Algorithm-circuit co-design for detecting symptomatic patterns in biological signals

Himanshu Markandeya

Purdue University

Follow this and additional works at: http://docs.lib.purdue.edu/open_access_dissertations

Part of the Biomedical Engineering and Bioengineering Commons, and the Electrical and Computer Engineering Commons

Recommended Citation

http://docs.lib.purdue.edu/open_access_dissertations/330

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Himanshu Markandeya

Entitled
Algorithm-circuit Co-design for Detecting Symptomatic Patterns in Biological Signals

For the degree of Doctor of Philosophy

Is approved by the final examining committee:
KAUSHIK ROY
ANAND RAGHUNATHAN
BYUNGHOO JUNG
PEDRO IRAZOQUI

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification/Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University’s “Policy on Integrity in Research” and the use of copyrighted material.

KAUSHIK ROY

Approved by Major Professor(s): ________________________________

______________________________

Approved by: M. R. Melloch 12/11/2014

Head of the Department Graduate Program Date
ALGORITHM-CIRCUIT CO-DESIGN

FOR

DETECTING SYMPTOMATIC PATTERNS IN BIOLOGICAL SIGNALS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Himanshu Markandeya

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2014

Purdue University

West Lafayette, Indiana
This thesis is dedicated to my loving parents, Maya and Suhas Markandeya.
ACKNOWLEDGMENTS

I want to thank all the people who have contributed in various ways to my research presented in this dissertation. Foremost is my deep appreciation and gratitude to my advisor Prof. Kaushik Roy. His invaluable guidance and patience throughout the course of my doctoral studies have been a major factor in the conduct of this research work. I would forever be grateful for the advice and constant encouragement, I have received from him. I would also like to thank my committee member, Prof. Pedro Irazoqui. He gave me an opportunity to work with his research group and collaborate with them. I would also like to acknowledge my PhD committee members, Prof. Anand Raghunathan and Prof. Byunghoo Jung for their experience and time; effort and helping me focus on my dissertation.

I would like to thank my friend and collaborator from the Department of Biomedical Engineering, Dr. Shriram Raghunathan. It was a very pleasant and educationally rewarding experience to work with him. I would also like to thank Dr. Georgios Panagopoulos and Dr. Sang Phill Park, for sharing their knowledge with me and helping me with numerous technical issues. I am fortunate to be a part of a laboratory consisting of diverse and intelligent individuals. I would like to thank all my colleagues from the Nanoelectronics Research Laboratory group and my friends at Purdue University for encouraging me and often having intellectually stimulating discussions. I also want to show my appreciation to Rwitti Roy, who made us all feel at home in Purdue University.

Finally, I would like to thank my parents, Maya and Suhas Markandeya. There unconditional love, support always helped me get back on my feet from difficult situations. Thank you for being the pillars of strength during this huge endeavor.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>ABBREVIATIONS</td>
<td>x</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.1 Epileptic Seizure Detection: An Implantable Application</td>
</tr>
<tr>
<td></td>
<td>1.1.1 Epilepsy – Background</td>
</tr>
<tr>
<td></td>
<td>1.1.2 Seizure Detection – Literature Survey</td>
</tr>
<tr>
<td></td>
<td>1.2 Health monitoring system: A Wearable Application</td>
</tr>
<tr>
<td>2 DISCRETE WAVELET TRANSFORM AND QUASI-AVERAGE BASED ALGORITHM FOR SEIZURE DETECTION</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>2.1 Discrete Wavelet Transform: Background</td>
</tr>
<tr>
<td></td>
<td>2.2 Design Methodology</td>
</tr>
<tr>
<td></td>
<td>2.2.1 Neural Data Recording</td>
</tr>
<tr>
<td></td>
<td>2.2.2 Algorithm</td>
</tr>
<tr>
<td></td>
<td>2.2.3 Training</td>
</tr>
<tr>
<td></td>
<td>2.2.4 Implementation</td>
</tr>
<tr>
<td></td>
<td>2.3 Algorithm and Hardware Implementation</td>
</tr>
<tr>
<td></td>
<td>2.3.1 DWT-QA Algorithm</td>
</tr>
<tr>
<td></td>
<td>2.3.2 Training for User-Specific Operation</td>
</tr>
<tr>
<td></td>
<td>2.3.3 Architecture Mapping</td>
</tr>
<tr>
<td></td>
<td>2.4 Efficacy and Efficiency Results</td>
</tr>
<tr>
<td></td>
<td>2.4.1 Algorithmic Efficacy</td>
</tr>
<tr>
<td></td>
<td>2.4.2 Hardware Efficiency</td>
</tr>
</tbody>
</table>
3 EPILEPTIC SEIZURE DETECTION PROCESSOR WITH MULTI-ALGORITHM PROGRAMMABILITY

