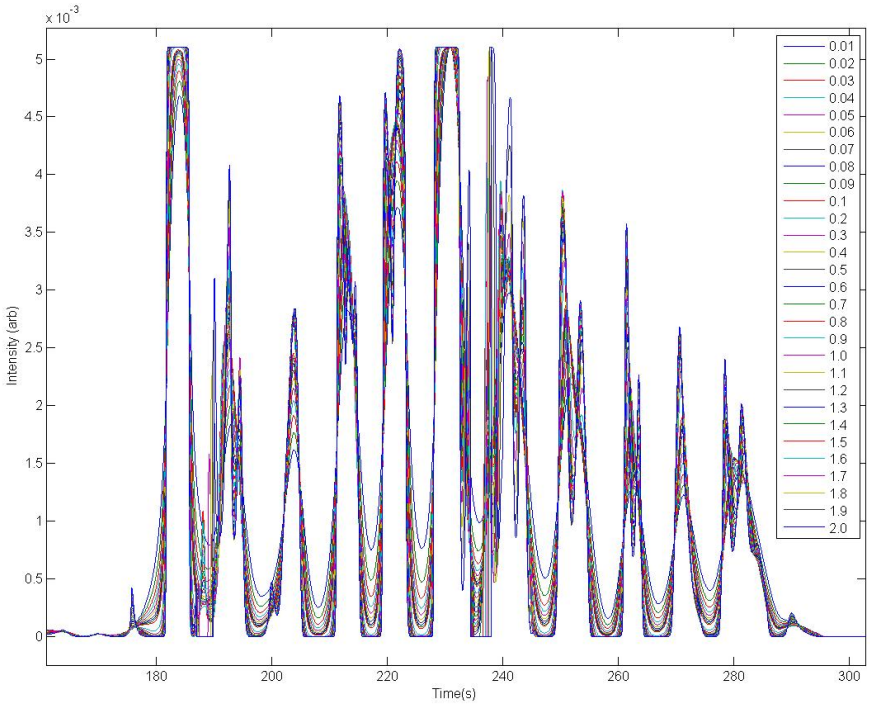


Figure 6.2: Parametric sweep of b from 0.01 to 2.

from 0.01 to 2 are shown in Fig. 6.2. We observed that as the relaxation constant b increased, Jansson's deconvolution separated the peaks more. If there was noise in the signal, and b was greater than 0.3, Jansson's deconvolution over deconvolved and created artifacts. But if b was smaller than 0.06, insufficient peak separation was achieved. It was crucial to find a value of b that minimized the creation of artifacts and maximized peak separation. From the zoomed-in deconvolved signal shown in Fig. 6.3, we observed that a b value of 0.01 provided the most peak separation without introducing artifacts in the signal.

We also investigated the effect of the number of iterations performed in Jansson's deconvolution. Fig. 6.4 shows the deconvolved signal at various iteration levels. We observed that as the number of iterations increased, the more the overlapping peaks were separated; however, artifacts were created at higher iteration

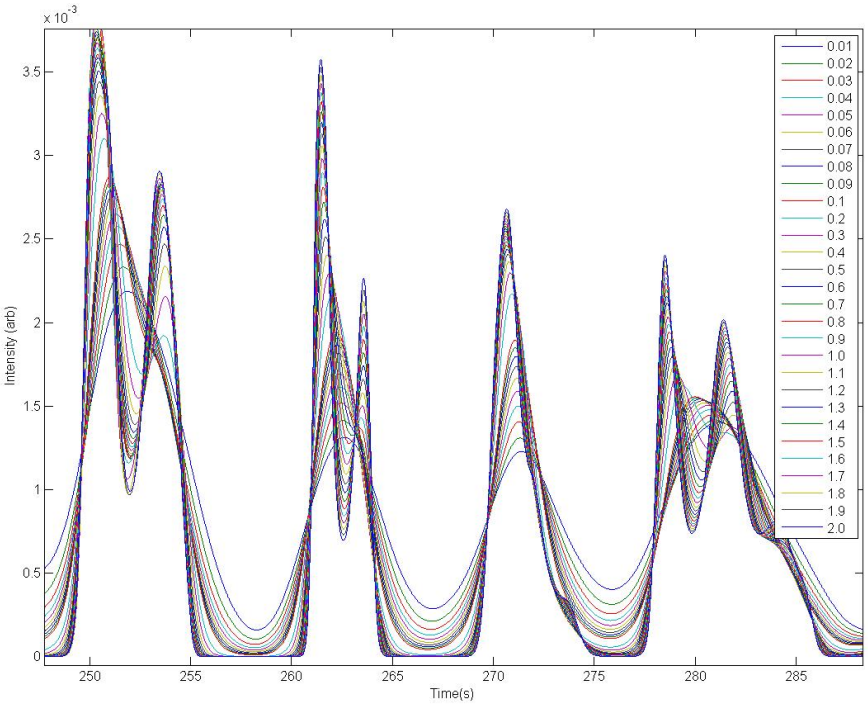


Figure 6.3: Zoomed-in peaks shown in Fig. 6.2

levels. This may not be an issue for high-precision and low-noise chromatography signals. But for a photodiode-based CE device, deconvolving with too many iterations creates artifacts. Since the main goal was to isolate overlapping peaks, only a few iterations were required to separate peaks.

To evaluate the effect of the number of iterations in Jansson's deconvolution, we plotted the deconvolution factor ρ against the number of iterations. The deconvolution factor is a measure of the difference between the current and the previous iteration of the signal and is defined by equation (6.7),

$$\rho = \frac{||x_k - x_{k-1}||}{x_{k-1}}. \tag{6.7}$$

As shown in Fig. 6.5, we found that ρ dropped significantly during the first few iterations and flattened out as the number of iterations increased.

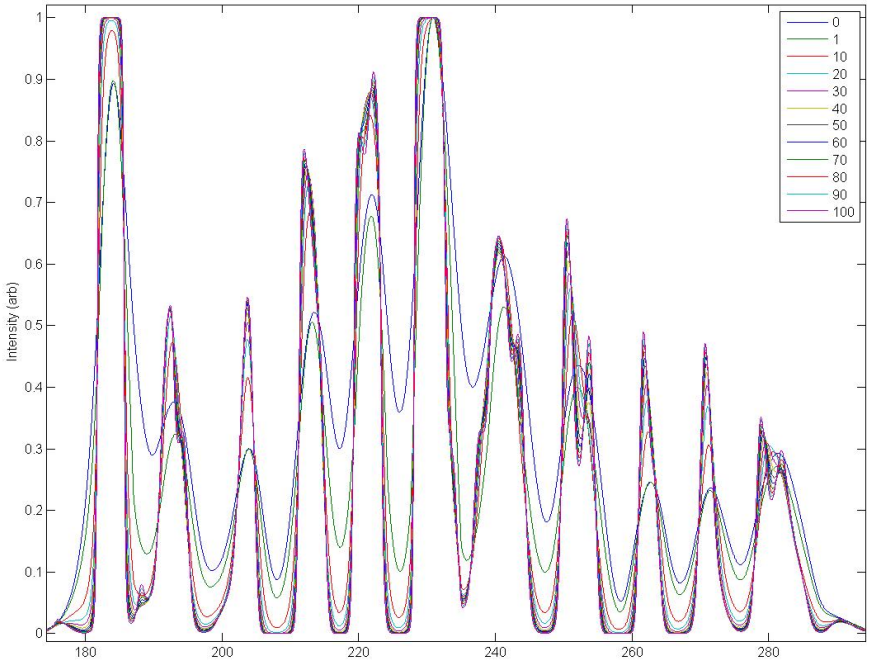


Figure 6.4: Deconvolved signal at various iteration level of Jansson's deconvolution.

6.2.2.3 Jansson's Deconvolution Modifications

Three modifications were made to Jansson's deconvolution to automate its use for CE signals. The first modification to Jansson's deconvolution was normalization of the signals and the PSF. Although a very simple process, normalization standardized all signals and the PSF to a height of 1. Normalization set all signals to the same range of values and allowed the signal bound parameters c in equation (6.4) and a in equation (6.5) to be set to constants. These parameters were obtained by trial and error testing on synthetic and experimental signals. We found that setting a to 1 and c to 0.025 provided the most peak separation without creating artifacts.

To minimize the creation of artifacts, we incorporated peak detection in each iteration of the Jansson's deconvolution. The number of detected peaks in the current

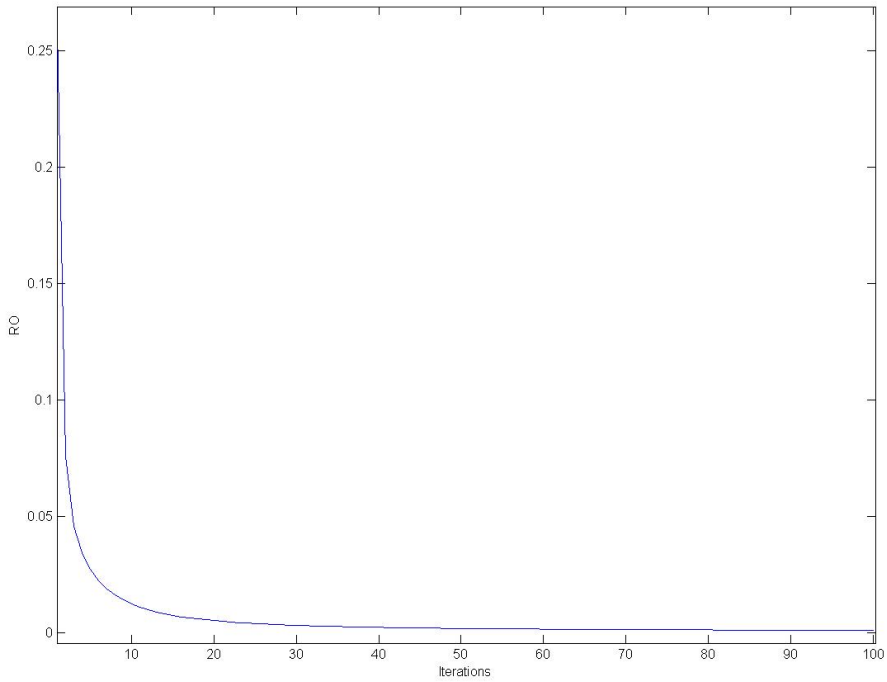


Figure 6.5: Deconvolution factor ρ versus number of iterations

iteration were compared to the number of peaks detected in the previous iteration. If they are not equal, then artifacts are created and the deconvolved signal from previous iteration is reverted and used as the final deconvolved output. To ensure adequate peak separation, a minimum of five iterations should be performed. An iteration maximum count of 100 is used to stop iterating if the number of peaks does not change through all of the iterations.

Another modification to Jansson's deconvolution was the use of deconvolution factor (ρ). ρ is defined by equation (6.7) and it calculates the similarity of the estimated signal between the current and the previous iterations. The deconvolution factor threshold (ρ_{thr}) is used to stop the iteration process when further iterating does not improve peak separation significantly. Using the deconvolution factor plot, as illustrated in Fig. 6.5, we set ρ_{thr} to 0.01. The implementation of a deconvolution

factor threshold reduces the creation of artifacts.

The use of peak detection and deconvolution factor between the iterations allowed Jansson's deconvolution to stop iterating once artifacts are detected or peaks are adequately separated in the deconvolved signal. Combined with the use of normalization, the parameters required for Jansson's deconvolution were set to constants. These modifications allowed Jansson's deconvolution to separate overlapping peaks in a CE signal processing algorithm.

6.3 Results

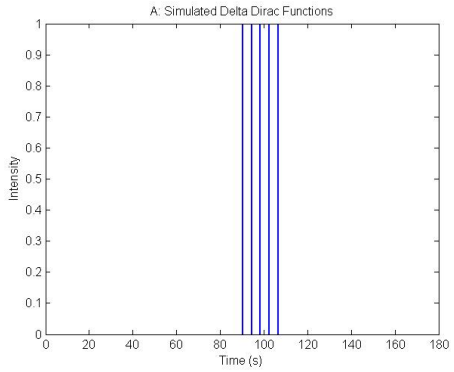
6.3.1 Synthetic Signals

6.3.1.1 Convolved and Deconvolved Synthetic Signal

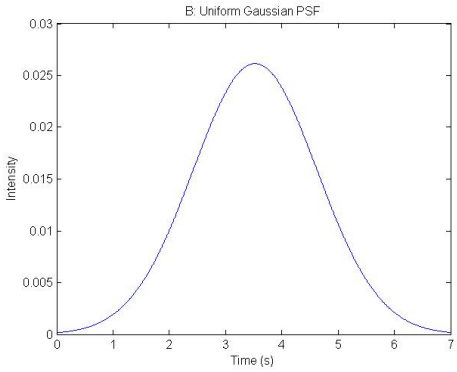
To verify that Jansson's deconvolution could be used to retrieve the true signal from a convolved signal, we synthesized a signal with five narrow pulses separated by 3.75 seconds and with widths of 0.3 seconds, as shown in Fig. 6.6(a). This signal was convolved with a uniform Gaussian PSF shown in Fig. 6.6(b) and the convolved signal is shown in Fig. 6.6(c). The deconvolved signal using Jansson's deconvolution is shown in Fig. 6.6(d). We demonstrated that Jansson's deconvolution could retrieve the original signal from a noiseless-convolved signal.

6.3.1.2 Resolution Limit

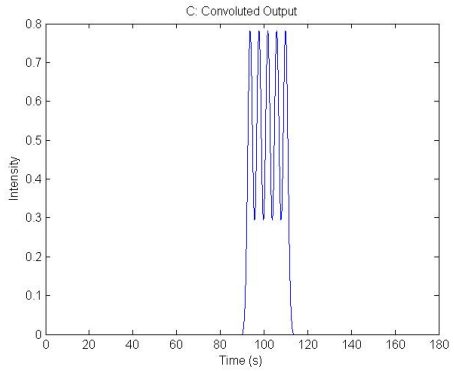
To test the resolution limit of Jansson's deconvolution, we synthesized and deconvolved signals with various resolutions. We found that the lowest signal resolution that Jansson's deconvolution could separate without distorting the peak shapes was 0.7. Fig. 6.7(a) and Fig. 6.7(b) show the original and deconvolved signals, respectively.



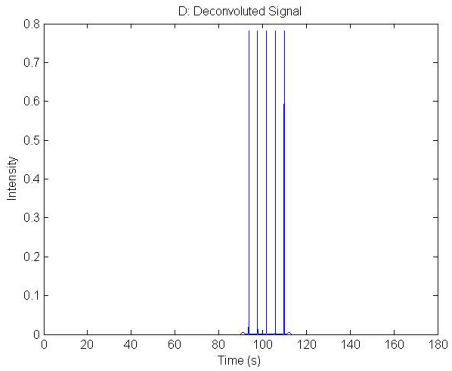
(a) Delta Dirac pulses



(b) Uniform Gaussian PSF model



(c) Convolved signal



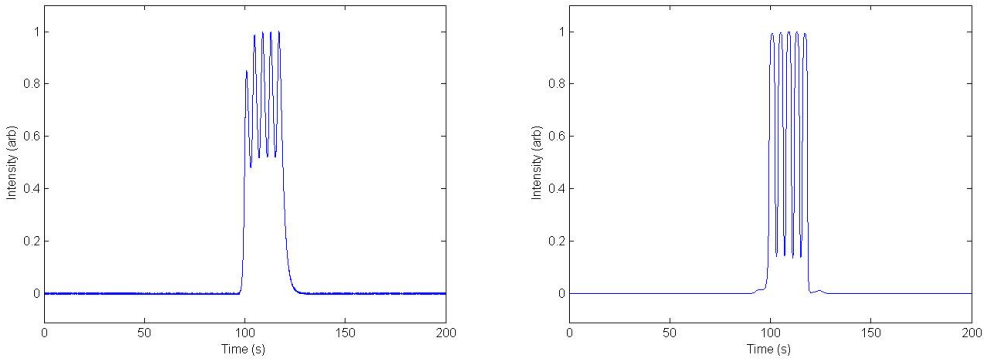
(d) Deconvolved signal output

Figure 6.6: Testing of convolution and Jansson's deconvolution using a synthetic signal.