3.1 System Overview

3.1.1 Algorithms

3.1.2 Boolean Logical Combination

3.1.3 Training Phase

3.2 Hardware Implementation

3.2.1 Algorithm Bank

3.2.2 Threshold Bank

3.2.3 Boolean Logic Selector

3.2.4 Built-in Self Test (BIST)

3.3 Efficacy and Efficiency Results

3.4 Conclusion

4 DETECTION OF EPILEPTIC SEIZURE USING MULTIPLE(2) STAGES

4.1 Overview

4.1.1 Algorithm

4.1.2 Training

4.1.3 Near-Threshold Voltage Operation

4.2 Hardware Implementation

4.2.1 Monitoring Stage-Coastline Parameter Algorithm

4.2.2 Detection Stage-DWT-QA Algorithm

4.2.3 Controller

4.3 Efficacy and Efficiency Metrics

4.4 Results

4.4.1 Algorithm Efficacy with minimal detection latency

4.4.2 Hardware Energy Efficiency

4.5 Conclusion
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Coefficient Representation using Pre-computation</td>
<td>24</td>
</tr>
<tr>
<td>2.2 Original and Modified Filter Coefficients</td>
<td>26</td>
</tr>
<tr>
<td>2.3 Algorithm Efficacy for 5 Different Animals with Daubechies 2\textsuperscript{nd}, 4\textsuperscript{th} and 6\textsuperscript{th} order wavelets</td>
<td>32</td>
</tr>
<tr>
<td>2.4 Power Dissipation and Area Improvement using IBM 90 nm bulk-Si Technology</td>
<td>33</td>
</tr>
<tr>
<td>3.1 Measured Results for Power</td>
<td>51</td>
</tr>
<tr>
<td>4.1 Hardware Efficiency for Typical Test Data</td>
<td>68</td>
</tr>
<tr>
<td>5.1 Classification accuracy for five acoustic symptomatic patterns</td>
<td>90</td>
</tr>
<tr>
<td>5.2 Power and Area of Hardware Implementation</td>
<td>91</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Towards implantable medical solutions</td>
<td>3</td>
</tr>
<tr>
<td>1.2</td>
<td>Towards wearable health monitoring solutions</td>
<td>6</td>
</tr>
<tr>
<td>2.1</td>
<td>Wave vs. Wavelet</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>Time-Frequency Resolution of STFT and DWT</td>
<td>11</td>
</tr>
<tr>
<td>2.3</td>
<td>Mallat Algorithm</td>
<td>12</td>
</tr>
<tr>
<td>2.4</td>
<td>Daubechies 4&lt;sup&gt;th&lt;/sup&gt; order scaling and wavelet function</td>
<td>13</td>
</tr>
<tr>
<td>2.5</td>
<td>Comparison of High Pass Filter response for Daubechies family of wavelets</td>
<td>14</td>
</tr>
<tr>
<td>2.6</td>
<td>Design Methodology for Epileptic Seizure Detection System</td>
<td>15</td>
</tr>
<tr>
<td>2.7</td>
<td>Six-Level DWT decomposition of a neural signal (LFP)</td>
<td>21</td>
</tr>
<tr>
<td>2.8</td>
<td>Seizure Detection Using DWT-QA Algorithm</td>
<td>21</td>
</tr>
<tr>
<td>2.9</td>
<td>Block Diagram of System Architecture</td>
<td>23</td>
</tr>
<tr>
<td>2.10</td>
<td>Implementation of Select/Shift Add Block (SSA)</td>
<td>25</td>
</tr>
<tr>
<td>2.11</td>
<td>Comparison of Filter Response for Modified filter coefficients</td>
<td>27</td>
</tr>
<tr>
<td>2.12</td>
<td>Block Diagram for Quasi-Averaging Technique</td>
<td>29</td>
</tr>
<tr>
<td>2.13</td>
<td>LFP Neural Recording System Implanted in a live Rat</td>
<td>31</td>
</tr>
<tr>
<td>3.1</td>
<td>Seizure Detection with independent algorithm and Boolean combination</td>
<td>41</td>
</tr>
<tr>
<td>3.2</td>
<td>Block Diagram of Seizure Detection Processor</td>
<td>43</td>
</tr>
<tr>
<td>3.3</td>
<td>Block Diagram of Algorithms in Algorithm Bank</td>
<td>44</td>
</tr>
<tr>
<td>3.4</td>
<td>Modes of operation of the multi-algorithm processor</td>
<td>46</td>
</tr>
<tr>
<td>3.5</td>
<td>Die Photo of Seizure Detection Processor</td>
<td>48</td>
</tr>
<tr>
<td>3.6</td>
<td>System Test Setup</td>
<td>48</td>
</tr>
<tr>
<td>3.7</td>
<td>Efficacy of the Algorithms for Large Animal Study</td>
<td>50</td>
</tr>
<tr>
<td>3.8</td>
<td>Average Efficacy vs. Normalized Hardware Cost</td>
<td>50</td>
</tr>
<tr>
<td>4.1</td>
<td>Algorithm Methodology and Flow Diagram</td>
<td>57</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>4.2</td>
<td>Typical Plot to demonstrate the effect of $V_{DD}$ scaling on Energy dissipation</td>
<td>60</td>
</tr>
<tr>
<td>4.3</td>
<td>Block Diagram of Two Stage Algorithm</td>
<td>61</td>
</tr>
<tr>
<td>4.4</td>
<td>Detection of Epileptic Seizure with marked FP FN and Seizure</td>
<td>64</td>
</tr>
<tr>
<td>4.5</td>
<td>Algorithmic Efficacy for the Two-Stage Algorithm and its comparison with individual stages and logical combination</td>
<td>67</td>
</tr>
<tr>
<td>5.1</td>
<td>Algorithm methodology for pattern detection in acoustic non-speech human signals</td>
<td>72</td>
</tr>
<tr>
<td>5.2</td>
<td>Algorithm flow for symptom detection in audio biological signals</td>
<td>73</td>
</tr>
<tr>
<td>5.3</td>
<td>Acoustic signals used for symptomatic pattern detection</td>
<td>73</td>
</tr>
<tr>
<td>5.4</td>
<td>Wavelet decomposition of ‘Wheeze’ acoustic signal</td>
<td>75</td>
</tr>
<tr>
<td>5.5</td>
<td>Continuous increase in amplitude in ‘Vomit’ pattern</td>
<td>76</td>
</tr>
<tr>
<td>5.6</td>
<td>Frequency spectrum for typical cough and sneeze signals</td>
<td>78</td>
</tr>
<tr>
<td>5.7</td>
<td>Conversion between Mel scale and Frequency scale</td>
<td>79</td>
</tr>
<tr>
<td>5.8</td>
<td>Block diagram for standard MFCC algorithm</td>
<td>80</td>
</tr>
<tr>
<td>5.9</td>
<td>Frequency response of Mel filter bank</td>
<td>81</td>
</tr>
<tr>
<td>5.10</td>
<td>Block diagram for modified MFCC algorithm</td>
<td>82</td>
</tr>
<tr>
<td>5.11</td>
<td>Mapping of the $D_3$ wavelet coefficient on the Mel filter bank</td>
<td>83</td>
</tr>
<tr>
<td>5.12</td>
<td>Block diagram for symptomatic pattern detection in acoustic non-speech signals</td>
<td>85</td>
</tr>
<tr>
<td>5.13</td>
<td>Symptomatic pattern detection for Sneeze signal</td>
<td>89</td>
</tr>
<tr>
<td>5.14</td>
<td>Symptomatic pattern detection for Cough signal</td>
<td>89</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>VLSI</td>
<td>Very Large Scale Integration</td>
<td></td>
</tr>
<tr>
<td>AED</td>
<td>Anti-Epileptic Drugs</td>
<td></td>
</tr>
<tr>
<td>VNS</td>
<td>Vagus Nerve Stimulation</td>
<td></td>
</tr>
<tr>
<td>FDA</td>
<td>Food and Drug Administration</td>
<td></td>
</tr>
<tr>
<td>EEG</td>
<td>Electroencephalogram</td>
<td></td>
</tr>
<tr>
<td>ECoG</td>
<td>Electrocorticogram</td>
<td></td>
</tr>
<tr>
<td>LFP</td>
<td>Local Field Potential</td>
<td></td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
<td></td>
</tr>
<tr>
<td>STFT</td>
<td>Short-Time Fourier Transform</td>
<td></td>
</tr>
<tr>
<td>DWT</td>
<td>Discrete Wavelet Transform</td>
<td></td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
<td></td>
</tr>
<tr>
<td>KiMs</td>
<td>Kids Health Monitoring System</td>
<td></td>
</tr>
<tr>
<td>TSMC</td>
<td>Taiwan Semiconductor Manufacturing Company</td>
<td></td>
</tr>
<tr>
<td>LAS</td>
<td>Large Animal Study</td>
<td></td>
</tr>
<tr>
<td>DB</td>
<td>Daubechies</td>
<td></td>
</tr>
<tr>
<td>QA</td>
<td>Quasi-Averaging</td>
<td></td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative Distribution Function</td>
<td></td>
</tr>
<tr>
<td>FIR</td>
<td>Finite Impulse Response</td>
<td></td>
</tr>
<tr>
<td>CSHM</td>
<td>Computation Sharing Multiplier</td>
<td></td>
</tr>
<tr>
<td>CSE</td>
<td>Common Sub-expression Elimination</td>
<td></td>
</tr>
<tr>
<td>SSA</td>
<td>Select/Shift and Adder</td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>Adder Level</td>
<td></td>
</tr>
<tr>
<td>LPF</td>
<td>Low-Pass Filter</td>
<td></td>
</tr>
<tr>
<td>HPF</td>
<td>High Pass Filter</td>
<td></td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------</td>
<td></td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
<td></td>
</tr>
<tr>
<td>FN</td>
<td>False Negative</td>
<td></td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
<td></td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
<td></td>
</tr>
<tr>
<td>ADR</td>
<td>Average Detection Rate</td>
<td></td>
</tr>
<tr>
<td>BIST</td>
<td>Built-In Self Test</td>
<td></td>
</tr>
<tr>
<td>CL</td>
<td>Coastline</td>
<td></td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel Frequency Cepstrum Coefficient</td>
<td></td>
</tr>
<tr>
<td>DCT</td>
<td>Discrete Cosine Transform</td>
<td></td>
</tr>
<tr>
<td>DC</td>
<td>Direct Current</td>
<td></td>
</tr>
<tr>
<td>VHSIC</td>
<td>Very High Speed Integrated Circuit</td>
<td></td>
</tr>
<tr>
<td>VHDL</td>
<td>VHSIC Hardware Description Language</td>
<td></td>
</tr>
</tbody>
</table>
ABSTRACT

Markandeya, Himanshu PhD, Purdue University, December 2014. Algorithm-Circuit Co-design for Detecting Symptomatic Patterns in Biological Signals. Major Professor: Kaushik Roy.

The advancement in scaled Silicon technology has accelerated the development of a wide range of applications in various fields including medical technology. It has immensely contributed to finding solutions for monitoring general health as well as alleviating intractable disorders in the form of implantable and wearable systems. This necessitates the development of energy efficient and functionally efficacious systems. This thesis has explored the algorithm-circuit co-design approach for developing an energy efficient epileptic seizure detection processor which could be used for implantable epilepsy prosthesis. Novel wavelet transform based algorithms are proposed for accurate detection of epileptic seizures. Energy efficient techniques at circuit level such as power and clock gating are utilized along with error resiliency at algorithm level to implement these algorithms in TSMC 65nm bulk-Si technology. Furthermore, the methodology is extended to develop a generic pattern detection system, which could be used for health monitoring. The wavelet transform along with mathematical metrics and Mel cepstrum are used to develop an algorithm which can detect generic patterns in biological audio signals. The application of algorithm-circuit co-design methodology helps in practically implementing this system into a low power design. Using approximation of coefficients and multiplier-less implementation, the Mel cepstrum algorithm is modified to optimize the hardware cost without losing its functional efficacy. The system is user-specific and scalable for detecting various patterns in biological signals. The methodologies mentioned in this thesis are intended
towards development of user-scalable, energy efficient and highly efficacious systems for detection of patterns in variety of biological signals.
1. INTRODUCTION

Silicon technology has evolved over the past few decades into being a prominent part of human life. The advancements made in the field of Very Large Scale Integration (VLSI) technology have resulted in development of novel engineering applications in numerous fields such as computers, communication and even medical science. Technology scaling has enabled the integration of an ever increasing number of transistors on a single silicon chip. This has resulted in the design of systems and applications of very high complexity. The domain of medical science has benefited immensely with these systems. VLSI technology has enabled development of systems which can be used not only for health monitoring but also to alleviate certain intractable disorders. These systems may be wearable or implantable on human body. However, the high complexity of such a system often requires significant amount of computation. This translates to power consumption of insurmountable proportions. Furthermore, scaling of devices into the nanometer regime has brought forth numerous design challenges in the form of short channel effects. As the devices scale down in size, the leakage power due to sub-threshold current becomes a dominant component of the total power consumed. This is undesirable and a cause of concern, especially, in the design of biomedical applications. In health monitoring systems, such a high power consumption will drain the battery in shorter period of time and eventually cause a functional failure. In implantable biomedical systems, this is even more undesirable due to critical nature of operation. In such applications, the power consumption of a system is an important factor in not only determining its design but also its practical feasibility. This is because these implantable systems are powered by a limited energy source (implanted battery) and have to ensure reliable operation to avoid any catastrophic failure. Hence, the design of any biomedical implant has stringent constraints on the power consumption. Increased durability of the implant ensures reduced frequency
of replacement of the energy source, which might be expensive and require intrusive surgery. Apart from prolonging the battery life, it also ensures that the on-chip temperatures are at acceptable levels in order to prevent any damage to biological tissue surrounding the implant. This thesis addresses the feasibility of a system design for a wearable or implantable biomedical application. Furthermore, a generic scalable system is also developed to be used in wearable health monitoring products. The thesis intends to show that using simultaneous optimization at algorithm and circuit levels of design abstraction, it is possible to develop highly efficacious and low-power implantable and wearable biomedical systems. Such algorithm-circuit co-design techniques not only provide with optimal functionality but also pragmatic feasibility.

1.1 Epileptic Seizure Detection: An Implantable Application

In this research, “Epileptic Seizure Detection” has been targeted as an implantable biomedical application. In order to better understand the nature of the problem, this section briefly discusses the nature of the disorder and possible remedies available currently.

1.1.1 Epilepsy – Background

Epilepsy is a chronic neurological disorder affecting about 50 million people worldwide [1]. It is the most common disorder next only to Alzheimer's disease and stroke [2]. It is characterized by synchronous firing of a group of neurons in a section or the whole of brain tissue. This manifests into physical convulsions which are known as epileptic seizures or “fits”. However, unlike other medical disorders, epilepsy can only be controlled using medications but it cannot be cured. Even the best available pharmaceutical therapies can only help to reduce the frequency of occurrence of the seizures. However, about 30% of the epileptics do not respond positively to these pharmacological treatments. These treatments include Anti-Epileptic Drugs
(AED), which suppress neural activity and resective surgery, wherein a section of the identified brain tissue is surgically removed [1], [2].

Epilepsy can be classified into various types depending on the cause, location of the origin in brain, external stimulus (inducing seizures) etc. In this research, the type of epilepsy under consideration is ‘Focal Epilepsy’, wherein the seizure originates from a small section in the brain. The point of origin is the epileptogenic focus. Research has shown that neurostimulation at the epileptogenic focus in terms of chemical, electrical or optical intervention has emerged to be a promising therapy for the set of medically refractory patients, who do not respond well to available therapies [2], [3]. Neurostimulation can either be ‘continuous’ or ‘responsive’. ‘Continuous neurostimulation’ stimulates the epileptogenic focus irrespective of the occurrence or knowledge of an impending seizure. On the other hand, ‘responsive neurostimulation’ administers the therapy only if a possible onset of seizure is detected. Currently, the Vagus Nerve Stimulator (VNS therapy) is the only anti-epilepsy prosthesis which is commercially available and approved by the Food and Drug Administration (FDA) [4], [5]. VNS is based on the modulation of seizure activity by continuous electrical stimulation. It shows varying levels of efficacy among the users [3]. However, it has

Fig. 1.1. Towards implantable medical solutions
been debated that continuous neurostimulation is detrimental to the neuronal tissue in the long run [6]. It has also been shown, that after a certain period of time, the neurons adapt to the pattern of stimulation rendering the therapy ineffective [6]. Comparatively, responsive stimulation, which delivers focal therapy in response to the onset of a seizure, has been shown to be more effective [7], [8]. The Neuropace device developed by Medtronic Inc. is the only known responsive neurostimulation based epilepsy prosthesis and is under review by FDA [8], [9].

An effective responsive therapy necessitates the development of an efficacious technique for detecting the onset of seizure. This, coupled with the idea of providing implantable anti-epilepsy prosthesis and hence enabling the mobility of the patient, has resulted in great interest in this field. However, in order to detect the onset of seizure with a high efficacy, it is necessary to extensively process the neural data. This significantly increases the computational load on the detection system. This is directly translated to the amount of power consumed. Scaling of silicon technology makes it feasible to implement complex seizure detection systems. However, the battery technology has not scaled at the same rate. Hence an increased power consumption of an implantable system would drain the implanted battery heavily and affects the its longevity. To that effect, it is increasingly important to develop efficient algorithm-circuit co-design strategies to detect the onset of epileptic seizures.

1.1.2 Seizure Detection – Literature Survey

Decades of research has resulted in design of numerous algorithms for detecting epileptic seizures [10], [11]. These algorithms promise high efficacy of detection. However, majority of the algorithms in the literature utilize complex mathematical or statistical techniques which make it too expensive to implement in hardware. This is due to the stringent power constrains in the design as mentioned previously [12], [13]. Any modifications to these algorithms, in order to make them pragmatic from implementation perspective, could affect their detection efficacy adversely. Moreover, these
algorithms are based on scalp recordings, viz. Electroencephalogram (EEG) or Electrocorticogram (ECoG) [12] – [16]. This implies that by the time the onset of seizure is detected, it has already spread through a majority of the region of the brain. Apart from that, due to non-linear and time-variant nature of the neural signals, detection of the seizure onset using EEG requires significant amount of redundant signal processing. This increases the power consumption of the system and introduces a latency of detection. In this research, the neural signals are recorded in terms of Local Field Potentials (LFP). LFP is recording of the neuronal electrical activity from a small volume of brain tissue [17]. This makes it feasible to record very close to the origin of the seizure and hence reduce the detection latency. Furthermore, since the recorded signal is not modulated by the neural signals from other regions of the brain, the complexity of the required signal processing is expected to reduce significantly.