6.3.2 Experimental Signals

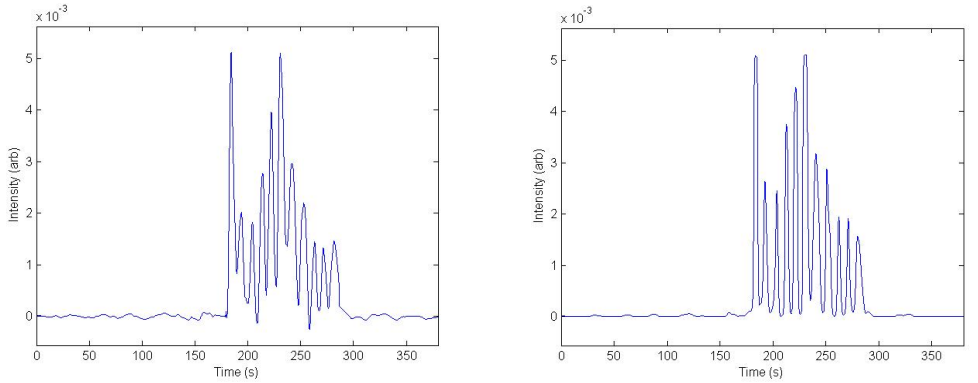
6.3.2.1 Low Resolution Signal

The CE signal of a 1μL of DNA ladder sample was used to test and verify that Jansson's deconvolution can be used to separate overlapping peaks. As shown in Fig. 6.8(a), this DNA ladder sample has 11 peaks: a primer peak (50 bp) and 10 product peaks (100 - 550 bp). It is the same signal as the signal shown in Fig. 6.1. We showed it again here to compare the signal with and without Jansson's deconvolution. Fig. 6.8(b) shows the deconvolved signal by our Jansson's deconvolution.



(a) Original signal (b) Deconvolved signal

Figure 6.7: Limit of Jansson's deconvolution signals with resolution of 0.7.



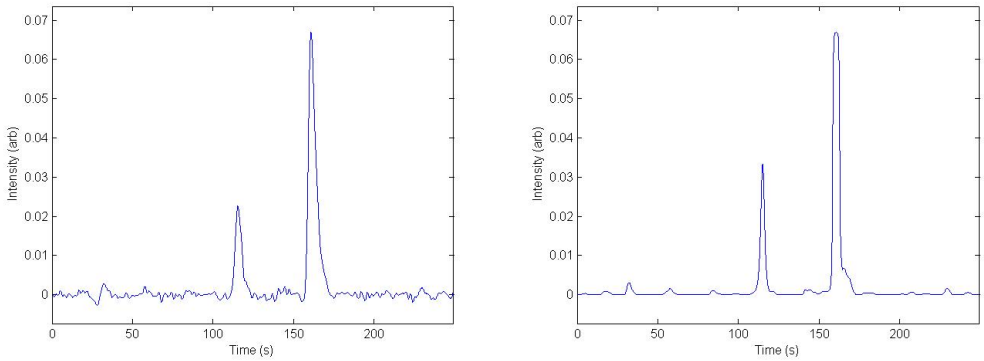
(a) Denoised Signal (b) Deconvolved signal

Figure 6.8: Jansson's deconvolution of a DNA ladder, a low resolution signal.

6.3.2.2 High Resolution Signals

Since Jansson's deconvolution will be applied to all signals in our signal processing algorithm, the effect of Jansson's deconvolution on high resolution signals was tested. Fig. 6.9(a) shows a two non-overlapping peak (high resolution) CE signal. This data was collected by the TTK using 0.5 μL of a thermo-cycled PCR product of BK virus. The primer has a length of 26 bp and the product has a length of 299 bp. The deconvolved signal is shown in Fig. 6.9(b) and showed that our modified

Jansson's deconvolution can also be used for CE signals with high resolution.



(a) Denoised Signal (b) Deconvolved Signal

Figure 6.9: Jansson's deconvolution of a high resolution signal.

6.3.3 Limitations and Assumptions

A limitation to Jansson's deconvolution is that it requires the accurate modelling of $h(t)$. As described, we obtained $h(t)$ by fitting an ETG function to the CE signal of a Cy-5 end-labelled DNA (primer) sample with a concentration of 0.749 ng/μL. For the TTK, we found that the $h(t)$ used could separate peaks for both signals with high and low resolution. If it was incorrectly modelled, artifacts can be easily created by Jansson's deconvolution.

It is very important to note that the TTK's $h(t)$ can vary from experiment to experiment. This is because the optical detection subsystem requires very precise alignment. Slight focus deviation from the optimal detection location can result in a signal with a much lower SNR for the same experiment. There are also variabilities in the manufacturing process of the microfluidic chips. For example, defects in the microfluidic chips can limit the mobility of DNA molecules during CE injection or separation. This changes the leading and trailing edges of a DNA peak and thus, changing $h(t)$. In this thesis, we assumed that all of the data collected was performed with no deviation from the CE protocol. This means that if we were to

find the $h(t)$ for every data set, our signal processing algorithm will be even more reliable and separate signals with lower resolution.

6.4 Conclusion

In this chapter we described the third and final step of our signal processing algorithm for CE signals. We modified Jansson's deconvolution algorithm and its parameters to enable overlapping peak separation for CE signals. By adding peak detection, deconvolution factor and normalization to Jansson's deconvolution, we minimized artifacts and maximized peak separation. We showed that Jansson's deconvolution retrieved the original signal and separated peaks with resolution as low as 0.7 for synthetic signals. We also demonstrated that Jansson's deconvolution can separate peaks in experimental signals with both low and high resolutions.

Chapter 7

Complete Capillary Electrophoresis Signal Processing Algorithm

7.1 Introduction

So far we have separately presented various algorithmic components for CE signal processing. In this chapter we will combine the previous algorithms together and investigate the resulting effectiveness for both synthetic and experimental CE signals. We identified high-frequency noise, baseline variations, and low resolution as three of the challenges in the signals recorded with the TTK.

Few previous researchers have addressed all of the issues (low SNR, baseline variations, low resolution) in non-confocal CE genetic analysis instruments. Most methods tackle only some of the problems. Shackman *et al.* developed a fast algorithm for processing large amounts of CE data [97]. The algorithm developed in this work possessed very impressive speed and throughput, but the technique used a straight line to estimate baseline variations and it did not remove baselines with high variation. Kaigala and Behnam *et al.* used Fourier filtering and a medium subtraction to process their CE signals, but in cases of overlapping peaks, their signal processing yielded signals with peaks dipping below the baseline (or negative peaks), which are non-physical signals [29].

We present a signal processing algorithm for CE signals with various levels of

Description	Variable	Value
Wavelet type	*wtname*	Sym8
Wavelet decomposition level	*level*	8
Wavelet threshold selection rule	*TPTR*	heuristic SURE
Wavelet threshold method	*SORH*	Soft

Table 7.1: Wavelet denoising parameters.

SNR, baseline variations, and resolutions. We will show that with proper signal processing, we can lower the TTK's LOD to that of a commercial system that utilizes expensive confocal optics.

7.2 Complete Capillary Electrophoresis Signal Processing Algorithm

High-frequency noise in raw CE signal is removed by a Symlet 8 (sym8) wavelet with level 8 decomposition. The denoised signal is then passed into the IPBF and peak region detection algorithms, where baseline variations are removed. A modified Jansson's deconvolution is used to improve peak resolution by separating overlapping peaks. Tables 7.1, 7.2, 7.3, and 7.4 summarise the parameters used for wavelet denoising, IPBF, peak region detection, and Jansson's deconvolution, respectively. We derived these parameters in Chapters 4, 5, and 6 of this thesis.

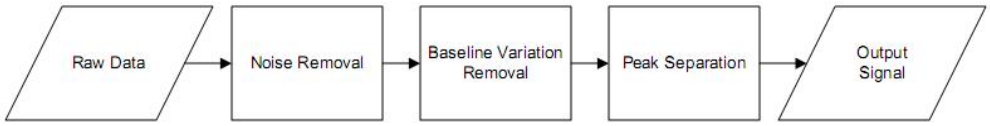


Figure 7.1: Overview of our complete CE signal processing algorithm consisting of noise removal, baseline removal, and peak separation.

Description	Variable	Value
Polynomial degree	n	20
Maximum iteration level	k_{max}	300
Iteration error threshold	ρ_{thr}	1.00×10^{-6}

Table 7.2: Iterative polynomial baseline fit parameters.

Description	Variable	Value
Zero threshold multiplier	ZTM	2
Region threshold multiplier	RTM	3
Maximum peak separation time (s)	$t_{PeakSepMax}$	5
Minimum span time (s)	$t_{SpanTimeMin}$	1.5

Table 7.3: Peak region detection algorithm parameters.

7.3 Test Results and Discussions

This section discusses the test results of our signal processing algorithm on experimental CE data samples collected using the TTK by M.Sc student Allison Bidulock. The CE protocol used is described in Section 4.3.2 and references [29] and [94]. We showed four examples of pre-processing and post-processing signals in Fig. 4.21. The dataset tested consists of CE runs with single and multiple peaks with low and high SNR and, near constant and varying baselines. We show a subset of the processed signals in this section. In each of the figures with sub-figures shown in this section, sub-figure (a) shows the raw signal; sub-figure (b) shows the signal with the

Description	Variable	Value
Relaxation factor	b	0.01
PSF scaling function	a	0.025
Signal bound	c	1
Maximum iteration count	k_{max}	100
Peak detection parameter	δ	0.1
Deconvolution factor threshold	ρ_{thrs}	0.02

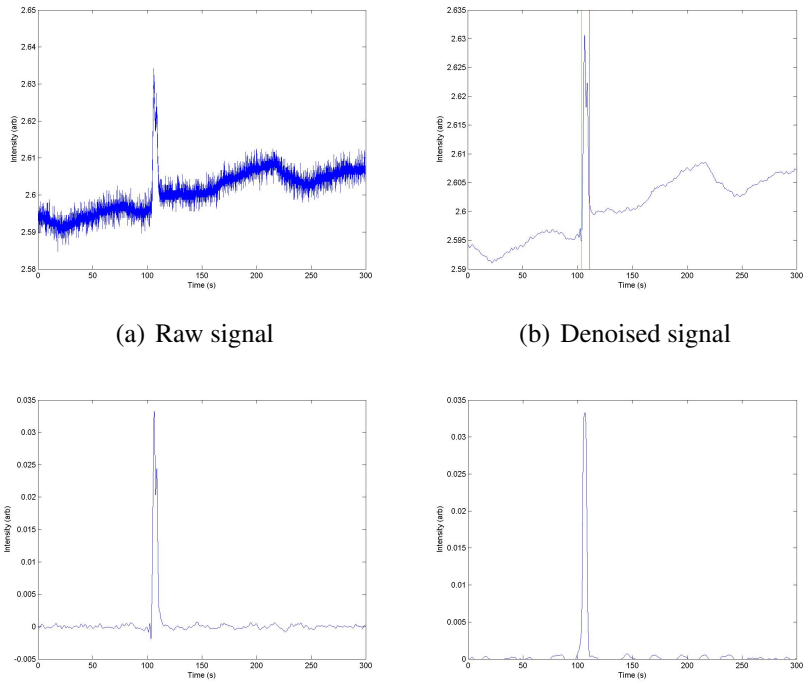
Table 7.4: Jansson's deconvolution parameters.

noise removed by wavelet denoising and the detected peak region(s); sub-figure (c) shows the signal after the baseline variations have been removed with IPBF; sub-figure (d) shows the signal with the peaks separated by Jansson's deconvolution.

7.3.1 Single Peak Signals

With our signal processing algorithm, we processed CE signals of single peak Cy-5 end-labelled (or primer) with DNA concentrations of 0.749 ng/μL and 0.498 ng/μL. Figs. 7.2, 7.3, and 7.4 show the signal processing stages for single peak signals with small baseline variations and SNR of 22.5, 7.7, and 4.6 V/V, respectively. Our signal processing algorithm removed noise and baseline variations and increased the post processing SNR to 342, 67, and 98 V/V, respectively. We measured our SNR by calculating the ratio between the height of the peak and the standard deviation of the baseline before the peak arrivals (typically the first 50 seconds of the signal).

We also tested our signal processing algorithm on a single peak signal with high baseline variation as shown in Fig. 7.5(a). This single peak signal has a baseline variation 15% higher than the signal's peak. Our peak region detection algorithm detected the peak region correctly as shown by the vertical lines in Fig. 7.5(b). The peak region detection algorithm is based on the rate of change of the signal (or y') rather than the magnitude of the signal. Since the rate of change of the baseline variation did not exceed the rate of change of the peak region threshold, our algorithm identified the peak region correctly and removed the baseline variation, as shown in Fig. 7.5(c). These processed samples showed that our signal processing algorithm could successfully remove the baseline variation (caused by laser intensity fluctuations) and the baseline offset (caused by autofluorescence) in single peak signals with various levels of SNR and baseline variations.



(a) Raw signal (b) Denoised signal

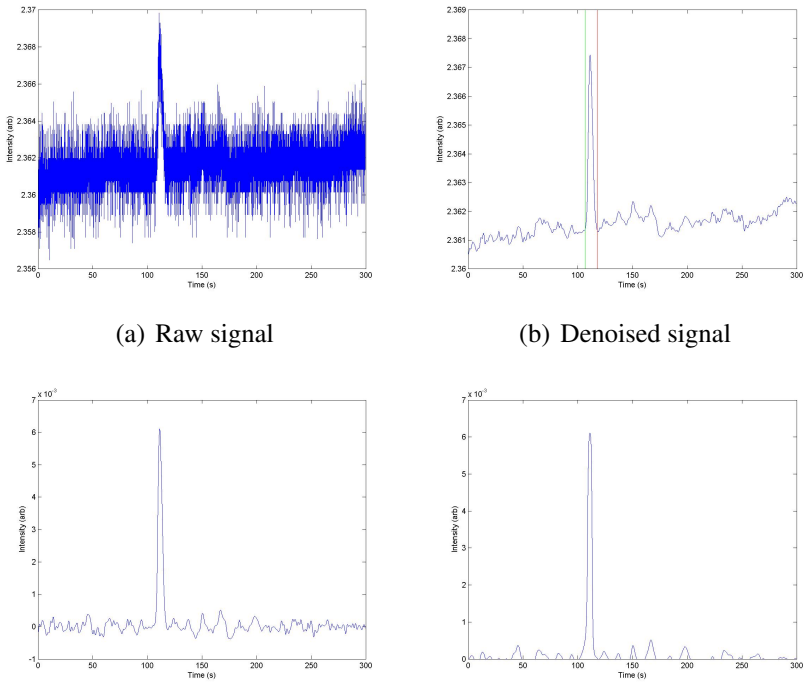
(c) Baseline variation removed signal (d) Deconvolved signal

Figure 7.2: Electropherogram of a 1 μL of Cy-5 end-labelled primer with concentration of 0.749 ng/μL. The unprocessed signal for this CE electropherogram had a medium SNR (22.5 V/V) and a varying baseline. The peak region was successfully detected, the baseline variation was removed, and the signal SNR was improved to 342 V/V post processing.

7.3.2 Multiple Peak Signals

We also performed signal processing on CE runs with multiple peaks. Figs. 7.6, 7.7, 7.8 and 7.9 show four different three-peak BK virus (BKV) CE runs with various level of SNR, baseline variations, and primary to secondary peak height ratio. The samples used in Figs. 7.6 and 7.7 were 0.5μL of thermo-cycled PCR product mixed with 3.5μL of 0.01xTTE and the samples used in Figs. 7.6 and 7.7 were 4μL of PCR product from the TTK.

In all of these runs, the first peak is the primer peak (25 or 26 bp), the second peak is the product peak (299 bp), and the third peak is an unspecific peak [29]. In



(a) Raw signal

(b) Denoised signal

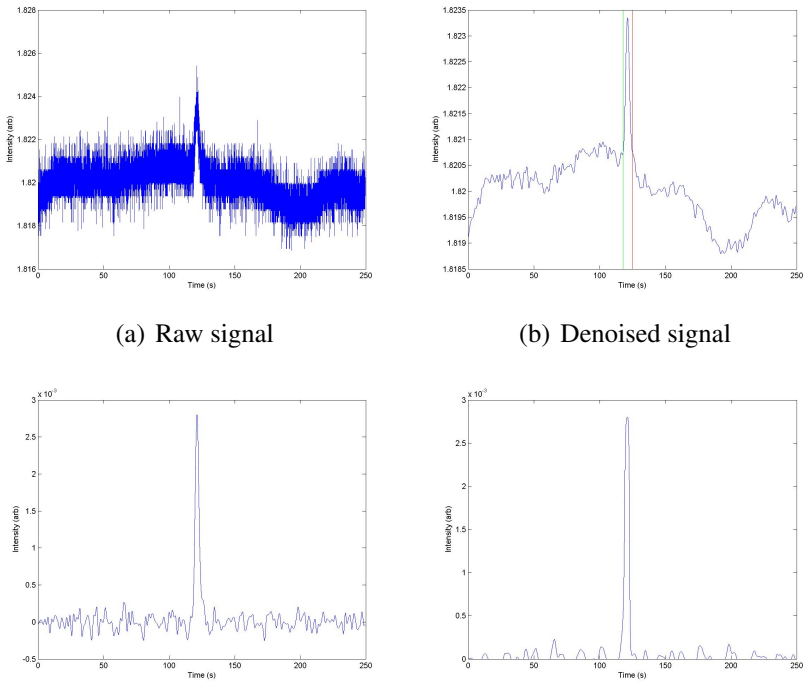
(c) Baseline variation removed signal

(d) Deconvolved signal

Figure 7.3: Electropherogram of a 1 μL of Cy-5 end-labelled primer with concentration of 0.749 ng/μL. The raw signal in this CE electropherogram had a low SNR of 7.7 V/V and near constant baseline. The signal processing was able to remove the noise in the signal and improve the signal SNR to 67 V/V after baseline removal.

the electropherograms with three peaks shown in Fig. 7.6 and Fig. 7.7, the peak height for all peaks is similar in magnitude. The difference between the two signals is that the signal in Fig. 7.6(a) has a higher SNR than the signal in Fig 7.7(a). In both electropherograms, our signal processing algorithm removed the noise and correctly identified the peak regions and no artifacts were created. Fig. 7.8 and Fig. 7.9 show three-peak signals where the height of the primary peaks are significantly higher than the heights of the secondary peaks. Despite the peak height differences, our signal processing algorithm removed the noise, detected the peak regions and removed the baseline variations in these signals.

We continued to test our signal processing algorithm on a 1μL of DNA ladder

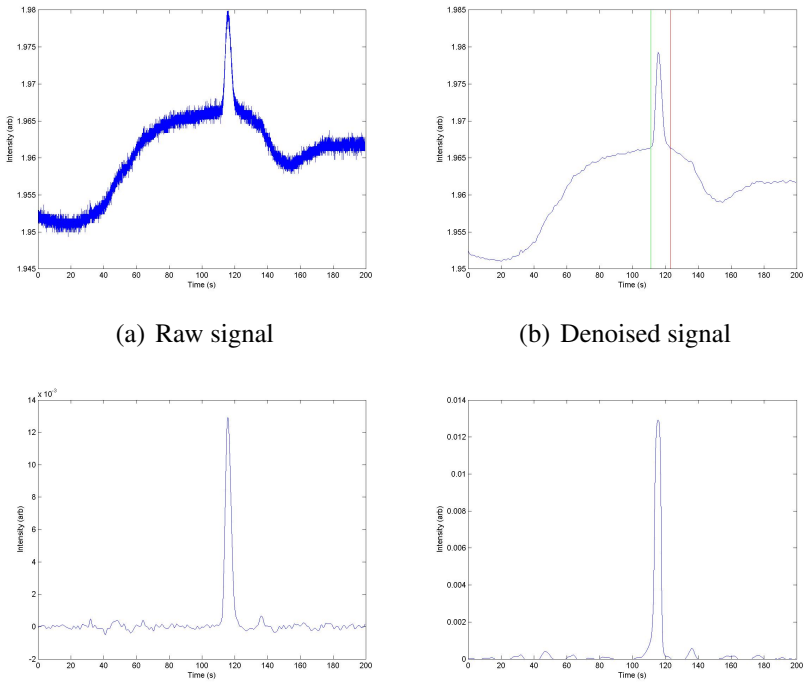


(a) Raw signal (b) Denoised signal

(c) Baseline variation removed signal (d) Deconvolved signal

Figure 7.4: Electropherogram of a 1 μL of Cy-5 end-labelled primer with concentration of 0.498 ng/μL. The signal processing stages for a CE signal with a low SNR (4.6 V/V) and a varying baseline. Our signal processing algorithm improved the SNR to 97.6 V/V and removed the baseline variations.

sample mixed with 3μL of 0.01xTTE, as shown in Fig. 7.10(a). This sample has 11 peaks (50 - 500 bp), where the first peak is the primer peak and the all other peaks are product peaks (separated by 50 bp). The denoised signal and the identified peak region are shown in Fig. 7.10(b). Because the peaks overlapped, our signal processing algorithm grouped the peaks as one peak region. The IPBF algorithm removed the baseline variations in the signal as shown in Fig. 7.10(c), and Jansson's deconvolution separated the overlapping peaks as shown in Fig. 7.10(d). The post-processing signal clearly showed 11 peaks. Our signal processing algorithm removed the noise and baseline variations, and increased the resolution of a DNA ladder sample. The same signal was processed by the current TTK software



(a) Raw signal (b) Denoised signal

(c) Baseline variation removed signal (d) Deconvolved signal

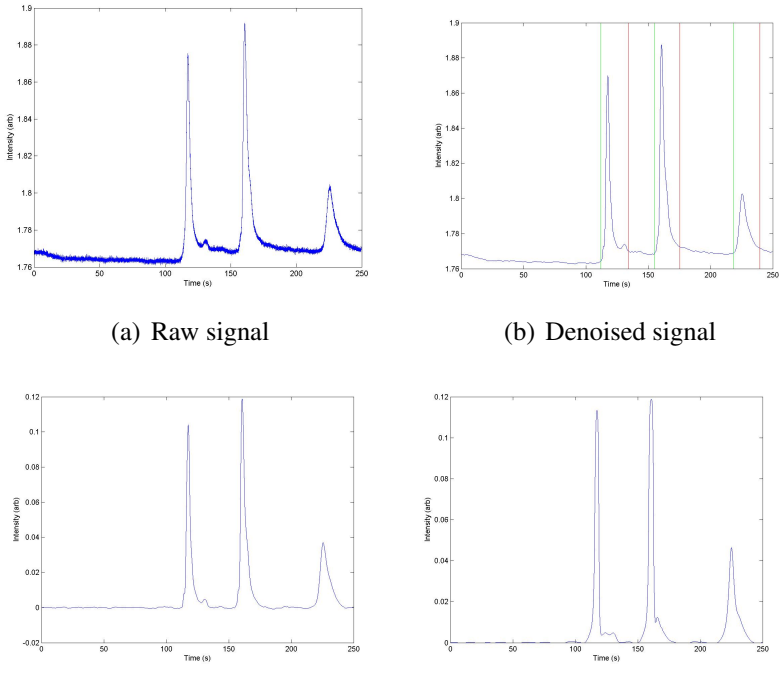
Figure 7.5: Electropherogram of a 1 μL of Cy-5 end-labelled primer with concentration of 0.749 ng/μL. Signal processing of a single peak signal with a medium SNR and a baseline variation magnitude greater than that of the signal peak.

as shown in Fig. 7.11 [101]. Median subtraction used in the TTK signal processing caused the peaks to dip below the baseline (negative peaks), which is a non-physical signal.

7.3.3 Reliability Study

A total of 117 experimental CE runs performed by M.Sc student Allison Bidulock with the TTK were analysed by our signal processing algorithm. Forty-four of the CE runs were single-peak DNA primer samples and 73 of the CE runs were multiple-peak BKV or DNA ladder samples.

Of the 44 runs with single-peak samples, our signal processing algorithm removed the noise in the signal, correctly identified the peak regions, and success-



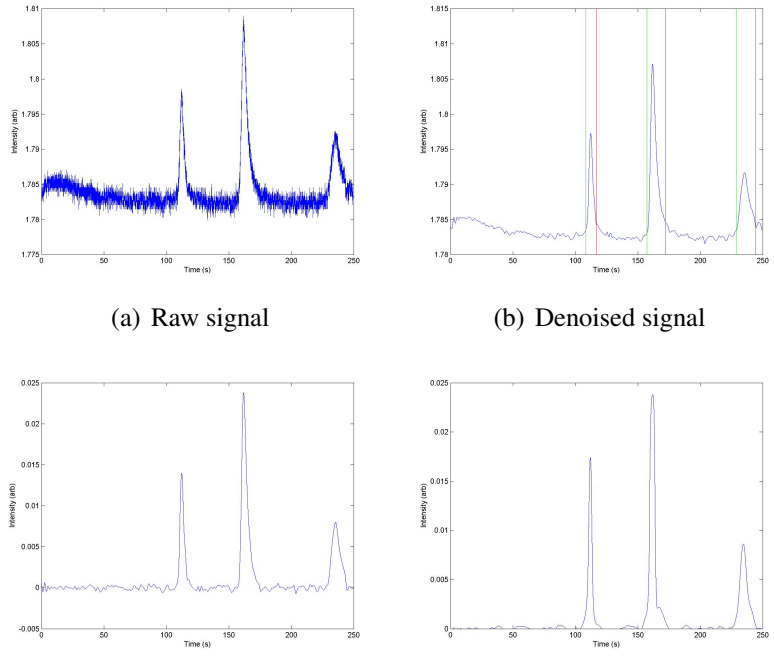
(a) Raw signal (b) Denoised signal

(c) Baseline variation removed signal (d) Deconvolved signal

Figure 7.6: Signal processing stages of BKV from thermo-cycler with three peaks with SNR of 69.2, 82, and 21.4 V/V. The baseline variation and noise were successfully removed.

fully removed the baseline variations and offsets in 32 runs. For the remaining 12 runs where additional peak regions were identified (due to abrupt baseline variations), four runs resulted in varying baselines and small artifacts were created in three runs.

An example in which our signal processing algorithm detected an additional peak region in a single peak signal (1μL of Cy-5 end-labelled DNA mixed with 3μL of 0.01xTTE) is shown in Fig. 7.12(a). Our signal processing algorithm identified an additional peak region near the beginning of the signal as shown in Fig. 7.12(b). As shown in the first derivative of the denoised signal (Fig. 7.13), the abrupt baseline variation at the beginning caused the first derivative of the signal to exceed the upper and lower thresholds for more than the prespecified minimum peak width.



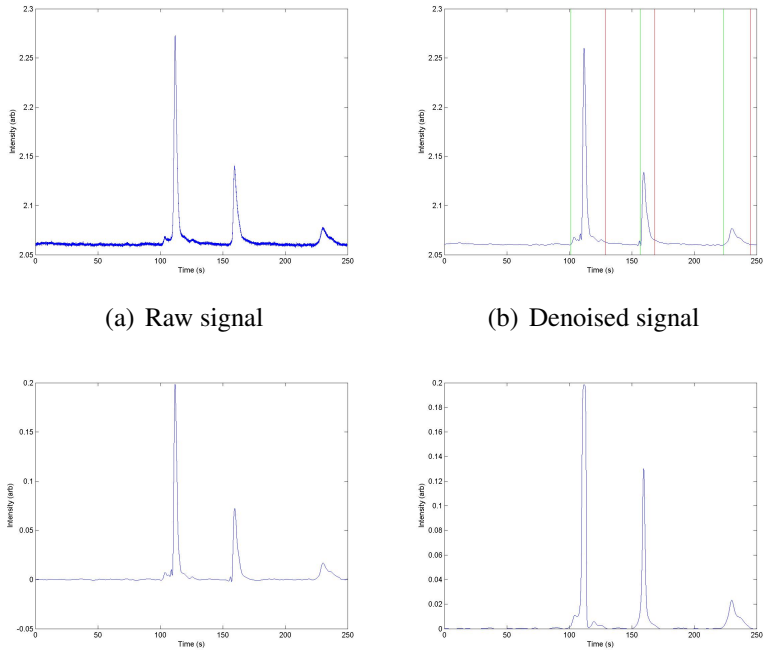
(a) Raw signal (b) Denoised signal

(c) Baseline variation removed signal (d) Deconvolved signal

Figure 7.7: Signal processing stages of BKV from thermo-cycler with 3 peaks with SNR of 14.4, 24.7, 7.5 V/V. The baseline variation and noise were successfully removed.

This caused a false peak region detection. The error was amplified by deconvolution, which increased the height of artifacts. Our signal processing algorithm did not remove abrupt baseline variations with long durations.

In the 74 CE runs with multiple peak samples, there were only four runs where peak regions were not correctly detected. This led to the creation of artifacts and varying baselines in these four runs. Fig. 7.14(a) shows the electropherogram of a 0.5 μL of BKV PCR product from a thermo-cycler mixed with 3.5 μL of 0.01xTTE. It resulted in three-peak CE signal with a large baseline variation in the beginning and a small unspecific peak near the end of the signal. As shown in Fig. 7.14(b), our signal processing algorithm failed to detect the last peak because the first derivative of the third peak did not exceed the peak region threshold. This could be corrected



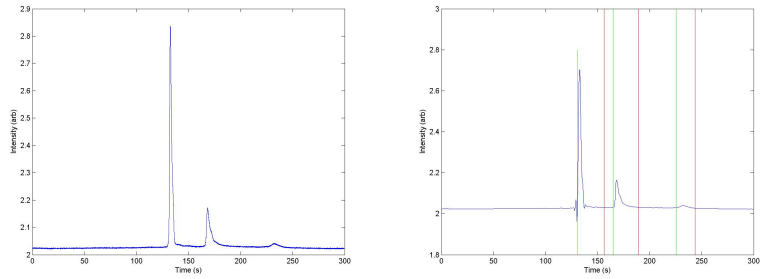
(a) Raw signal (b) Denoised signal

(c) Baseline variation removed signal (d) Deconvolved signal

Figure 7.8: Signal processing stages of a on-chip BKV sample with distributed peak heights. Our algorithm denoised the signal and identified the peak regions.