Clinically, an epileptic seizure is characterized by a gradual surge in the amplitude of the signal in specific frequency bands. This fact is exploited by existing algorithms, which analyze the Fast Fourier Transform (FFT) or the Short-time Fourier Transform (STFT) of the recorded neural signals [13], [14]. However, FFT based algorithms are not useful for implementing a system when data is streaming in continuously. This is due to the loss of temporal information in the FFT. STFT alleviates this by windowing the data into smaller sections. However, the resolution of the signal into its component frequency is not that accurate. It should also be noted that the FFT engine is computationally intensive and may require power hungry architectures [15]. Any modifications made to the architecture of FFT or STFT algorithm may change the efficacy of the system in terms of increased false detections, occurring due to “seizure-like” artifacts. An increased false detection results in redundant administration of the therapy and a wastage of the limited energy resource. This is detrimental to the neuronal environment as well. Some existing work uses the Discrete Wavelet Transform (DWT) to process signal efficiently and detect the onset of seizure [13], [18], [19]. However, most DWT based algorithms suggest the use of an artificial neural network (ANN) for subsequent processing [20]. This results in
a power hungry architecture which shadows the high efficacy of the algorithm due to lack of feasible implementation for an implantable prosthesis. There are other algorithms proposed, which are based on detection of spikes present in the neural signal [17], [21], [22]. The efficacy of these algorithms is debatable over a wider range of patients. This is because of inter-patient variability of the neural signals.

1.2 Health monitoring system: A Wearable Application

In addition to implantable biomedical application, rapid advancements in low-power VLSI methodologies have also spawned numerous types of wearable health monitoring systems. These systems record and analyze various bio markers such as temperature, blood-pressure, heart rate etc. Abnormalities in these parameters are a precursor to an impending health issue. These systems are also used to monitor patients in rehabilitation. Due to wearable nature of this system, it does not impede the mobility of the user. However, these system also have to depend on a limited energy source. Hence, there is a need for the design to be power efficient, high efficacy notwithstanding.
In literature, several such health monitoring systems have been proposed \cite{23}–\cite{26}. It is well known that acoustic symptoms such as cough, sneeze, belching etc. are early markers of an impending health issues such as influenza, diarrhea, whooping cough especially among children \cite{25} \cite{26}. If repetitive occurrence of these symptoms is detected in advance, it is possible for the patient or the healthcare personnel to commence remedial action prior to aggravation of the problem. The Kids Health Monitoring System (KiMS) system uses wearable sensors and acoustic signal processing in order to provide health monitoring in children \cite{26}. Using neural network based processing, the KiMS classifies various symptoms and activities and subsequently transmits the record to a parent or doctor for further analysis \cite{26}. Other proposed wearable systems in literature focus on a particular type of symptom such as cough \cite{27} \cite{28}. The primary limitations of such systems and algorithms is that; apart from feasibility of implementation from the power consumption perspective, these algorithms are too limited in terms of application and hence are not scalable.

This dissertation takes into consideration, the above mentioned challenges and develops algorithm-circuit co-design strategies in order to develop efficacious and pragmatic solutions for symptomatic pattern detection systems for use in implantable and wearable biomedical systems. In Chapter 2, a new DWT based algorithm and circuit is developed to detect epileptic seizures. This algorithm utilizes the frequency-time resolution property of the DWT in order to make the algorithm user-scalable. It also uses a quasi-averaging technique in order to enable a memory-less and feasible low-power implementation of the algorithm for implantable systems. Chapter 3 presents the silicon implementation of a multiple algorithm technique, which is also programmable to individual user. The implementation in 65nm bulk-Si TSMC technology also includes a self testing block, which ensures the correct operation of the system. In Chapter 4, a circuit level technique is developed where two stages are used in order to detect the onset of seizure. The technique is based on the principle that redundant complex computations can be avoided in the baseline period of the neural signals. This results in energy savings as well as improvement in detection efficacy. In
Chapter 5 the algorithm-circuit co-design strategy is applied to a human non-speech acoustic pattern detection system. This system processes five types of audio signals such as cough, sneeze etc. in order to correctly classify them. Correct classification can be then used in a wearable health monitoring system. The conclusion is drawn and presented in Chapter 6.
2. DISCRETE WAVELET TRANSFORM AND QUASI-AVERAGE BASED ALGORITHM FOR SEIZURE DETECTION

In this chapter, we have used the algorithm-circuit co-design strategy to find solution for designing an implantable application. The biomedical application chosen to illustrate this is “Epileptic Seizure Detection”. Based on certain characteristics of the disorder, viz. Epilepsy, a novel algorithm is developed to detect the onset of epileptic seizure., which utilizes Discrete Wavelet Transform (DWT) in order to detect the onset of the epileptic seizure. In the development of the algorithm and the corresponding circuit, it is carefully ensured that the implementation is optimized for low-power operation. At the same time the algorithm is also designed to achieve maximum efficacy of detection. This chapter begins with the basic concepts of discrete wavelet transform and the justification for its use in case of seizure detection. Subsequently, the DWT based algorithm and the training associated with it is explained in detail. The technique of quasi-averaging is also explained along with its advantages in terms of computational efficiency. Furthermore, the hardware implementation and circuit/architectural techniques used to implement the algorithm and reduce the power consumption are presented. Using in-vivo recorded neural data in the form of local field potentials (LFP), the algorithm is applied and the corresponding detection results are obtained for large animal (rat) studies (LAS). These results are presented and the system is evaluated on the basis of algorithmic efficacy and hardware power efficiency.

2.1 Discrete Wavelet Transform: Background

Discrete Wavelet Transform (DWT) is a powerful mathematical tool which has been used for signal processing in various applications. The principle of DWT is that,
a signal is represented using an orthogonal basis. This is similar to Fast Fourier Transform (FFT). However, the difference between FFT and DWT is that while Fourier transform uses the sine and cosine functions or “infinite” waves as the representation basis, DWT uses a family of wavelets which are “localized” waves. The energy of these wavelets is concentrated in time and space (Fig. 2.1). The advantage of such representation is utilized for an accurate time-frequency resolution of a non-stationary signal.

Additionally, DWT resolution preserves the spectral as well as temporal information contained in the signal. In the case of FFT, the temporal information is lost. It can be argued that Short-time Fourier transform (STFT), wherein FFT is computed over moving window of data, can be used to analyze a non-stationary signal. However, STFT uses a constant resolution at all the frequencies. On the other hand, DWT analyzes the signal using varying frequency resolutions. This can be seen in Fig. 2.2. Furthermore, unlike STFT, the width of the wavelet function (basis function) used in DWT changes with each spectral component in accordance with a scaling function. This multi-resolution technique is more accurate in order to decompose the signal into its constituent frequencies. The sub-band frequencies of the signal are represented in terms of DWT coefficients. DWT generates two types of coefficients, viz.
Fig. 2.2. Time-Frequency Resolution of STFT and DWT
“approximate” and “detail” coefficients. The “approximate” coefficients correspond to the low frequencies contained in the signal. The “detail” coefficients, also known as the ‘wavelet coefficients’, correspond to the higher frequencies present in the signal [29]. Mallat’s algorithm is used in this research to compute the discrete wavelet transform [29].

Mallat’s algorithm simplifies the expression for DWT into a filter representation. This is shown in Fig. 2.3 [30]. By using successive stages of low-pass (H) and high-pass filters (G) separated by intermediate down sampling operation, the approximate and detail coefficients of DWT can be computed. The output of the H filters is the approximate coefficients ($c_j$) and the output of the G filters is the detail coefficients ($d_j$). The filters in Mallat algorithm are designed such that the entire range of frequencies present in the signal is covered. These filters have a very sharp cut off in the frequency response. DWT has been popularly used in image compression. In the case of epileptic seizure detection, the resolved signal frequencies can be used to identify the bands showing seizure-like activity. By processing the corresponding frequency bands of interest, a decision can be made regarding the occurrence of the seizure.
As described previously, in order to compute the DWT, two types of basis functions are used viz. ‘scaling’ and ‘wavelet’ function (also known as mother wavelet). These basis functions are grouped into various wavelet families of mother wavelet [29]. Daubechies (DB) family of wavelets is one of the most popular orthogonal wavelets used for signal decomposition and reconstruction. These wavelets exhibit maximum range of smoothness or regularity [12]. The DB-mother wavelets are available with range of orders (2, 3, 4, 5...). The order of the Daubechies wavelet represents the complexity of the mother wavelet used in DWT computation. These DB wavelets can be efficiently implemented into filters in accordance with Mallat’s Algorithm. Daubechies 4th order (DB-4) wavelet has been shown to be the most appropriate for analyzing EEG neural data (Fig. 2.4) [18].

The order of the Daubechies wavelet is directly related to the order of the filter used to implement it in accordance with Mallat’s algorithm. Hence, the selection of the mother wavelet has a proportionally equivalent effect on the power consumption of the system. The higher order of the wavelet also indicates that the signal will be decomposed with higher detail. This implies that more information is preserved.
in its coefficients. A comparative study of the filter response for the high pass filter corresponding to each order of DB wavelet is shown in the Fig. 2.5. The output of the high pass filter is the wavelet coefficient and hence is of prominence for any signal processing. It can be seen that the higher order of mother wavelet changes the response close to the cut-off band. The difference is prominent only in terms of the cut-off slope. The Daubechies $6^{th}$ order wavelet has a sharper cut-off as compared to Daubechies $2^{nd}$ order mother wavelet. In this research, the Daubechies $4^{th}$ order mother wavelet is selected for analysis due to its spiking and smoothing consistency with the recorded neural signals. This justifies the use of Daubechies $4^{th}$ order mother wavelet to be used for analyzing neural signals. The algorithm methodology and procedure is described in the subsequent section.
2.2 Design Methodology

In the previous section, the selection of DWT as signal processing method and the appropriate mother wavelet was explained. The methodology for designing the wavelet based algorithm is explained in Fig. 2.6 This section describes the principle of the wavelet and quasi-average based algorithm (DWT-QA) [31]. It also describes the training and implementation strategy involved in order to tune the algorithm for a particular user.

2.2.1 Neural Data Recording

The first step in order to design a epileptic seizure detection algorithm is recording of the neural signals. In this thesis, the type of epilepsy under study is the focal human temporal lobe epilepsy. Kainate treated sprague-dawley rats are used to model the neural signals for testing the algorithm [17]. Traditionally, electroencephalograph (EEG) or electrocorticogram (ECoG) has been the basis of studying.
any neural recordings related to epilepsy. In Chapter 1, the limitations of EEG as compared to LFP, pertaining to detection of epileptic seizure, were highlighted. It is important to note that the use of LFP instead of EEG helps to avoid sensing numerous undesirable seizure-like signals originating from various regions of brain. This ensures that the recorded neural data is minimally modulated by these brain signals. This, thereby, helps in reducing the amount of signal processing needed post-recording. Moreover, the use of LFP reduces the latency of detection as the LFP is recorded using a micro-electrode tip placed close to epileptogenic focus [17]. Thus, the implantable epilepsy prostheses can make good use of LFP recordings for early detection and quick mitigation of seizure. In this research, the LFPs were recorded from kainate treated rats at a sampling frequency of 1526 Hz [17]. This sampling frequency ensures that the bandwidth is more than sufficient to capture the seizure signals. As will be shown in following sections, the seizure signals occur in specific frequency bands. However, these bands will vary between different animals. Hence, it is essential to have a large enough bandwidth in order to maintain user scalability. Along with the LFP, a simultaneous video footage was obtained for correct identification of seizures. The seizures were then marked visually and electrographically by inspection of data and corresponding video footage. A team of neurologists at the Indiana University School of Medicine (Indianapolis, IN, USA) verified the seizure patterns and the onset times were marked. The algorithm and system developed in this research is supposed to detect this onset of seizure. This classified data is then use to train and test the algorithm to evaluate its efficacy [17], [31].

2.2.2 Algorithm

In section 2.1, the time-frequency resolution property of the DWT was explained. The DWT-QA algorithm utilizes the DWT as a mathematical tool in order to process the recorded LFP. The Mallat’s algorithm, with its filter stages, is used for computation of DWT. The DWT coefficients at successive stages correspond to the various
constituent frequency sub-bands of the signal. It is clinically observed that during the onset of seizure, there is a gradual surge in the amplitude of the signal in a very narrow frequency band (20 Hz bandwidth). Since the seizure activity is confined to specific frequency bands, it is possible to select the corresponding wavelet coefficients in order to isolate this seizure-like activity and use it for detection of the onset of seizure [18], [31]. Mallat’s algorithm utilized for this purpose requires the implementation of multiple levels of high pass and low pass filters [19]. As mentioned in section 2.1, the DB-4 mother wavelet and its corresponding filter implementation is used to decompose the recorded LFP. Apart from DWT, a technique called “quasi-averaging” is also used. The processed LFP recordings are quasi-averaged over a predefined window size in order to smoothen out the spikes. Such spikes might result in false detection, thereby reducing the efficacy of detection. This operation is performed over a continuously moving window. The use of quasi-average, rather than the standard statistical average is advantageous in implementing the moving window in hardware without storing any data within the system. Finally, the averaged values are weighted, added and compared to a predefined threshold to register a detection of seizure.

2.2.3 Training

The neural recordings are known to exhibit a great degree of inter-subject variability. Hence, the algorithm used for detection of seizures in one particular subject may not show a similar efficacy in another subject. However, it is essential to make the seizure detection algorithm user-scalable. The training stage in the design methodology is necessary in order to achieve this. In the training phase, certain parameters used in the algorithm are kept tunable according to user specification. By adjusting the values of these parameters for each user, the same algorithm can be used over a range of users. In the proposed algorithm, the DWT coefficients, corresponding to the frequency sub-bands, the weights in quasi-averaging operation and the final threshold
are specific to each user. In order to decide the values taken by these parameters, the algorithm has to be trained. These parameters are selected such that the circuit implementation results in a significantly reduced hardware.

2.2.4 Implementation

Subsequent to the training of the algorithm, all the user-specific parameters are set to predefined values; the algorithm can now be implemented into hardware. Design techniques are utilized at the architectural level where computations are shared or truncation is used to reduce complexity of computational blocks. At the circuit level, filters are implemented in multiplier-less configurations, greatly reducing power consumption. Apart from that, quasi-averaging operation which enables memory-less implementation adds to the power savings. It should also be noted that since the data sampling frequency is very low at 1.5 KHz (approx), $V_{DD}$ scaling techniques can be employed to further reduce power consumption. Such an implementation would lead to a low-power epileptic seizure detection system which is more suitable to be used in a battery powered implantable prosthesis.

2.3 Algorithm and Hardware Implementation

Based on the design methodology described in section 2.2, in this section, the DWT-QA algorithm is described in detail. Subsequently, this section also presents the implementation methods along with the architecture and circuit-level techniques for optimizing power savings.

2.3.1 DWT-QA Algorithm

It is clinically observed that during the onset of seizure, there is a gradual surge in the amplitude of the signal in very narrow band of LFP frequency spectrum (20 Hz bandwidth). Analysis shows that, in order to filter such narrow bandwidth with a
standard FIR filter having a sharp cut off, a very high order filter (order of 300) is needed, which is practically unfeasible, especially for achieving a low power design. It was described in section 2.1 that in Mallat’s algorithm, the filter stages are separated by intermediate downsampling. The interleaving property of DWT due to this intermediate down sampling by factor of 2 helps overcome this constraint. It enables the use of wider bandwidth filters in succession to effectively achieve a very narrow bandwidth. Thus, the frequency-band specific, characteristic change in the amplitude of the LFP recordings can be captured by the algorithm in specific frequency components (wavelet coefficients). In this thesis, six stages of wavelet decomposition are needed in order to capture the seizure like activity. Hence, six levels of DWT coefficients are obtained for the sampling frequency of 1.526 KHz. Among these wavelet coefficients, the coefficients of interest are identified in the training phase. The selected DWT coefficients of interests are then converted to their unsigned values to give an estimate of absolute amplitude surge. These coefficients are then windowed over a predefined period of time. It should be noted that the lengths of the wavelet coefficient vectors at subsequent levels in DWT differ by a factor of two and hence need to be equalized. This is achieved by sizing the window appropriately. Subsequently, these coefficients are quasi-averaged over the continuously moving window.