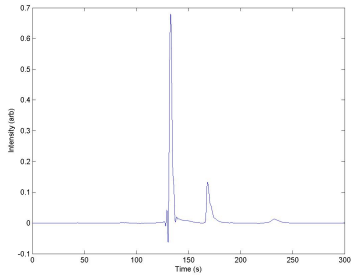
by lowering the zone threshold (via a zone threshold multiplier), but it could lead to longer peak region times for all peaks. Our signal processing algorithm did not detect peaks with very small peak amplitudes relative to baseline variations.

Another limitation to our signal processing algorithm was that it did not completely separate the peaks for signals with low resolution. Jansson's deconvolution can separate signals with low resolution by setting the relaxation factor high or increasing the maximum number of iterations. Because our goal was to separate overlapping peaks without creating artifacts, we used a high deconvolution factor threshold value. As a result, when peak resolution was low, our signal processing algorithm did not separate the peaks completely, as shown in Fig. 7.15(d). This sample shown in this figure is a 4 μL of BKV PCR product obtained from a on-chip PCR performed on the TTK.

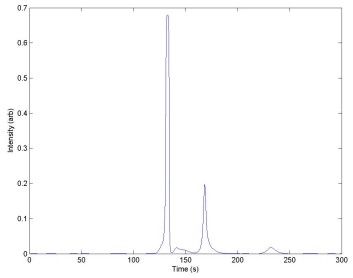


(a) Raw signal

(b) Denoised signal



(c) Baseline variation removed signal

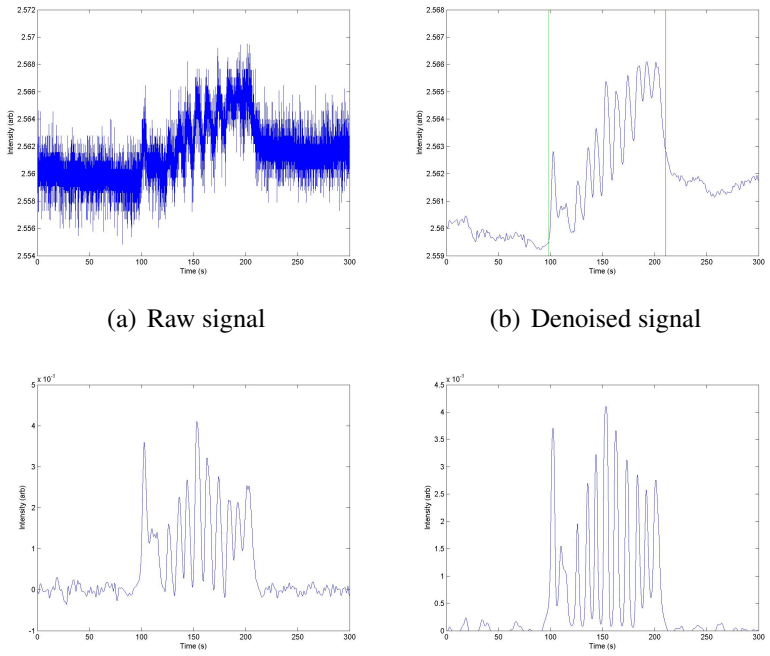


(d) Deconvolved signal

Figure 7.9: Signal processing stages of a on-chip BKV sample with distributed peak heights. Our signal processing algorithm determined the peak region of each peak.

Using our signal processing algorithm on 117 CE runs, 14 (12%) of the post-processing runs resulted in varying baselines or created artifacts because the SNR of the raw signal was too low or the baseline variation to peak height ratio was too high. However, signal processing success rate is hard to quantify and compare. From our literature review of denoising, baseline variation removal, and overlapped peaks separation algorithms, there was not a single journal article reported on their signal processing success rate for synthetic or experimental signals. Clearly the success rate of a signal processing is not easy to quantify.

Although a success rate of 88% might appear low, we believe it is because we calculated our signal processing's success rate based on the entire CE dataset collected by M.Sc student Allison Bidulock. We did not take the external environments that might have affected the quality of CE runs into account. External environments



(a) Raw signal (b) Denoised signal

(c) Baseline variation removed signal (d) Deconvolved signal

Figure 7.10: DNA ladder CE electropherogram with a low SNR (about 6 V/V on the highest peak). The automated signal processing algorithm denoised the signal, removed the baseline and separated the overlapping peaks. The SNR of the highest peak was found to be 396 V/V after processing.

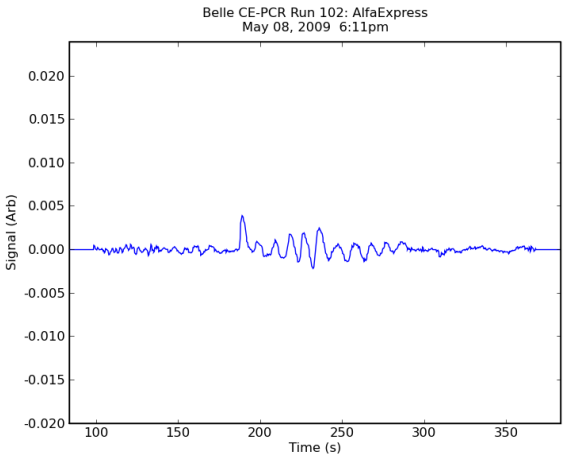
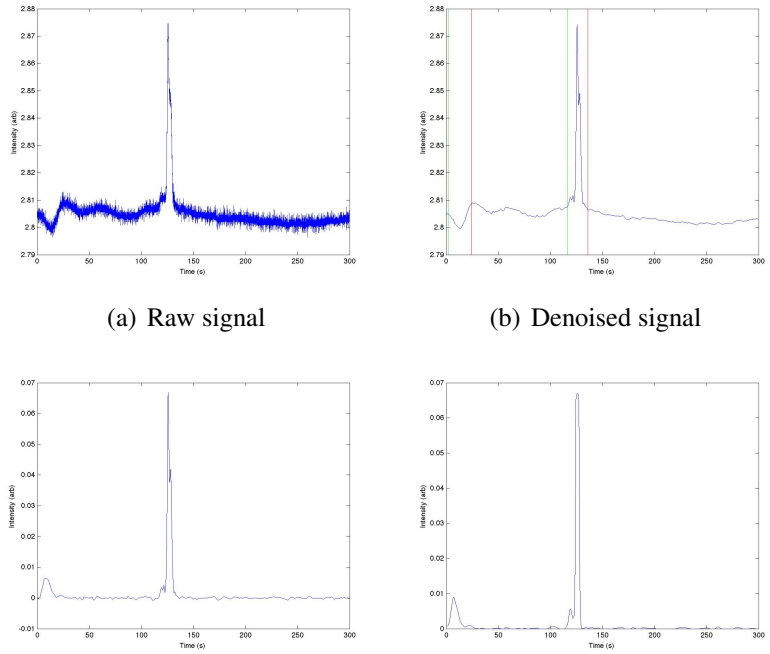


Figure 7.11: Ladder sample processed by the current TTK software.



(a) Raw signal (b) Denoised signal

(c) Baseline variation removed signal (d) Deconvolved signal

Figure 7.12: Signal processing stages of a single peak signal. An abrupt baseline variation at the beginning of the signal created an artifact in the post signal processing.

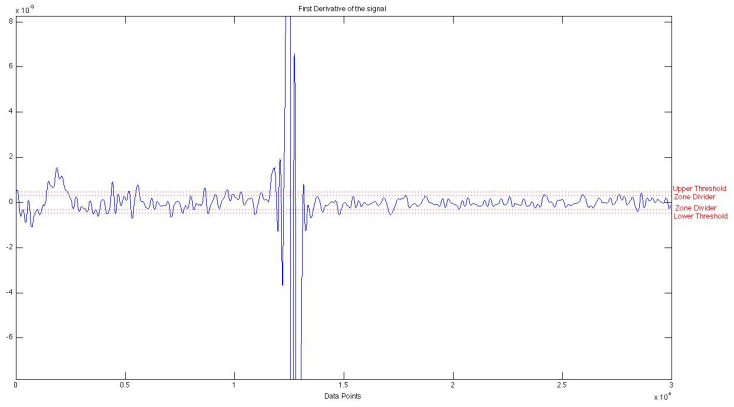
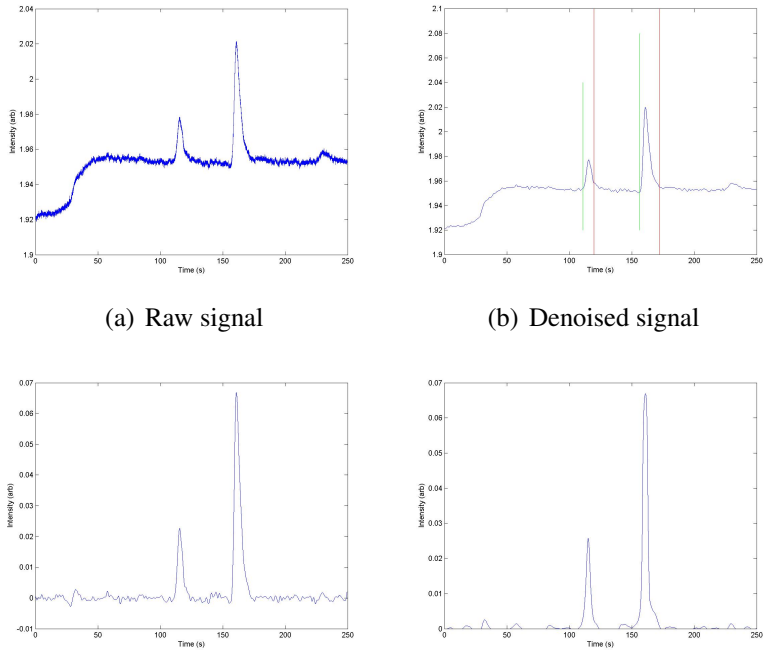


Figure 7.13: The first derivative of the signal shown in Fig. 7.12(b).



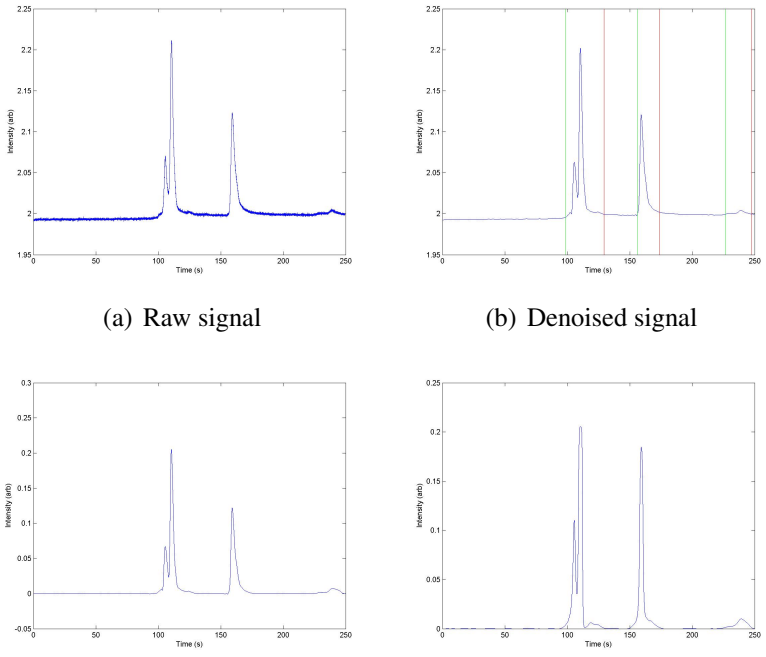
(a) Raw signal (b) Denoised signal

(c) Baseline variation removed signal (d) Deconvolved signal

Figure 7.14: Signal processing stages of a three-peak CE signal with a varying baseline. The small peak's region could not be detected due to its low SNR compared to the baseline variation.

such as laser misalignment, electronic noise (excluding amplifier and photodiode), poor microfluidic channel coating, and bad reagents, all can have a huge effect on the SNR and the baseline variation in the recorded electrophoergrams. We believe that our signal processing will have a much higher success rate if we filtered out CE runs with these external experimental errors. We also believe that the results we have shown here are not phenomenological because we have successfully processed (i.e/ removed noise and baseline variations, and separated overlapping peaks) a wide array of CE data from the TTK, including CE runs that might have affected the signal quality because of external factors.

Lastly, the performance of our complete CE signal processing algorithm could be improved by changing the algorithm parameters for specific signals. However,



(a) Raw signal (b) Denoised signal

(c) Baseline variation removed signal (d) Deconvolved signal

Figure 7.15: A case where our signal processing algorithm did not completely separate the overlapping peaks.

changing a set of parameters might work for one signal but could have adverse effects on other signals. For example, the zone threshold multiplier can be lowered to increase peak region sensitivity, but lowering it can lead to false peak region detection in noisy signals. Using a higher level decomposition in WT removes more noise in noisy signals but it distorts peak shapes. Increasing the relaxation factor or decreasing the deconvolution factor threshold in Jansson's deconvolution increases peak separation, but artifacts are more easily created.

7.3.4 Resolution

The most recent published resolution for the TTK was 12 bp [29]. This journal article used median subtraction to remove baseline offset and variations. As shown by the processing of a ladder sample with 50-500 bp in Fig. 7.11, median subtraction

caused the peaks to dip below the baseline reference and resulted in a non-physical signal [101]. Because median subtraction distorts signals with wide or overlapping peaks, the resolution for the TTK was calculated using BKV samples with separated peaks.

The same ladder sample was processed with our signal processing algorithm and the post-processing signal is shown in Fig. 7.10(d). The peak locations and heights for the post processing signal are shown in Table 7.5. We calculated the resolution (in bp) of the TTK using equation (7.1) [102],

$$RES(bp) = \frac{(w1+w2)}{2}\frac{\Delta M}{\Delta t} \tag{7.1}$$

where $w1$ and $w2$ are the full width at half maximum (FWHM) of the peaks, Δt is the separation time between the peaks, and ΔM represents the size difference between the DNA molecules in bp. Because there are multiple peaks (or DNA molecules) in this sample, we calculated the resolution of all adjacent peaks. The average resolution for all adjacent peaks was calculated to be 20.5 bp. We believe that our signal processing has a degraded resolution compared to median subtraction because the median values calculated in the rising and falling regions of isolated peaks, have higher values than the actual baseline. Subtracting the median values in these regions effectively reduces peak widths. But as discussed, median subtraction distorts signals with wide or overlapping peaks and cannot be used to calculate the resolution for the DNA ladder sample.

7.3.5 Limit of Detection Improvement

We reprocessed experimental data recently published by the AML [29] and recalculated the LOD of the TTK using our signal processing algorithm. We used a standard procedure to determine the LOD [103]. Three different concentrations (0.749, 0.498, and 0.249 ng/μL) of Cy-5 DNA primer were used, and each concentration was loaded twice and two CE runs were performed with each DNA primer

Peak Number	1	2	3	4	5	6
Peak Amplitude (V)	0.0041	0.0018	0.0025	0.0032	0.0038	0.0042
Peak Time (s)	102.3	109.9	125.6	135.5	143.6	153.6
Peak FWHM (s)	3.8	8.5	2.6	3.2	3.1	5.1
Size (bp)	-	50	100	150	200	250
Peak Number	7	8	9	10	11	
Peak Amplitude (V)	0.0038	0.0034	0.0031	0.003	0.0028	
Peak Time (s)	162.3	173	183.4	192	200.2	
Peak FWHM (s)	4.3	3.8	3.5	3.2	6.5	
Size (bp)	300	350	400	450	500	

Table 7.5: Post-processing of red ladder peak information

load.