Quasi-averaging is an approximation technique for averaging operation, which accurately models the average of a continuously moving window [31]. A preliminary analysis and comparison with standard “average” definition over a randomized set of data of about 100000 elements shows that the mean square error involved in this approximation is of the order $10^{-10}$. The approximation is based on the assumption, that each individual data point in a window (being averaged) can be represented by the average itself. Such a relaxation in the definition of “averaging” operation helps in a memory-less implementation in hardware. The quasi averages of the selected coefficients are then weighted and added to generate the ‘onset detection signal’. The weights help in normalizing the amplitudes of the coefficients which are amplified due to successive filtering. Finally, the onset detection signal is compared with a preset
threshold to identify and detect a seizure. However, the above designed algorithm uses several parameters, such as the DWT coefficients, the weights for quasi-averaging and the threshold value. These parameters are tuned or selected based on user specific requirements in the training phase.

2.3.2 Training for User-Specific Operation

In the training stage, the algorithm is tuned to a particular subject (user). This is achieved by processing a set of data, i.e. training data, with the DWT-QA algorithm. The training data is a selected set of LFP recording consisting of both the baseline and the seizure signals. The baseline signal refers to the part of neural signal when there is no electrographic or visual evidence of seizure. It is essential to include the baseline in the training data set in order to train the algorithm to minimize detection of false positives. This data is analyzed by implementing the DWT-QA algorithm in Matlab. It should be noted that the amount of training data available directly affects the efficacy of the algorithm. This data is processed using the algorithm described in section 2.3.1. The six-level decomposition of a sample training data set is shown in Fig. 2.7. The coefficients of interest are selected from this decomposed transform. Due to a large inter-subject variability in the neural recordings, the coefficients selected for the detection of onset would vary between subjects. As can be observed in Fig. 2.7, the three coefficients viz. $D_4$, $D_5$ and $D_6$ show significant seizure like activity in the case of this data set. These are the DWT coefficients in the $4^{th}$, $5^{th}$ and $6^{th}$ level decomposition of the wavelet transform. Since the sampling frequency of the depicted neural recording is $1.526 \, kHz$, the decomposition levels ($D_4$, $D_5$ and $D_6$) correspond to frequency bands of $48 - 95 \, Hz$, $24 - 48 \, Hz$ and $12 - 24 \, Hz$ respectively. The above mentioned DWT coefficients are selected and subjected to a statistical analysis using CDF (Cumulative Distribution Function) in the next step for training. A window size of 4096 raw input data values is used for this analysis. The CDF curves are then cut off at 98 % point. Note that the cut-off point can be
Fig. 2.7. Six-Level DWT decomposition of a neural signal (LFP)

Fig. 2.8. Seizure Detection Using DWT-QA Algorithm
changed depending on the required sensitivity and the epileptic history of the subject. Next, the ratio of the magnitudes of $D_4 : D_5 : D_6$ is evaluated at the cut-off point. In the decomposition shown in Fig. 2.7, this ratio was computed and approximated to be $0.5 : 1 : 2$. Such approximation helps in multiplier less implementation as the products can be obtained by simple shifting of data. Subsequently, the equalizing weights are chosen according to the ratio obtained from the statistical analysis. These weights are used to generate the onset detection signal using the algorithm described in section 2.3.1. Finally, a threshold is chosen so as to maximize the efficacy of detection. In this research, the threshold is selected by optimizing for the entire range of the onset detection signal. Fig. 2.8 shows the detection signal obtained using the algorithm for a sample training data set. It can be deduced visually that it consists of 3 seizure-like events. Note that, during the seizure event, the detection signal shows a gradual increase in the amplitude. This amplitude surge, in combination with a properly selected threshold, can be used to raise a detection flag. The wavelet transform filters away the high frequency components comprising the occasional spiking activity which could have led to false detection. The occasional spikes in the selected frequency bands are further smoothed out by the quasi-averaging operation over the moving window. The algorithm along with user-specific parameters can be implemented into circuits. The corresponding hardware and the circuit level techniques used are described in the following section.

2.3.3 Architecture Mapping

In this section, the algorithm is implemented into power efficient hardware. Using special circuit and architecture level techniques a low power implementation is made feasible. The functionality of the algorithm is unaffected. This ensures that the energy consumption of the system is reduced, without degrading the efficacy of detection. A top level block diagram of the system is shown in Fig. 2.9. It consists of a Wavelet
Decomposition block followed by the Quasi-averaging and the Thresholding block. These are described in following sections.

**Wavelet Decomposition Block**

The wavelet decomposition block is computationally, the most, intensive and power hungry block in the system. It consists of the discrete wavelet transform block to compute the DWT of the input and a counter to synchronize its operation with the streaming input signal. There are various ways suggested in the literature to implement efficient wavelet transform architecture [32]. These architectures are generally well suited for data compression applications and result in the loss of the intermediate derivable coefficients. This is because the final stage wavelet coefficient represents the maximum compression possible. However, the requirement of this system is to retain these coefficients in real time in order to analyze them for seizure like patterns. Hence, the folded architectures cannot be utilized for this application without the use
Table 2.1.
Coefficient Representation using Pre-computation

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Decomposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_0 = 01100111(103)$</td>
<td>$c_0 \cdot x = 2^5 \cdot (0011) \cdot x + (0111) \cdot x$</td>
</tr>
<tr>
<td>$c_1 = 10001011(139)$</td>
<td>$c_1 \cdot x = 2^7 \cdot (0001) \cdot x + (1011) \cdot x$</td>
</tr>
</tbody>
</table>

of memory elements and are ruled out [32]. In contrast to this, a simpler approach can be taken to implement the wavelet transform in accordance to Mallat algorithm (Fig. 2.3) [30].

In order to implement the algorithm, which evaluates on the basis of $D_4$, $D_5$ and $D_6$ DWT coefficients, 6 cascaded stages of the G and H filters are needed. Due to the selection of DB-4 mother wavelet each of the filters is of the $8^{th}$ order [29]. However, the number of G filters can be reduced as $D_1$ through $D_3$ coefficients are unused. Hence a total of 9 filters are required (6 LPF and 3 HPF). If these filters are implemented using standard FIR filter architecture, each of these filters would require a large number of power intensive multiplier and adder blocks. In order to overcome this problem, energy efficient wavelet transform block can be constructed by using multiplier-less architecture for FIR filter (G and H). We achieve this using the Computation Sharing Multiplier (CSHM) architecture for the filters [33]. Another commonly used multiplier-less technique for achieving power savings is by elimination of the common sub-expressions within the filter coefficient vector. Such approach is referred to as “Common Sub-expression Elimination” implementation (CSE) [34].

The main principle of CSHM is that, in vector scaling operations, any scalar $s_i$ can be decomposed into smaller bit sequences $a_k$ (alphabets). These alphabets are such that $s_i$ can be rebuilt from these sequences by few shift and add operations. Using these alphabets the coefficient vector ($C$) can be constructed spanning the entire set of filter coefficients. [33]. For instance, an alphabet set consisting of $\{1, 3, 7, 11\}$ can be used to represent the coefficients $\{103 & 139\}$ as shown in Table 2.1. This reduces
the entire multiplication operation of the filter to shift/select and add operations using Select/Shift and Adder (SSA) units instead of the power intensive multipliers. Fig. 2.10 shows the SSA unit utilized in implementation of the seizure detection system [33]. The pre-computer computes the product of alphabets and input vector $X$ in advance and stores them for re-use. The shifter inside the SSA selects the appropriate pre-computed product depending upon the bit-pattern of the coefficient alphabet. The pre-computed product is then shifted according to the decomposition of coefficient (Table 2.1).