The raw data from these CE experiments were processed by our signal processing algorithm, as shown in Fig. 7.16. The average post-processing peak heights for 0.749, 0.498, and 0.249 ng/μL primer concentration were 0.0659 V, 0.0442 V, and 0.0372 V, respectively. The average standard deviation of the post-processing noise (first 50 seconds) was 50 μV. The LOD was calculated by fitting a best fit line to the peak height against concentration plot. Linear extrapolation was used to find the concentration where the peak height is three times of the noise. We found the TTK's LOD to be 1.7 pg/μL with our signal processing algorithm.

Our signal processing algorithm reduced the LOD of the TTK from 6 pg/μL [29] to 1.7 pg/μL. The LOD using our signal processing algorithm is smaller than the LOD of the μTK, which is a commercial CE genetic analysis instrument with a LOD of 2.3 pg/μL [29]. No other LOD record for commercial devices could be found during our review.

One of the key contributing factors for the LOD improvement is wavelet denoising. As discussed in Chapter 4, wavelet denoising is excellent at peak information preservation while removing noise. Wavelet denoising allowed for the removal of noise while maintaining peak heights, hence increasing the post-processing SNR

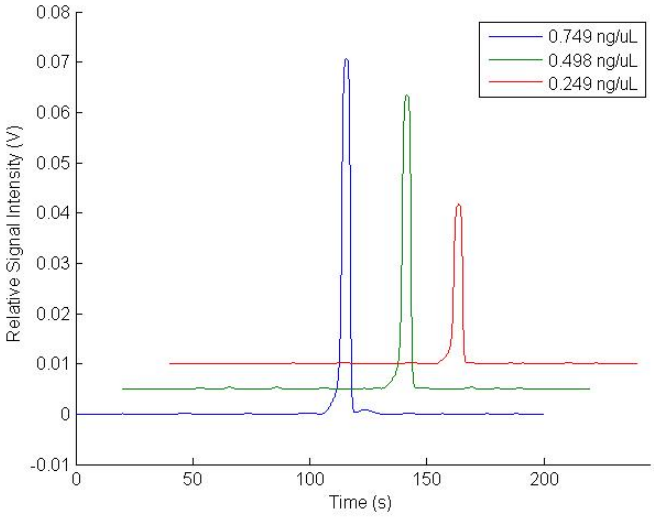


Figure 7.16: Limit of detection using our signal processing algorithm of 0.749, 0.498, and 0.249 ng/μL of DNA.

and decreasing the LOD.

7.3.6 Algorithm Limitations

Using AML's standard CE protocol and our signal processing algorithm, the lowest concentration of Cy-5 end-labelled DNA was 0.498 ng/μL. In synthesised CE signals, the lowest SNR our algorithm was able to process was 2 V/V. We also showed that our algorithms can remove baseline variations for signals with peak to baseline ratio as low as 1.4 V/V.

However, the implementation of Jansson's deconvolution to separate overlapping peaks and to improve resolution requires an accurate model of the TTK's point spread function. Our signal processing algorithm runs on parameters that was tune specificly for TTK's CE signals. If the point spread function was incorrectly modelled or the experimental conditions are changed, then our algorithm may create artifacts.

7.3.7 Executable Signal Processing Function

Using the MATLAB compiler and the Microsoft Visual Studio 2008 compiler, we turned our signal processing scripts into an executable program for Windows computers. MATLAB is not required to run this program. This executable program opens a dialog box and allows the user to select a text file saved by the TTK software v3.2x from a CE run. It automatically parses through the text file and extracts the raw data and performs wavelet denoising, IPBF, and Jansson's deconvolution. The executable saves the stages of processing in a JPEG image, as well as the denoised signal and the peak locations in text files. The user guide can be found in Appendix A.7.

Our signal processing algorithm requires a powerful computer. Running our signal processing algorithm on a modern computer with 2.8 GHz quad-core with 6 GB of memory, processing a sample signal took 16 seconds. Running on an older computer with Pentium 4 1.7 GHz and 768 MB of memory, processing the same signal took 140 seconds.

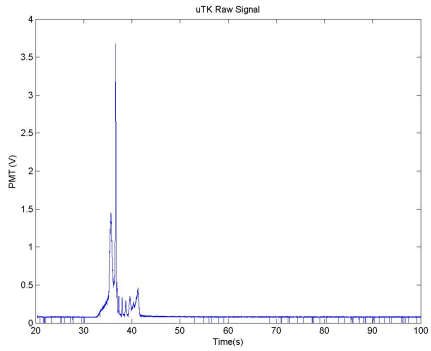
7.4 Applicability to the μTK

To demonstrate that our signal processing algorithm is generally applicable for different signals, we processed and analysed some CE runs performed on a μTK by a M.Sc student Samira Movahedi. Wavelet denoising of CE signals with various number of peaks and resolutions are shown in Fig. 7.17. Wavelet denoising was able to remove the noise in these signals while preserving peak information. Baseline offset and variation removal are not required for μTK signals because the area of detection of PMT device is small and the baseline is so low that the laser variation has a negligible affect. We also performed Jansson's deconvolution on the denoised μTK signals. We found Jansson's deconvolution created artifacts in all of the signals we tested. This is because the point spread function (PSF) of the μTK is

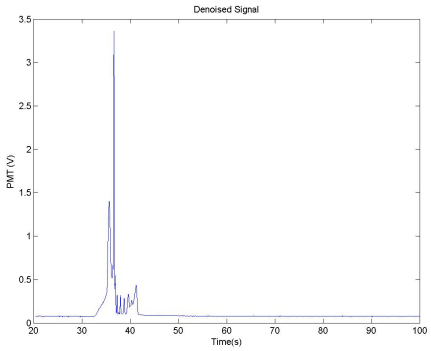
different than that of the TTK's, which is used in our signal processing algorithm. We believe that Jansson's deconvolution can separate overlapping peaks for μTK signals if an accurate model of the PSF was used.

7.5 Conclusion

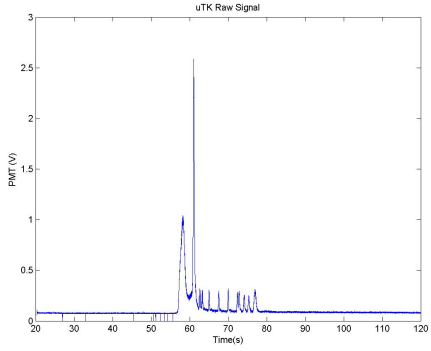
In this chapter, we presented our signal processing algorithm for CE. Our algorithm consists of wavelet denoising, iterative polynomial baseline fit, and Jansson's deconvolution. Without additional parameter tweaking, our CE signal processing algorithm improved SNR, resolution, and removed baseline variations for various types of CE signals gathered from an inexpensive non-confocal LIF-CE instrument. Using our signal processing algorithm, we lowered the LOD of the TTK from 6 pg/μL to 1.7 pg/μL.



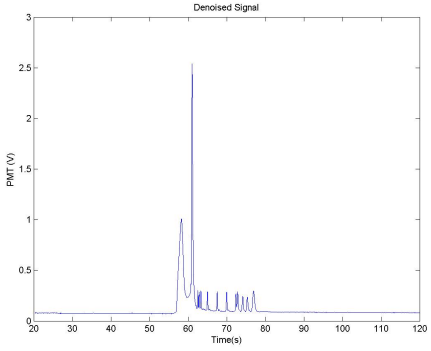
(a) Raw signal



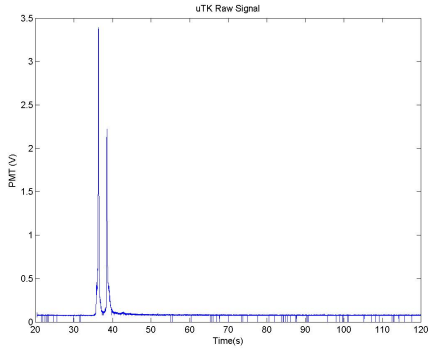
(b) Denoised signal



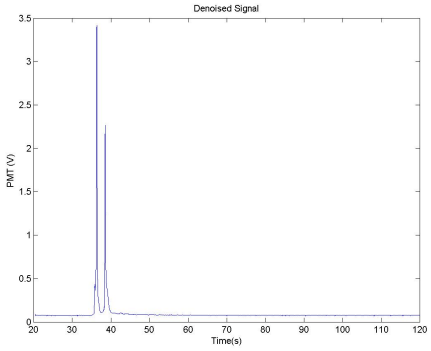
(c) Raw Signal



(d) Denoised Signal



(e) Raw Signal



(f) Denoised signal

Figure 7.17: Wavelet denoising of signals from a μTK.

Chapter 8

Conclusion

8.1 Summary

In this thesis, we looked at the extraction of information from the TTK. We have made contributions to the TTK in two important areas: electronics and signal processing algorithms.

Chapter 3 presented our contribution to the electronics which involved the design and test of a multistage amplifier that extends the bandwidth to extract extra signal information compared to the TTK's current electronics. This multistage amplifier has the same current to voltage gain (-1×10^9 V/A) as the existing single-stage amplifier, with improved bandwidth (160 Hz from 1.5 Hz) and voltage swing (4 V from 3.5 V). We performed a detailed noise analysis on our design and compared it to the single-stage amplifier design. Our analysis showed that because we were successful in extending the bandwidth, we noticed an increased presence of noise. We then made recommendations that will help to lower the noise in a future multistage amplifier design for the TTK.

In Chapter 4 we made a contribution to the signal processing aspects by studying the parameters of wavelet transform denoising that would be most applicable to the TTK system. We developed a method based on noise removal efficiency and peak preservation metrics to find the most reliable wavelet denoising parameters for a variety of CE signals. From our analysis, we found that the Symlet8 wavelet with

level 8 decomposition provided the most reliable denoising among those signals we tested. By comparing the peak preservation and noise removal performance between wavelet denoising, Fourier filtering, Savitzky-Golay filtering, and moving average, we found that wavelet denoising outperformed all the other noise removal algorithms we tested. We showed that wavelet denoising provided a low root mean square error (RMSE) of 0.0889, an average peak shift of 0.058 seconds, and a peak height error of 2.5% for a synthetic CE signal with a SNR of 5 V/V. We then demonstrated wavelet denoising for experimental CE signals.

Although wavelet transforms can be used to remove high-frequency noise, low frequency noise, such as baseline variation in CE signals, requires a different treatment. In Chapter 5 we presented an algorithm capable of removing baseline variation and found parameters that are applicable to the TTK. By combining our peak region detection algorithm with an iterative polynomial baseline fit (IPBF) algorithm, we accurately estimated and removed the baseline variations in synthetic CE signals with SNR and peak-to baseline variation ratio as low as 1.4 V/V and 0.75 V/V, respectively. This algorithm was used in experimental signals and showed that it can remove the baseline variations for signals with single and multiple peak regions.

The focus Chapter 6 was to separate overlapping peaks observed in CE signals to increase the number of biological applications for TTK. In this chapter we presented a modified Jansson's deconvolution algorithm. By incorporating normalization, peak detection, and a deconvolution factor into Jansson's deconvolution, we minimized artifacts while maximizing peak separation for CE signals. Our modified Jansson's deconvolution isolated overlapping peaks in synthetic CE signals with a resolution as low as 0.7 as well as experimental CE signals with low and high resolutions.

In Chapter 7 we combined the methods we previous studied and used them to perform signal processing on a large number of experimental CE signals. We

showed that the combined algorithm improved SNR and resolution for various CE experimental signals without additional tweaking of algorithm parameters. The combined algorithm removed the noise and baseline variations in single peak signals with a SNR as low as 2.9 V/V. More importantly, the TTK's LOD was reduced from 5.8 pg/μL to 1.7 pg/μL, which is better than the μTK, a commercial CE instrument used in the AML.

In conclusion, four main issues in regards to the processing of TTK's CE signal processing are presented and addressed in this thesis. First, we designed and tested a multistage amplifier for the optical detection subsystem in the TTK to extract additional signal information. Then we implemented wavelet transform and IPBF to remove the high-frequency noise and baseline variation, the conditions commonly observed in non-confocal CE systems. Lastly, we separated overlapping peaks to enable future miniaturisation of microfluidics.

An inexpensive and portable genetic analysis instrument with a low LOD could be used in a point-of-care (POC) mode in hospitals, clinics, and homes. Greater accessibility to molecular medical diagnosis would help practitioners diagnose potential problems early and lower the cost of diagnostic procedures.

8.2 Future Improvements

Our signal processing algorithm is currently implemented in MATLAB and an executable program for Windows PC. The current TTK software is written in Python. An improvement to this project would be integrating the signal processing algorithm with the TTK software by having the software call the executable program automatically.

Another improvement to this project would be to build the recommended multistage amplifier and use it in a TTK to perform CE experiments. If this amplifier to be successfully implemented, the higher bandwidth achieved has a potential to provide more insight of noise characteristics and more advanced signal processing

techniques can be used. In terms of future miniaturisation, integration of the multistage amplifier onto a CMOS circuit alongside a digital signal processing chip to perform real time signal processing could lower cost and improve speed.

Bibliography

[1] P. Yager, T. Edwards, E. Fu, K. Helton, K. Nelson, M. R Tam, and B. H Weigl, "Microfluidic diagnostic technologies for global public health," *Nature*, vol. 442, no. 7101, pp. 412–418, 2006.

[2] Centers for Disease Control and Prevention, " Seasonal Influenza (Flu) - Antiviral Drugs - Side Effects," `http://www.cdc.gov/flu/protect/antiviral/sideeffects.htm`, July 2009, [Accessed May 2010].

[3] S. Brin, "Sergey Brin Online Blog, LRRK2," `http://too.blogspot.com/2008/09/lrrk2.html`, September 2008, [Accessed June 2010].

[4] X. Zhang, L. Li, D. Wei, Y. Yap, and F. Chen, "Moving cancer diagnostics from bench to bedside," *Trends in Biotechnology*, vol. 25, no. 4, pp. 166–173, 2007.

[5] L. A. Gloeckler Ries, M. E. Reichman, D. R. Lewis, B. F. Hankey, and B. K. Edwards, "Cancer Survival and Incidence from the Surveillance, Epidemiology, and End Results (SEER) Program," *Oncologist*, vol. 8, no. 6, pp. 541–552, Dec 2003.

[6] O. J. Semmes, Z. Feng, B. L. Adam, L. L. Banez, and et al., "Evaluation of Serum Protein Profiling by Surface-Enhanced Laser Desorption/Ionization Time-of-Flight Mass Spectrometry for the Detection of Prostate Cancer: I. Assessment of Platform Reproducibility," *Clin Chem*, vol. 51, no. 1, pp. 102–112, 2005.

[7] J. R. Webster, M. A. Burns, D. T. Burke, and C. H. Mastrangelo, "Monolithic Capillary Electrophoresis Device with Integrated Fluorescence Detec-

tor," *Analytical Chemistry*, vol. 73, no. 7, pp. 1622–1626, Apr. 2001.

[8] J. Kling, "Moving diagnostics from the bench to the bedside," *Nature Biotechnology*, vol. 24, no. 8, pp. 891–898, 2006.

[9] P. S. Dittrich and A. Manz, "Single-molecule fluorescence detection in microfluidic channelsthe Holy Grail in TAS?," *Analytical and Bioanalytical Chemistry*, vol. 382, no. 8, pp. 1771–1782, 2005.

[10] W. R. Rodriguez, N. Christodoulides, P. N. Floriano, S. Graham, S. Mohanty, and et al., "A Microchip CD4 Counting Method for HIV Monitoring in Resource-Poor Settings," *PLoS Med*, vol. 2, no. 7, pp. 182, July 2005.

[11] J. P. Landers, "Molecular Diagnostics on Electrophoretic Microchips," *Analytical Chemistry*, vol. 75, no. 12, pp. 2919–2927, June 2003.

[12] G. V. Kaigala, *Genetic Analysis Using Lab-on-a-Chip Technolgies*, Ph.D. thesis, University of Alberta, Edmonton, AB, 2009.

[13] V. Olazabal, L. Prasad, P. Stark, and J. A. Olivares, "Application of wavelet transforms and an approximate deconvolution method for the resolution of noisy overlapped peaks in DNA capillary electrophoresis," *The Analyst*, vol. 129, no. 1, pp. 73–81, 2004.

[14] B. Pokric, N. M. Allinson, E. T. Bergstrom, and D. M. Goodall, "Dynamic analysis of capillary electrophoresis data using real-time neural networks," *Journal of Chromatography A*, vol. 833, no. 2, pp. 231–244, Feb. 1999.

[15] A. Fanigliulo, F. Bortolotti, J. Pascali, F. Tagliaro, and M.J. Bogusz, "Chapter 15 Forensic toxicological screening with capillary electrophoresis and related techniques," in *Forensic Science*, vol. Volume 6, pp. 513–534. Elsevier Science B.V., 2007.

[16] P. Pittet, J. M. Galvan, G. N. Lu, L. Blum, and B. D. Leca-Bouvier, "CMOS LIF detection system for capillary analysis," *Sensors and Actuators B-Chemical*, vol. 97, no. 2-3, pp. 355–361, Feb. 2004.

[17] E. Verpoorte, "Microfluidic chips for clinical and forensic analysis," *Elec-*

trophoresis, vol. 23, no. 5, pp. 677–712, 2002.

[18] M. U. Kopp, H. J. Crabtree, and A. Manz, "Developments in technology and applications of microsystems," *Current Opinion in Chemical Biology*, vol. 1, no. 3, pp. 410–419, 1997.

[19] R. Stewart, A. Wee, D. B. Grayden, and Y. Zhu, "Capillary electrophoresis (CE) peak detection using a wavelet transform technique," in *Biomedical Applications of Micro- and Nanoengineering IV and Complex Systems*, Melbourne, VIC, Australia, 2009, vol. 7270, p. 12, SPIE.

[20] A. T. Woolley, K. L., A. N. Glazer, and R. A. Mathies, "Capillary Electrophoresis Chips with Integrated Electrochemical Detection," *Analytical Chemistry*, vol. 70, no. 4, pp. 684–688, Feb. 1998.

[21] A. Shevchenko, M. Wilm, O. Vorm, and M. Mann, "Mass Spectrometric Sequencing of Proteins from Silver-Stained Polyacrylamide Gels," *Analytical Chemistry*, vol. 68, no. 5, pp. 850–858, 1996.

[22] T. Shimomura, C. Izawa, and T. Matsui, "A highly sensitive, highly reproducible laser-induced fluorescence detection system with optical pickup," *Measurement Science and Technology*, vol. 19, no. 8, 2008.

[23] K. Uchiyama, H. Nakajima, and T. Hobo, "Detection method for microchip separations," *Analytical and bioanalytical chemistry*, vol. 379, no. 3, pp. 375–382, 2004.

[24] V. Srinivasan, V. K. Pamula, and R. B. Fair, "An integrated digital microfluidic lab-on-a-chip for clinical diagnostics on human physiological fluids," *Lab on a Chip*, vol. 4, no. 4, pp. 310, 2004.

[25] C. J. Backhouse, T. Footz, S. Adamia, and L. M. Pilarski, "Microfluidic chips for the molecular analysis of human cancer," Banff, Canada, 2003, pp. 377–382, International Conference on MEMS, NANO and Smart Systems.

[26] F. C. Huang, C. S. Liao, and G. B. Lee, "An integrated microfluidic chip for DNA/RNA amplification, electrophoresis separation and on-line optical

detection," *Electrophoresis*, vol. 27, no. 16, pp. 3297–3305, 2006.

[27] P. A. G. Butler, B. Mills, and P. C. Hauser, "Capillary electrophoresis detector using a light emitting diode and optical fibres," *The Analyst*, vol. 122, no. 9, pp. 949–953, 1997.

[28] N. Manaresi, A. Romani, G. Medoro, L. Altomare, A. Leonardi, M. Tartagni, and R. Guerrieri, "A CMOS chip for individual cell manipulation and detection," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 12, pp. 2297–2305, 2003.

[29] G. V. Kaigala, M. Behnam, A. C. Bidulock, C. Bargen, R. W. Johnstone, D. G. Elliott, and C. J. Backhouse, "A scalable and modular lab-on-a-chip genetic analysis instrument," *The Analyst*, vol. 135, pp. 1606–1617, 2010.

[30] M. E. Johnson and J. P. Landers, "Fundamentals and practice for ultrasensitive laser-induced fluorescence detection in microanalytical systems," *Electrophoresis*, vol. 25, no. 21-22, pp. 3513–3527, 2004.

[31] F. B. Myers and L. P. Lee, "Innovations in optical microfluidic technologies for point-of-care diagnostics," *Lab on a Chip*, vol. 8, no. 12, pp. 2015–2031, 2008.

[32] M. Behnam, G. V. Kaigala, M. Khorasani, P. Marshall, C. J. Backhouse, and D. G. Elliott, "An integrated CMOS high voltage supply for lab-on-a-chip systems," *Lab on a Chip*, vol. 8, no. 9, pp. 1524–1529, 2008.

[33] M. Behnam, G. V. Kaigala, J. M. Quijada, S. Choi, and C. J Backhouse, "Highly compact instrumentation using CMOS technology for genetic amplification and microchip electrophoresis Poster," in *The 14th International Conference on Miniaturized Systems for Chemistry and Life Sciences*, Jutu, Korea.

[34] T.A. Heumier and J. L. Carlsten, "Application Note: Mode hopping in semiconductor lasers," *Dept of Physics, Montana State University, Bozeman, MT, USA*, pp. 1–9.

[35] Y. Sun and Y. C. Kwok, "Polymeric microfluidic system for DNA analysis," *Analytica Chimica Acta*, vol. 556, no. 1, pp. 80–96, 2006.

[36] M. Behnam, "Integration of High Voltage CMOS and Microfluidic Technologies for Genetic Analysis Systems," Master of Science Thesis, University of Alberta, 2007.

[37] R. C. Ebersole and R. M. McCormick, "Separation and isolation of viable bacteria by capillary zone electrophoresis," *Nature Biotechnology*, vol. 11, no. 11, pp. 1278–1282, 1993.

[38] G. V. Kaigala, V. N. Hoang, A. Stickel, J. Lauzon, D. Manage, L. M. Pilarski, and C. J. Backhouse, "An inexpensive and portable microchip-based platform for integrated RT–PCR and capillary electrophoresis," *The Analyst*, vol. 133, no. 3, pp. 331–338, 2008.

[39] P. Horowitz and W. Hill, *The Art of Electronics*, Cambridge University Press, Cambridge, UK, 1989.

[40] Hamamatsu Corporation, "Photodiode Technical Guide," `http://sales.hamamatsu.com/assets/html/ssd/si-photodiode/index.htm`, [Accessed March 2010].

[41] R. Sarpeshkar, T. Delbruck, and C. A. Mead, "White noise in MOS transistors and resistors," *IEEE Circuits and Devices Magazine*, vol. 9, no. 6, pp. 23–29, 1993.

[42] "Noise Analysis in Operational Amplifier Circuits, Digital Signal Processing Solutions," `http://focus.ti.com/lit/an/slva043b/slva043b.pdf`, 2007, [Accessed March 2010].

[43] H. Zhu, S. M. Clark, S. C. Benson, H. S. Rye, A. N. Glazer, and R. A. Mathies, "High-Sensitivity capillary electrophoresis of Double-Stranded DNA fragments using monomeric and dimeric fluorescent intercalating dyes," *Analytical Chemistry*, vol. 66, no. 13, pp. 1941–1948, July 1994.

[44] H. Swerdlow, J. Z Zhang, D. Y Chen, H. R Harke, R. Grey, S. Wu, N. J

Dovichi, and C. Fuller, "Three DNA sequencing methods using capillary gel electrophoresis and laser-induced fluorescence," *Analytical Chemistry*, vol. 63, no. 24, pp. 2835–2841, 1991.

[45] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures.," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[46] C. Perrin, B. Walczak, and D. L. Massart, "The Use of Wavelets for Signal Denoising in Capillary Electrophoresis," *Analytical Chemistry*, vol. 73, no. 20, pp. 4903–4917, 2001.

[47] B. F. Liu, Y. Sera, N. Matsubara, K. Otsuka, and S. Terabe, "Signal denoising and baseline correction by discrete wavelet transform for microchip capillary electrophoresis," Germany, Sept. 2003, vol. 24 of *Electrphoresis (Germany)*, pp. 3260–3265, Wiley-VCH.

[48] D. Feuerstein, K. H. Parker, and M. G. Boutelle, "Practical Methods for Noise Removal: Applications to Spikes, Nonstationary Quasi-Periodic Noise, and Baseline Drift," *Analytical Chemistry*, vol. 81, no. 12, pp. 4987–4994, June 2009.

[49] T. O'Haver, "Introduction to Signal Processing: Differentiation," `http://terpconnect.umd.edu/~toh/spectrum/Differentiation.html`, [Accessed August 2009].

[50] V. Dohnal, F. Zhang, H. Li, and J. Havel, "Quantitative chiral analysis in capillary electrophoresis from unresolved peaks using derivative electropherograms, experimental design, and artificial neural networks," *Electrophoresis*, vol. 24, no. 15, pp. 2462–2468, 2003.

[51] S. B. Cheng, C. D. Skinner, J. Taylor, S. Attiya, W. E. Lee, G. Picelli, and D. J. Harrison, "Development of a multichannel microfluidic analysis system employing affinity capillary electrophoresis for immunoassay," *Anal. Chem*, vol. 73, no. 7, pp. 1472–1479, 2001.

[52] G. Vivo-Truyols, J. R. Torres-Lapasio, A. M. van Nederkassel, Y. Vander

Heyden, and D. L. Massart, "Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals part I: peak detection," *Journal of Chromatography A*, vol. 1096, no. 1-2, pp. 133–145, 2005.

[53] X. G. Shao, A. K. M. Leung, and F. T. Chau, "Wavelet: A new trend in chemistry," *Accounts of Chemical Research*, vol. 36, no. 4, pp. 276–283, Apr. 2003.

[54] I. Daubechies, "Orthonormal bases of wavelets with finite support-connection with discrete filters," in *Wavelets. Time-Frequency Methods and Phase Space*, 1989, p. 38.

[55] S. G. Mallat, "Multiresolution approximations and wavelet orthonormal bases of L 2 (R)," *Transactions of the American Mathematical Society*, vol. 315, no. 1, pp. 69–87, 1989.

[56] B. K. Alsberg, A. M. Woodward, M. K. Winson, J. Rowland, and D. B. Kell, "Wavelet denoising of infrared spectra," *The Analyst*, vol. 122, no. 7, pp. 645–652, 1997.

[57] F. Zhang and H. Li, "Resolution of overlapping capillary electrophoresis peaks by using chemometric analysis: Quantification of the components in compound reserpine tablets," *Electrophoresis*, vol. 26, no. 9, pp. 1692–1702, 2005.

[58] W. Cao, X. Chen, X. Yang, and E. Wang, "Discrete wavelets transform for signal denoising in capillary electrophoresis with electrochemiluminescence detection," *Electrophoresis*, vol. 24, no. 18, pp. 3124–3130, 2003.

[59] H. F. Zhang, X. W. Liui, X. H. Shao, X. L. Wang, and B. Zhou, "The study of the improved wavelet thresholding with translation invariant de-noising on capillary electrophoresis signal," in *International Conference on Nano/Micro Engineered and Molecular Systems*, Los Alamitos, CA, USA, 2009, vol. 0, pp. 1099–1102, IEEE Computer Society.

[60] F. Gan, G. Ruan, and J. Mo, "Baseline correction by improved iterative

polynomial fitting with automatic threshold," *Chemometrics and Intelligent Laboratory Systems*, vol. 82, no. 1-2, pp. 59–65, May 2006.

[61] M. S. Friedrichs, "A model-free algorithm for the removal of baseline artifacts," *Journal of Biomolecular NMR*, vol. 5, no. 2, pp. 147–153, Feb. 1995.

[62] L. Keselbrener, M. Keselbrener, and S. Akselrod, "Nonlinear high pass filter for R-wave detection in ECG signal," *Medical Engineering & Physics*, vol. 19, no. 5, pp. 481–484, July 1997.

[63] A. K. Atakan, W. E. Blass, and D. E. Jennings, "Elimination of Baseline Variations from a Recorded Spectrum by Ultra-low Frequency Filtering," *Applied Spectroscopy*, vol. 34, pp. 369–372, May 1980.

[64] P. A. Mosier-Boss, S. H. Lieberman, and R. Newbery, "Fluorescence Rejection in Raman Spectroscopy by Shifted-Spectra, Edge Detection, and FFT Filtering Techniques," *Applied Spectroscopy*, vol. 49, pp. 630–638, May 1995.

[65] Y. G. Hu, T. Jiang, A. G. Shen, W. Li, X. P. Wang, and J. M. Hu, "A background elimination method based on wavelet transform for Raman spectra," *Chemometrics and Intelligent Laboratory Systems*, vol. 85, no. 1, pp. 94–101, 2007.

[66] G. Schulze, A. Jirasek, M. M. L. Yu, A. Lim, R. F. B. Turner, and M. W. Blades, "Investigation of selected baseline removal techniques as candidates for automated implementation," *Applied Spectroscopy*, vol. 59, no. 5, pp. 545–574, May 2005.

[67] A. Wee, D. B. Grayden, Y. G. Zhu, K. Petkovic-Duran, and D. Smith, "A continuous wavelet transform algorithm for peak detection," *Electrophoresis*, vol. 29, no. 20, pp. 4215–4225, Nov. 2008.

[68] K. R. Coombes, S. Tsavachidis, J. Morris, K. Baggerly, M. C. Hung, and H. Kuerer, "Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform," *Pro-*

teomics, vol. 5, no. 16, pp. 4107–4117, 2005.

[69] J. P. Landers, *Handbook of capillary electrophoresis*, CRC Press, Boca Raton, FL, USA, 1997.

[70] J. A. Pierce, R. S. Jackson, K. W. Van Every, P. R. Griffiths, and H. Gao, "Combined deconvolution and curve fitting for quantitative analysis of unresolved spectral bands," *Analytical Chemistry*, vol. 62, no. 5, pp. 477–484, 1990.

[71] W. F. Maddams, "The scope and limitations of curve fitting," *Applied Spectroscopy*, vol. 34, no. 3, pp. 245–267, 1980.

[72] J. K. Kauppinen, D. J. Moffatt, H. H. Mantsch, and D. G. Cameron, "Fourier self-deconvolution: a method for resolving intrinsically overlapped bands," *Applied Spectroscopy*, vol. 35, no. 3, pp. 271–276, 1981.