Finally, these shifted pre-computed values are added to generate the SSA output $(C \ast X)$ [33]. The FIR filter can be implemented by using adequate number of alphabets and replacing the multiplier units with the SSA units. It should be noted that the number of alphabets directly translates to power dissipation of the pre-computer unit. Moreover, the number of communication buses (coming out of the pre-computer) is also proportional to the number of alphabets. Hence, power can be further reduced in such an architecture by minimizing the number of alphabets.
Table 2.2.
Original and Modified Filter Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Low Pass Filter</th>
<th>High Pass Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original</strong></td>
<td><strong>Modified</strong></td>
<td><strong>Original</strong></td>
</tr>
<tr>
<td>−3</td>
<td>−3</td>
<td>−59</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>183</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>−162</td>
</tr>
<tr>
<td>−48</td>
<td>−48</td>
<td>−7</td>
</tr>
<tr>
<td>−7</td>
<td>−8</td>
<td>48</td>
</tr>
<tr>
<td>162</td>
<td>162</td>
<td>8</td>
</tr>
<tr>
<td>183</td>
<td>184</td>
<td>−8</td>
</tr>
<tr>
<td>59</td>
<td>59</td>
<td>−3</td>
</tr>
</tbody>
</table>

However, in order to achieve this it is necessary to ensure that the reduced number of alphabets do not affect the frequency response of the filter significantly and provide adequate output quality. Since the CSHM based FIR filters are the basic component of the DWT based seizure detection circuit, it is essential that any change in filter frequency response must not degrade the overall detection efficacy. In order to achieve this a sensitivity analysis is performed by changing the number of alphabets, thereby, the filter coefficients, trying to ensure minimum quality degradation. Table 2.2 shows the initial filter coefficients used in DWT and the modified ones. It can be observed that alphabets \{1, 3\} are sufficient in order to represent all the coefficients with minimal effect on filter response. Fig. 2.11 shows the LPF filter response obtained with conventional and modified coefficient vector. It is evident that the error in the filter response is negligible. Note that by using the method described in [33] the optimality of the modified coefficient vector is verified in terms of chebychev error.

Apart from CSHM that tries to increase the reusability of common sub-expressions through alphabets, several other methods have been developed to identify and elim-
inate common sub-expressions within the filter coefficients [34]. One of the effective methods explored in this research is level-constrained common sub-expression elimination (CSE) algorithm [34], which can constrain the number of adders along with the number of adder levels (AL) required to compute each of the coefficient outputs. This reduction in number of AL translates to lower complexity architecture and thus, lower power. This method can be briefly explained as follows. Consider two functions \( F_1 \) and \( F_2 \), where \( F_1 = 13 \times X \) and \( F_2 = 29 \times X \). Both \( F_1 \) and \( F_2 \) can be represented in terms of shift and adds in the following manner:

\[
F_1 = X + X \ll 2 + X \ll 3 \\
F_2 = X + X \ll 2 + X \ll 3 + X \ll 4
\]

Both the expressions, \( F_1 \) and \( F_2 \) have some common terms,

\[
D = X + X \ll 2 + X \ll 3
\]

Therefore, \( F_1 \) and \( F_2 \) can be rewritten as:

\[
F1 = D
\]
\[ F2 = D + X \ll 4 \]

Reusing D in both the expressions reduces the computation overhead and the number of adders required to implement both expressions. Filter implementation using CSE can result in significant power savings by reducing number of adders and shifters. However, note that higher sharing might come at a cost of increased computational path thus reducing the operating frequency. There is a trade-off between power consumption and the frequency requirements in case of CSE-based implementation. However, since the operating frequency of the seizure detection system is very low (KHz) due to low sampling frequency (1.5 KHz), this trade off can be utilized positively.

The CSE algorithm used in this paper results in constraining the computational time of the critical path by controlling the number of ALs [34]. The DWT coefficients are constrained to be implemented within 4 ALs. This provides good trade-off between computational delay and the number of adders. Note that apart from sharing of computations, other minute approximations, such as inverting a bit in order to reduce the number of ‘1’s can be made. This results in computationally efficient filter coefficients by minimizing the number of computed products needed for multiplier-less architectures. Eventually, this manifests as reduced power consumption and area of the hardware used to implement the same. Using the above mentioned circuit-level techniques, the DWT circuit can be implemented for the wavelet decomposition block.

In this application, a total of 6 Low pass filters (H) and 3 High pass filters (G) are used. The filter coefficients are digitized and approximated to binary powers. This minimizes the number of ‘1’s, and the number of pre-computed products needed. This reduces the power and area consumption of the implemented hardware. In accordance with Mallat’s algorithm, the down sampling operation between every successive stage of the DWT is performed using a ripple counter which acts as a clock divider [30], [29]. Due to this dyadic scaling in DWT operation, the adjacent output bits of a counter are used to clock the successive filter stages in the DWT block. The DWT coefficients along with corresponding clock signals are propagated to the quasi averaging blocks for further signal processing in accordance to the DWT-QA algorithm.
The Quasi-Averaging (QA) block consists of two parts, viz. an absolute value component followed by quasi averaging units for each of the selected DWT coefficients. The DWT coefficients are computed in the 2’s complement form. Since the input is a non-stationary signal with random variation, calculating a moving window average on its component frequency will yield near zero value. To overcome this, the coefficients are converted into their unsigned absolute magnitudes for further analysis. The unsigned coefficients obtained are subsequently provided to the QA blocks to calculate the quasi average over a continuously moving window. Fig. 2.12 shows the block diagram depicting the principle of quasi averaging. The function of a continuous moving-window quasi-averaging operator is explained as follows. If $S_{i:i+w}$ is the sum of elements $x_i$ in the window $W_k$ of window size $w$, then the average of $W_k$ is given by

$$\langle W_k \rangle = \frac{1}{w} S_{i:i+w} \quad (2.1)$$

The average of the next window $W_{k+1}$ can be calculated as

$$\langle W_{k+1} \rangle = \frac{1}{w} (S_{i:i+w} - x_i + x_{i+w+1}) \quad (2.2)$$

However, this would require storing the first $w$ elements for subsequent usage. The QA technique approximates this calculation by the assumption that “the average of

![Fig. 2.12. Block Diagram for Quasi-Averaging Technique](image)
a window is the true representation of all the elements contained in it”. The Eq.(2.2) can then be written as

\[ \langle W_{k+1} \rangle = \frac{1}{w} (S_{i:i+w} - \langle W_k \rangle + x_{i+w+1}) \]  

(2.3)

This helps implement the moving window in a much simplified circuit and more importantly, without the usage of memory elements. This technique greatly facilitates implementation to achieve real time operation, for epileptic seizure detection application. Three such QA modules are used for $D_4$, $D_5$ and $D_6$. The corresponding window size used are 512, 256 and 128 respectively. These window sizes corresponds to a window of 4096 samples in the raw LFP data sequence. The selected window sizes simplify the hardware implementation by simplifying the divider operation. Due to the window size being powers of 2, division is carried out by dropping the least significant bits corresponding to those powers. The windowed quasi averages of each detail coefficients $\langle D_4 \rangle$, $\langle D_5 \rangle$, $\langle D_6 \rangle$ are then weighted with the weights of 2, 1, and 0.5 respectively. The weighted sum of the quasi averages forms the onset detection signal, which is passed to the next stage to compare against a threshold and generate the seizure detection flag.

**Thresholding Block**

In the Thresholding Block, a simple comparator, made out of XOR gates, compares the onset detection signal to the prefixed threshold value. A flag signal is generated and pulled to ‘1’ at the onset of the seizure. Note that the frequency of operation of the entire circuit is the same as the sampling frequency of the LFP recording. Since this frequency is very low, $V_{DD}$ scaling techniques can be utilized to further reduce power dissipation due to quadratic dependence of power on $V_{DD}$. The obtained efficacy and power results from circuit simulations are discussed in the following section.
2.4 Efficacy and Efficiency Results

The DWT-QA algorithm based epileptic seizure detection system was designed and simulated using 90nm (IBM) bulk-Si technology library. As mentioned in previous section, the neural data was obtained using an implantable LFP neural recording system from live animals (Fig 2.13) [17]. The functionality of the algorithm was verified by comparing the results from software simulation in Matlab with hardware simulations. The system was evaluated on two aspects viz. Algorithmic Efficacy and Hardware Efficiency.