[73] W. J. Yang, P. R. Griffiths, D. M. Byler, and H. Susi, "Protein conformation by infrared spectroscopy: resolution enhancement by Fourier self-deconvolution," *Applied spectroscopy*, vol. 39, no. 2, pp. 282–287, 1985.

[74] X. Q Zhang, J. B Zheng, and H. Gao, "Comparison of wavelet transform and Fourier self-deconvolution (FSD) and wavelet FSD for curve fitting," *The Analyst*, vol. 125, no. 5, pp. 915–919, 2000.

[75] L. Smeller, K. Goossens, and K. Heremans, "How to minimize certain artifacts in Fourier self-deconvolution," *Applied Spectroscopy*, vol. 49, no. 10, pp. 1538–1542, 1995.

[76] L. Jiao, S. Gao, F. Zhang, and H. Li, "Quantification of components in overlapping peaks from capillary electrophoresis by using continues wavelet transform method," *Talanta*, vol. 75, no. 4, pp. 1061–1067, May 2008.

[77] P. B. Crilly, "Evaluation of Jansson's method for resolving overlapped gas chromatographic peaks," *Journal of Chemometrics*, vol. 1, no. 3, 1987.

[78] P. B. Crilly, "Numerical Deconvolution of Gas Chromatography Peaks Using Janssons Method," *Journal of Chemometrics*, vol. 1, no. 2, 1987.

[79] P. B. Crilly, "The use of a cross-correlation technique to enhance Janssons deconvolution procedure," *Journal of Chemometrics*, vol. 4, no. 4, 1990.

[80] L. Li and T. P. Speed, "Parametric deconvolution of positive spike trains," *The Annals of Statistics*, vol. 28, no. 5, pp. 1279–1301, 2000.

[81] L. Li and T. P. Speed, "Deconvolution of sparse positive spikes: is it ill-posed," *University of California at Berkeley, Department of Statistics Technical Report*, vol. 586, 2000.

[82] P. B. Crilly, A. Bernardi, P. A. Jansson, and L. E. B. da Silva, "Improving the convergence rate of Jansson's deconvolution method," *IEEE Transactions on Instrumentation and Measurement*, vol. 51, no. 6, pp. 1142–1144, 2002.

[83] Texas Instruments, "OPA129 Datasheet," `http://www.datasheetcatalog.com/datasheets_pdf/O/P/A/1/OPA129.shtml`, [Accessed October 2009].

[84] Edmund Optics, "Silicon detectors," `http://www.edmundoptics.com/onlinecatalog/displayproduct.cfm?productid=1305`, [Accessed May 2010].

[85] National Semiconductors, "LMV792 datasheet,," `http://www.national.com/ds/LM/LMV791.pdf`, June 2008, [Accessed August 2010].

[86] E. Cheong, *Electrospray Ionization for Applications in Ion Mobility Spectrometry*, Master of science, University of Alberta, 2003.

[87] eCircuit Center, "Op Amp Bandwidth," `http://www.ecircuitcenter.com/circuits/op_bandwidth1/op_bandwidth1.htm`, 2003, [Accessed June 2010].

[88] Ron Mancini, Bruce Carter, and Texas Instruments Incorporated, *Design Reference: Op Amps for Everyone*, Newnes/Elsevier, Amsterdam; Boston, 3rd edition, 2009.

[89] OSI Optoelectronics, "Photovoltaic Se-

ries," http://www.osioptoelectronics.com/Libraries/Product-Data-Sheets/Photovoltaic-LowNoise-Photodiodes.sflb.ashx, [Accessed July 2011].

[90] W. M. Leach, "On the calculation of noise in multistage amplifiers," *IEEE Transactions on Circuits and Systems I Fundamental Theory and Applications*, vol. 42, no. 3, pp. 176–178, 1995.

[91] S. G. Stacphane and G. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, Elsevier /Academic Press, Amsterdam; Boston, sparse ed edition, 2009.

[92] M. Misiti, Y. Misiti, G. Oppenheim, and J. M Poggi, *Wavelet toolbox: for use with MATLAB: user's guide: version 3*, Math Works, Natick, MA, USA, 2008.

[93] Charles M. Stein, "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, vol. 9, no. 6, pp. 1135–1151, 1981.

[94] G. V. Kaigala, V. N. Hoang, A. Stickel, J. Lauzon, D. Manage, L. M. Pilarski, and C. J. Backhouse, "An inexpensive and portable microchip-based platform for integrated RT-PCR and capillary electrophoresis," *The Analyst*, vol. 133, no. 3, pp. 331–338, Mar. 2008, PMID: 18299747.

[95] J. Li, "Development and evaluation of flexible empirical peak functions for processing chromatographic peaks," *Analytical Chemistry*, vol. 69, no. 21, pp. 4452–4462, 1997.

[96] M. C. Garcia-Alvarez-Coque, E. F. Sim-Alfonso, J. M. Sanchis-Mallols, and J. J. Baeza-Baeza, "A new mathematical function for describing electrophoretic peaks," *Electrophoresis*, vol. 26, no. 11, 2005.

[97] J. G. Shackman, C. J. Watson, and R. T. Kennedy, "High-throughput automated post-processing of separation data," *Journal of Chromatography A*, vol. 1040, no. 2, pp. 273–282, 2004.

[98] L. Kaur, S. Gupta, and RC Chauhan, "Image denoising using wavelet thresholding," in *Indian Conference on computer Vision, Graphics and Image Processing, Ahmedabad*. Citeseer, 2002.

[99] Z. Hua, S. Yan, W. Jie, W. Liqiang, and L. Zukang, "The analysis on the signals denoising and single base pair resolution of DNA sequencing," in *International Symposium on Biophotonics, Nanophotonics and Metamaterials, Metamaterials 2006, October 16, 2006 - October 18, 2006*, Hangzhou, China, 2006, pp. 118–121, Inst. of Elec. and Elec. Eng. Computer Society.

[100] S. W. Smith, *Digital Signal Processing: A Practical Guide for Engineers and Scientists*, Elsevier: Newnes, Amsterdam; Boston, 2003.

[101] A. Bidulock, "Scalable, Modular, Integrated Genetic Analysis Systems," Master of Science Thesis, University of Alberta, Expected Graduation: Fall 2011.

[102] C. J. Backhouse V. J. Sieben, "Rapid on-chip postcolumn labeling and high-resolution separations of DNA," *Electrophoresis*, vol. 26, no. 24, pp. 4729–4742, 2005.

[103] D. MacDougall and W. B Crummett, "Guidelines for data acquisition and data quality evaluation in environmental chemistry," *Analytical Chemistry*, vol. 52, no. 14, pp. 2242–2249, 1980.

Appendix A

MATLAB Code

A.1 Complete CE Signal Processing Algorithm

```
function completeSignalProcessingApp

tic

%Open a text file
[file,dir]=uigetfile('*.txt');


filename_loc=strcat(dir,file);
[raw_data,relay_off]=capture_data(filename_loc);

disp('Processing... Please wait')
%wavelet denoise parameters
wt_name='sym8';
wt_level=8;

%baseline removal parameters
n=20;  %10 is good
max_ilevel=300;
error=1E-6;

%deconvolution parameters
b=1;
k_max=100;
shift=1;
ddelta=0.1;

%noise region parameters
noise_b=0.01;  %noise start
noise_e=50;      %noise end

load('peak_for_conv_isolated.mat');
h=h2(1:1300); %ETG


%%%%%%%%%%%%%%% Thresholds %%%%%%%%%

%1 and 2 workd for RL1,2
ztm=2; %was2,
```

```
rtm=3;  %was 3

zone_mul=ztm;
threshold_mul=rtm;
%%%%%%%%%%%%%%%%%%%%


peak_sep_time_max=10; %was 5
min_span_time=1.5;  %was 1.5

%basic parameters used for debugging
Ts=0.01;
Fs=1/Ts;


file_name=[dir,file];
file_name = regexprep(file_name, '\.', '');
file_name=strrep(file_name, 'txt', '');

file_name2=[dir,file];
file_name2 = regexprep(file_name2, '\.', '');
file_name2=strrep(file_name2, 'txt', '');

[out1,out2,out3,maxtab,offset,bpeaks,dpeaks,peak_start,peak_end]
=complete_CE_sigal_processing(raw_data,wt_name,wt_level,n,max_ilevel,error,
noise_b,noise_e,h,b,k_max,ddelta,threshold_mul,
zone_mul,peak_sep_time_max,min_span_time,file_name,file_name2);
%plot(raw_data);

fid2= fopen([file_name,'_denoised.txt'],'w');
 fprintf(fid2,'%12.12f\n',out3);
 fclose('all');

disp('Processing completed')
toc

end
```

A.2 Wavelet Denoising

```
function [out_denoised,output_baseline,beg_p,end_p]
=wt_denoise_baseline(x,name,level,n,max_ilevel,error,threshold_mul,
zone_mul,peak_sep_time_max,min_span_time,file_name)
%function [out]=wt_denoise_baseline(x,name,level,t_original,n,max_ilevel,error,noise_b,noise_e)
%This function performs 'name' wavelet at level.  t_original is the signal
%length in time, n is the polynomial fitting order, max_ilevel isteh max
%number of iteration in the curve fitting, error is the error factor,
%noise_b and noise_e is the beggining and end of the noise period.beg_p and
%end_p are the begining and end of the peak region

Fs=100;

tptr='heursure';
sorh='s';
scal='one';

p1=wden(x,tptr,sorh,scal,level,name);
out_denoised=p1;

[peak_region_start,peak_region_end,startn,finishn] =
```

```
peak_region_location_mod3(x',threshold_mul, zone_mul,peak_sep_time_max,min_span_time,file_name);

beg_p=startn;
end_p=finishn;

[output,k,ro,b_k]=baseline_sub_poly_mod4(p1,beg_p,end_p,n,max_ilevel, error);

output_baseline=output;
```

A.3 Iterative Polynomial Baseline Fit

```
function [output,k,ro,b_k]=baseline_sub_poly(y,beg_p,end_p,n,max_ilevel, error)
%baseline_sub_poly(y,beg_p,end_p,t,n,max_ilevel, error)is a function that will
%remove baseline variations using an iterative polynomial fit technique to remove the
%baseline variations.  y_in is the input signal, beg_p is beinging peak and
%end_p is the end peak location (in seconds), n is the index of polynomial,
%t is the length of the signal,max_ilevel is teh maximum number of
%iterations. error is the error factor used for comparisons.

warning off;
Fs=100;


y_in=y;

y_leng=length(y_in);

x=(1/Fs:1/Fs:y_leng/Fs)';


mm=length(beg_p);
nn=length(end_p);
if (mm~=nn)
    error('number of peak start and end does not equal!');
end;


%form the peak region
peak_reg=zeros(length(y),1); %initial peak_reg variable for storag
for q=1:mm,
    for j=beg_p(q):end_p(q);  %set peak_reg(j) to one where j is i the region of the peaks
        peak_reg(j)=1;
    end;
end;



    %Step 1 and 2.  Caculate the polynomial fitting equations
    ro=1;
    k=1;


    y_km1=y_in;
    b_km1=y_in;


    while((ro>error)&&(k<max_ilevel))


        [a,s]=polyfit(x,y_km1,n); %calculates polynomial approx
```

```
        b_k=polyval(a,x); %calculates approximated y

        %si(k)=s;  %What is si?


        y_k=y_km1;


            for i=1:length(y_in) %scan the length of y
                if (peak_reg(i)==1)  %check to see of i is in the peak region area
                    if (y_km1(i)>b_k(i))
                        y_k(i)=b_k(i);
                    end
                end
            end


        %Step 4, check for error criterion
        ro=norm(b_k-b_km1)/norm(b_km1);  %for the duration of the signal
%
        rok(k)=ro;

        b_km1=b_k;
        y_km1=y_k;
        k=k+1;


    end;


    output = y_in-b_k;



mean(rok);


end
```

A.4 Peak Region Detection

A.4.1 Peak Region Location

```
function [peak_region_start,peak_region_end,startn,finishn]
= peak_region_location(y,threshold_multiplier_fd2,
zone_threshold_multiplier,peak_sep_time_max,min_span_time,file_name) %
%This function will return the peak region start and the end by inputting
%the input

file_name=[file_name,'_der'];
y1=y;

%Initialisation
Ts=0.01;
Fs=1/Ts;
degree=5;
dx=Ts;
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Step 1: High Order Wavelet Transform

 wt_name='sym8';
 wt_level=8;

TPTR='heursure';
SORH='s';
SCAL='one';

y=wden(y,TPTR,SORH,SCAL, wt_level , wt_name);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Step 2: Calculate the derivatives and optimal Smoothing
%Optimal Savitzky Golay Smoothing.  The optimal seeks for the optimal span when
%for the nth degree polynomial SG smoothing
% [yout, span ,DW2]= optimal_savitzky(y, degree);
 y_original=y;

y_mean=mean(y);

%Take the derivatives of the signal and use optimal smooth every stage.
yp=derivative_cwt(y,'gaus1',16,dx,1);

%End taking derivatives


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Step 3: Threshold Calculations

 thresd_divider=300;
 thrshd_multiplier=300;
 thresy1=y_mean+calculate_threshold(y,0.01,5,thrshd_multiplier);
 thresy2=y_mean+(max(y)-y_mean)/thresd_divider; %max of y divide by a set constant

 thresfd=calculate_threshold(yp,0.01,5,10); %this function looks for the

 thresfd2=thrsd_calc(yp,threshold_multiplier_fd2);  %the constant used to be 5 for initla testing
 zone_threshold=thrsd_calc(yp,zone_threshold_multiplier);

 thressd=thrsd_calc(ypp,5);

 threstd=thrsd_calc(yppp,5);

[peak_start,peak_end,thr,new_start,new_finish,index,zone_exceeds_threshold]
=region_detect(yp,thresfd2,peak_sep_time_max,min_span_time,zone_threshold,y);

%Step 7: Variable Assignment

 peak_region_start=peak_start;
 peak_region_end=peak_end;


 startn=new_start;
 finishn=new_finish;


end
```

A.4.2 Region Detect

```
function [startn,finish,thr,new_start,new_finish,index,zone_exceeds_threshold].
=region_detect(yp,thresfd2,peak_sep_time_max,min_span_time,zone_threshold,y)
%This function will find the peak region coordinates in a signal.

Fs=100;
if nargin < 3,
    peak_sep_time_max=5;
    min_span_time=1;
    %peak_sep_time_max=5;
end

peak_sep_time=peak_sep_time_max*Fs;
min_span_width=min_span_time*Fs;
degree=5;
dx=0.01;

%initialization.
startn=0;
finish=0;
peak_start=0;
peak_end=0;

% thr=mean(yp)*2;
thr=0;
thr=zone_threshold;
%thr=thresfd2*2/3;

index=crossing(yp,[],thr); %find all the zero crossing in the 1st derivative
l=length(index);

if l==0;
    error('no crossing!');
end;

thr;
index2=crossing(yp,[],-thr); %find all the zero crossing in the 1stderivative
l2=length(index2);

if l2==0;
    error('no crossing!');
end;

%combine and sort all of the crossing coordinates
index=[index,index2];

index=remove_duplicate(index); %remove same elements
index=sort(index);
l=length(index);

zone_exceeds_threshold=zeros(1,l-1);
%This loop loops through all the change in coordinates and examines to see
%which cross the threshold.  +1 if it exceeds on the high end, -1 if on the
%low end, and 0 if it doesnt cross threshold on both ends.
for i=1:l-1,

if((max(yp(index(i)+1:index(i+1)))>thresfd2) && ((index(i+1)-index(i)+1)>min_span_width))
        zone_exceeds_threshold(i)=1;
    %else if (min(yp(index(i):index(i+1)))<=-abs(thresfd2))
    else if
((min(yp(index(i)+1:index(i+1)))<=-abs(thresfd2)) && ((index(i+1)-index(i)+1)>min_span_width))
            zone_exceeds_threshold(i)=-1;
        else zone_exceeds_threshold(i)=0;
```

```
        end;
    end;
end;

kkk=1;
zone_exceeds_threshold;

q=length(zone_exceeds_threshold);  %number of zones
%If there are any zones that not not above the threshold that's within x
%seconds.
for i=1:q-1,
    if (zone_exceeds_threshold(i)~=0)
        for j=i+1:q, %look for the next zone that crosses the threshold
            %found the next above threshold zone
            %added an extra condition where the sign have to be different between the zones.
            if (zone_exceeds_threshold(j)~=0)%found the next above threshold zone
                if ((index(j)-index(i+1))<peak_sep_time)  %if the time difference between
                    %the above threhsold zones
                )  %if the time difference between the above threhsold zones
                    for k=i+1:j-1, %change all zones inbetwen the zones
                        zone_exceeds_threshold(k)=sign(yp(index(k)));

                    end;
                    break;  %exit the j=i+1:q-1 for loop
                end;
            end;
        end;
    end;
end;

kkk=2;
zone_exceeds_threshold;


%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%This funciton looks for at least one change in the zones that
%%%%%%%%%exceeded the threshold.
zone_exceeds_threshold_temp=zone_changes_check(zone_exceeds_threshold);
%save zones.mat zone_exceeds_threshold zone_exceeds_threshold_temp;%
zone_exceeds_threshold=zone_exceeds_threshold_temp;
%%%%%%%%%%%%%%%%%%%%%%%%%

kkk=3;
zone_exceeds_threshold;

%This loop gets rid of the regions that didn't cross the threshol on the
%plus and negative side so taht information can be easily interpreted.
k=0;
threshold_flag=0;


for i=1:l-1,
    if zone_exceeds_threshold(i)~=0,
        k=k+1;
        startn(k)=index(i)+1;
        finish(k)=index(i+1);


        zone(k)=zone_exceeds_threshold(i);
        %this is used to store the number of zone transitions that is above the threshold
        threshold_flag=1;
    end;
```

```
end


%This part of the code get rid of overlapping peaks
new_start=startn;
new_finish=finish;


     ooo=1;
     new_start;
     new_finish;

w=seek_for_repeat(startn,finish); %w stores the indexes of repeats
new_start(w+1)=[];
new_finish(w)=[];


if(length(new_finish)~=0)

     ooo=2;
     new_start;
     new_finish;

end;

%%%%%%%%%%%%%%%%%%%%This section checks for the real peak region
%min_span_time2=1*Fs;  %I think 3 seconds is good for determination of a
%real peak region
temp1=[];
temp2=[];
k=1;
l=length(new_start);
for i=1:l,
     if((new_finish(i)-new_start(i))>min_span_width*3)
          %finish(i)-new_start(i)
          temp1(k)=new_start(i);
          temp2(k)=new_finish(i);
          k=k+1;
     end;

end;


new_start=temp1;
new_finish=temp2;

     ooo=3;
     new_start;
     new_finish;
% %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% % % % % %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%This section of the code checks for
% peak average in each section of the loop.
% % % % % %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%region
temp1=[];
temp2=[];
k=1;
l=length(new_start);
ym=median(y)-.5*std(y);
%ym=median(y)
```

```
for i=1:l,
    median(y(new_start(i):new_finish(i)));
    if(median(y(new_start(i):new_finish(i)))>=ym)
        temp1(k)=new_start(i);
        temp2(k)=new_finish(i);
        k=k+1;
    end;
end;

new_start=temp1;
new_finish=temp2;

    ooo=4;
    new_start;
    new_finish;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%%%%%%%%%%%%%This section of the code gets sets startn to 1 and finish to
%%%%%%%%%%%%%%end of the signal's legnth when there are no peaks detected
startn;
finish;

if ((length(new_start)==0) || ((length(new_start)==1)
   && ((new_start==0) || (new_finish==0))))
   %if there are no peak regions detected.
  % Then use the length of the signal as the peak region

    new_start=1;
    new_finish=length(yp);
end;
```

A.5 Jansson's Deconvolution

```
function [out,ro,rms,peak_num]=jan_deconv_mod6_backup(y,h1,b,k_max,delta)
%jan_deconv(y,h1,a,b,c,k_max,shift)
%This function calculates the deconvolution of y with h1 by using Jansson's
%method.  a is the scaler multiple of h1, b is   ... c is critical and
%affects the performance, k_max is the maximum interations, and shift
%number in the alogrithm.

y_orig=y/max(y);
g=y; %preserve the captured image
h=(h1/max(h1)); %normalize to 1;
%h=h1;
scale=1/max(y);
y=y/max(y); % normalize captured image


shift=1;


    f=h(end:-1:1); %%%Does this equal to f(th(-t)?
    hm=conv(h,f);
    h=shift_conv(hm,h,f,shift);
    gm=conv(g,f);
    g=shift_conv(gm,h,g,-286); %286 random number based on the PSF i believe.
    g=g/max(g);
    h=h/max(h);
```

```
x_old=g; %step one.  Initial estimate
y=g;
% a=trapz(y)/trapz(x_old)
%a=10;

%step two
%for k=1:k_max,
    k=1;
    c=max(x_old);
    peak_num=zeros(1,k_max-1);
    ro=zeros(1,k_max-1);
    rms=zeros(1,k_max-1);
    [c_max,c_min]=peakdet(y_orig,delta);
    %figure; plot(y_orig); vline(c_max);
%while((ro>error)&&(k<k_max))



while(k<k_max)

    a=0.0025;
    temp=conv((a*h),x_old);

    temp=shift_conv_mod2(temp,h,x_old,shift+50);
    r2=relax(x_old,c,b);


    temp2=r2.*(y-temp);
    x_new=x_old+temp2;



    for i=1:length(x_new), %step four
        if (x_new(i)<0)
            x_new(i)=0;
            %x_new(i)=abs(x_new(i));
        end
    end



    %%%%%%%%%%%%Error between the current iteration and the origianl inpout
     ro(k)=norm(x_old-x_new)/norm(x_old);
    rms(k)=rmse(x_old,x_new);


    [c_max,c_min]=peakdet(x_old,delta);
    [d_max,d_min]=peakdet(x_new,delta);


    d_max(:,1);
    iter=['x_new at iteration ',num2str(k)];

    rok=ro(k);
    rmsk=rms(k);


    length(c_max);
    length(d_max);
    if((k>3) && (length(c_max)~=length(d_max) || (ro(k)<5E-3) ))
% % % % % % % % %          figure; plot(x_new); vline(d_max);
```

```
        x_new=x_old;
        break;
    end;



    peak_num(k)=length(d_max(:,1))-length(c_max(:,1));
    x_old=x_new;
    k=k+1;


end;
```

A.6 Wavelet Denoising for μTK

```
function waveletdenoiseuTK
%This function will be used to perform wavelet denoising on uTK data
%Prompt the user to select .txt files from uTK

[file,dir]=uigetfile('*.txt');

disp('Processing... Please wait')
filename_loc=strcat(dir,file);

%Delete the header rows  %this step is optional and would be ideal if it is
%possible
[PMT,t]=parse_utk2(filename_loc);

%Perform wavelet transform
name='sym8';
level=8;
tptr='heursure';
sorh='s';
scal='one';
denoise=wden(PMT,tptr,sorh,scal,level,name);

%Plot and save the data.
    plot(t,PMT);
    xlabel('Time(s)');
    ylabel('PMT (V)');
    title('uTK Raw Signal');
    saveas(gcf,strcat(filename_loc,'_raw.jpg'),'jpg');

    figure;
    plot(t,denoise);
    xlabel('Time(s)');
    ylabel('PMT (V)');
    title('Denoised Signal');

    saveas(gcf,strcat(filename_loc,'_denoised.jpg'),'jpg');

    fid2= fopen([filename_loc,'_denoised.txt'],'w');
    fprintf(fid2,'Time(s) \t PMT Raw (V) \t PMT Denoised (V) \n');
    p=[t;PMT;denoise];
    fprintf(fid2,'%f \t %fi \t %f \n',p);
    fclose('all');

disp('Processing completed')
```

```
function [PMT,t]=parse_utk(filename)

fid=fopen(filename) ; % the original file

% read column headers
C_text = textscan(fid, '%s', 30, 'delimiter', '|');

% read numeric data
C_data0 = textscan(fid, '%f %f %f %f %f %f %f %f %f %f %f %f %f %f %f %f %f %f %f %f %f %f %f %f %f %f %f %f');
fclose(fid);

t1=C_data0{1,25}(1:end,1);
PMT1=C_data0{1,26}(1:end,1);

i=find(~isnan(PMT1));

PMT1 = PMT1(~isnan(PMT1));
t1 = t1(~isnan(t1));

k=length(t1)+4;
t2=C_data0{1,1}(k:end,1);
PMT2=C_data0{1,2}(k:end,1);


PMT=[PMT1', PMT2'];
t=[t1',t2'];

%look for 5 consectutive zeros
z=zeros(5,1)';
index = findstr(PMT,z);
start=index(1)+50;

PMT=PMT(start:end);
t=t(start:end);
```

A.7 Executable Program User Guide

- Install the MATLAB Compiler Runtime (MCR), MCRInstaller.exe. This program contains the MATLAB default libraries and system files. This is required so that MATLAB compiled executable program could be run on computers without having MATLAB installed.
- After the MCR is installed, open TTKSignalProcessing.exe.
- Select the TTK output text file you wish to process.
- A command prompt window will pop up and while it is processing the signal, it will say “Processing... Please wait.” Wait for “Processing is completed” to proceed.

- A figure will appear once processing is done. Showing stages of processing: raw, wavelet denoised, baseline removed, and then peaks separated signals. This figure is automatically saved in the folder where the raw text file is located.
- Two text files are saved. One contains the post processing signal and the other contains the peak locations.

Appendix B

Iterative Polynomial Baseline Fit Pseudo-code

1. In the first iteration, set the signal estimate $y_1(t)$ and the baseline estimate $b_1(t)$ to the original measurement, $y(t)$.

2. Set the iteration count k to 2.

3. Calculate the n_{th} order polynomial coefficients p_n for $y_{k-1}(t)$. The polynomial coefficients are calculated using the MATLAB built-in function *polyfit*.

4. Construct the baseline estimation $b_k(t)$ with p_n using equation (B.1).

$$b_k(t) = \sum_{i=0}^{n} p_{i+1} t^{n-i}, \tag{B.1}$$

5. Update the estimated signal by setting $y_k(t)$ to $y_{k-1}(t)$, where $y_k(t)$ and $y_{k-1}(t)$ are the signal estimate in the k_{th} and $k-1_{th}$ iterations.

6. Remove the peaks by comparing $b_k(t)$ with $y_{k-1}(t)$ in the regions of signal peaks; if $y_{k-1}(i) > b_k(i)$, then set $y_k(i)$=$b_k(i)$, for $i = 1,2,...n$ where i is in the region where $y_{k-1}(t) > b_k(t)$.

7. Calculate the error factor between current and previous iterations of the baseline estimation, denoted by ρ and defined by equation (B.2). $b_k(t)$ and $b_{k-1}(t)$ are the baseline estimate with polynomials at k_{th} and $k-1_{th}$ iterations, respectively.

$$\rho = \frac{||b_k - b_{k-1}||}{b_{k-1}} \tag{B.2}$$

8. If ρ is smaller than a prespecified error factor threshold ρ_{thr} or k reaches k_{max}, then the baseline estimation is sufficient; otherwise, set $b_{k-1}(t)$ to $b_k(t)$, $y_{k-1}(t)$ to $y_k(t)$, and increment k. Return to Step 3.

Appendix C

PSPICE Circuit Simulation

```
OPMODEL1.CIR – OPAMP MODEL SINGLE–POLE
*
IS   1    0    AC   1   PWL(0US 0V   0.01US 1V 1US 1V 1.01US 0V)
Cin  1  0  8PF
XOP  0    1    3    OPAMP1
RF   1  3        10MEG
CF  1  3         100PF

XOP2      3    4    5    OPAMP1
R1   4    0    1K
R2   4    5    100K
C2   4    5  1PF

RL  5  0  10

*
* OPAMP MACRO MODEL, SINGLE–POLE
* connections:       non–inverting input
*                    |    inverting input
*                    |    |    output
*                    |    |    |
.SUBCKT OPAMP1       1    2    6
* INPUT IMPEDANCE
RIN  1    2    10MEG
* DC GAIN=100K AND POLE1=100HZ
* UNITY GAIN = DCGAIN X POLE1 = 10MHZ
EGAIN     3  0  1  2  1MEG
RP1  3    4    100K
CP1  4    0    1.5915UF
* OUTPUT BUFFER AND RESISTANCE
EBUFFER  5  0  4  0  1
ROUT      5    6    10
.ENDS

* ANALYSIS
.AC       DEC       5  1  10MEG
.TRAN     0.1US   1MS
.PLOT     AC   VM(3)
.PROBE
.END
```

Appendix D

Point Spread Function Modelling

D.1 Procedure

An approximation of the point spread function $h(t)$ can be obtained by curve fitting an empirically transformed Gaussian (ETG) function to a high-SNR experimental single peak signal. h(t) for a a CE system can be obtained by the following:

1. Obtain single-size sample. i.e/ 1 μL Cy-5 reversed primer (end-labelled DNA).

2. Perform standard CE with AML's CE protocol [94]: 1 μL of 4% linear polyacrylamide (LPA) sieving matrix and a 3 μL 0.01x Tris TAPS-EDTA (TTE) buffer were used to fill the microfluidic chip channels. First inject sample with a voltage of 200 V (or equivalent to a electric field of 222 V/cm) for 80 seconds, then separate the sample by a voltage of 600 V (67 V/cm) for 250 seconds. Detection was made 13 mm from the CE channel intersection. Make sure the signal has high SNR ($\geq$ 50 V/V).

3. Remove noise with sym8 wavelet.

4. Normalize the denoised signal to 1 by dividing the signal by its maximum value.

5. Fit curve with ETG. ETG functions are defined by equation (D.1),

$$h(t) = \frac{2He^{0.5}}{(1+\lambda_l e^{k_l*(t_l-t)})^{\alpha} + (1+\lambda_t e^{k_t*(t_t-t)})^{\beta} - 1}. \tag{D.1}$$

H is the maximum peak height, and because the signal is already normalized, set to 1. t_t and t_l are the half width times for the leading and trailing edge. Increase t_t and t_l to increase the length of the leading and trailing edge, respectively. k, λ, α and β can be used to adjust symmetrical properties of the peak. These parameters were manually tuned to fit the experimental signal peak's leading and trailing edge.

D.2 MATLAB Code For Generating PSF

```
function [output]=ETG(H,lamdal,kl,tl,alpha,lamdat,kt,tt,beta,time,Fs)

Ts=1/Fs;
t=Ts:Ts:time;

Hbb=2*H*exp(0.5);
a=(1+lamdal*exp((kl*(tl-t)))).^alpha;
b=(1+lamdat*exp((kt*(t-tt)))).^beta;
output=Hbb./(a+b-1);
```