2.4.1 Algorithmic Efficacy

Algorithmic Efficacy provides an insight to the correctness or accuracy of the detection. It is based on three interdependent parameters, viz. Sensitivity, Specificity and Average Detection Rate (ADR). The recorded in-vivo LFP data was used as test data. The hardware implementation of the algorithm in VHDL was simulated for a real time system level simulation using Modelsim. The seizures in the data were identified and time stamped by electrographic evidence and visual confirmation.
Based on this time stamped data, the detection was classified as a True Positive (TP), False Positive (FP), and True Negative (TN) and False Negative (FN). The most important among these classified detections are the FP and FN. It is essential to minimize both these parameters in order to increase the efficacy. The specificity, sensitivity and the ADR were calculated as per following equations (Eq. 2.4–2.6) [18].

\[
Sensitivity = \frac{TP}{TP + FN} \times 100\% 
\]  
(2.4)

\[
Specificity = \frac{TN}{TN + FP} \times 100\% 
\]  
(2.5)

\[
ADR = \frac{Sensitivity + Specificity}{2} 
\]  
(2.6)

The process was repeated over neural recordings obtained from five different animals. Three different mother wavelets viz. Daubechies 2\textsuperscript{nd}, 4\textsuperscript{th} and 6\textsuperscript{th} order were used in order to compare the result. The corresponding results are tabulated in Table 2.3. It should be noted that specificity and sensitivity are mutually related parameters. If during the training phase, if a much lower threshold is chosen, it would lead to a higher value of FP. This would also imply that the system is very sensitive to any seizure-like events in the recorded LFP data. In contrast, a higher threshold would result in reduced FP. However, this would lead to high probability of missing a seizure

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>ADR (with DB4)</th>
<th>ADR (with DB2)</th>
<th>ADR (with DB6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAS-2</td>
<td>100%</td>
<td>98.175%</td>
<td>99.09%</td>
<td>81.78%</td>
<td>73.33%</td>
</tr>
<tr>
<td>LAS-3</td>
<td>100%</td>
<td>94.85%</td>
<td>97.43%</td>
<td>71.30%</td>
<td>72.86%</td>
</tr>
<tr>
<td>LAS-4</td>
<td>96.88%</td>
<td>86.60%</td>
<td>91.74%</td>
<td>70.47%</td>
<td>71.89%</td>
</tr>
<tr>
<td>LAS-7</td>
<td>93.75%</td>
<td>99.69%</td>
<td>96.72%</td>
<td>78.75%</td>
<td>84.84%</td>
</tr>
<tr>
<td>LAS-8</td>
<td>100%</td>
<td>96.24%</td>
<td>98.12%</td>
<td>68.68%</td>
<td>92.10%</td>
</tr>
</tbody>
</table>
and hence result in increased FN. Since specificity and sensitivity are complementary to each other; the ADR is a good measure to calculate the overall efficacy of the system. As mentioned in section 2.1, and as it is clearly seen from Table 2.3 that the Daubechies 4th order mother wavelet is best suited for characterization of the neural signal as compared to 2nd and DB6th order. This is evident in all five cases. Moreover, DB-6 mother wavelet leads to a 12th order Mallat filter implementation, which would be higher in power consumption as compared to the DB-4 based filter. Although DB-6 has a higher order filter, the coarse and smooth features of the mother wavelet are not suitable for the neural recordings in these subjects. On the other hand, DB 2 wavelet, is least suitable for seizure detection. As can be seen from Table 2.3, the algorithm achieves a high value of average detection rate ranging from 91% to 99% with DB-4 wavelet. A seizure miss or an FN will have an adverse effect on the ADR. Although this can be compensated by lowering the threshold, such an adjustment would increase the sensitivity. Since there are far more seizure-like events which might trigger FP as compared to FN, an increased sensitivity in lieu of a decreased specificity will result in degradation of ADR which is undesirable [31].

Table 2.4.
Power Dissipation and Area Improvement using IBM 90 nm bulk-Si Technology

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Power (µW)</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$V_{DD} = 1$ V</td>
<td>$V_{DD} = 500$ mV</td>
</tr>
<tr>
<td>CSE (multiplier-less)</td>
<td>396</td>
<td>77</td>
</tr>
<tr>
<td>CSHM (multiplier-less)</td>
<td>400</td>
<td>75</td>
</tr>
<tr>
<td>CSA (with multiplier)</td>
<td>440</td>
<td>80</td>
</tr>
<tr>
<td>ANN based wavelet[14]</td>
<td>6.3 mW @ 1 V</td>
<td>6 X</td>
</tr>
</tbody>
</table>
2.4.2 Hardware Efficiency

As mentioned in previous section, apart from the algorithmic efficacy, it is equally important to see the feasibility of the algorithm in terms of a pragmatic implementation. The hardware implementation of the algorithm using the architectures discussed before was simulated using IBM 90nm bulk-Si library. The functionality was implemented in VHDL and checked using Modelsim. Subsequently, Synopsys Design Compiler was used to synthesize the system. The synthesized Verilog netlist was converted to Hspice netlist for transistor level simulation. Nanosim was used to simulate the system at transistor level to obtain the power consumption of the system. The power dissipation for the various architectures at nominal and scaled $V_{DD}$ is tabulated in Table 2.4. Note that the system was designed and optimized for a maximum clock speed of 1 MHz. However, due to requirement of low frequency of operation (1.5 KHz), $V_{DD}$ can be scaled down. This results in significant reduction of power consumption due to quadratic dependence on supply voltage. As seen from Table 2.4, the use of multiplier-less architectures along with $V_{DD}$ scaling results in reduction in power consumption by over 80%. It can also be observed that there is almost a 90% improvement in power consumption as compared to the Artificial Neural Network (ANN) and wavelet based seizure detection system. This is because of the fact that, the ANN based system uses a large number of ADCs in its wavelet processor [35]. There is also a significant reduction in area of the entire system due to the elimination of the multipliers from the FIR filter and sharing or elimination of common sub-expressions in CSHM/CSE [33], [34]. An FFT based system (based on EEG) is also shown to consume significantly more power owing to increased computation (2.1$\mu$W @0.5Hz, 1V, 180nm) [15]. Both the STFT and the ANN based algorithm would also have to be made user scalable due to the inter-subject variability to maintain a good efficacy at the given power. Comparatively, the DWT-QA algorithm and its implementation (in 90nm bulk Si) is better in performance and lower in power consumption making it a very good prospect for use in an anti-epilepsy prosthesis.
2.5 Conclusion

In this chapter, a novel low-power epileptic seizure detection system based on DWT and QA operation was developed [31]. The algorithm was designed to be programmable to user specific needs. The feasibility of the algorithm was maintained by utilizing power efficient design techniques at multiple levels of design abstraction. This chapter also highlighted the importance of the selection appropriate mother wavelet. At the algorithmic level user-specific critical parameters (DWT coefficients, weights and threshold) were identified to accurately detect the onset of seizure. These parameters were used to develop and train a highly efficacious algorithm. Furthermore, this algorithm was mapped and implemented into a low power hardware. Multiplier-less techniques were utilized at architectural level to reduce the power consuming logic elements. The resulting epileptic seizure detection system showed high detection efficacy, when tested on LFP data from in-vivo animal recordings. Simultaneously, low power operation was also achieved. It should be noted that due to inter-subject variations in neural LFP recordings, it is essential to provide flexibility in terms of maximum programmability to the algorithm. The proposed system has this feature in terms of the selected coefficients of interest, the weights for quasi averaging and thresholds. In the next chapter, an implementation of a technique using multiple algorithms will be presented to further enhance the programmability of the system. Such user-scalable systems which optimize efficacy and power consumption at both algorithm and circuit levels of abstractions are essential towards development of battery-powered implantable epilepsy prosthesis.
3. EPILEPTIC SEIZURE DETECTION PROCESSOR WITH MULTI-ALGORITHM PROGRAMMABILITY

In this chapter, a silicon implementation of a seizure detection processor is presented. The processor is designed with multiple algorithms in it for accommodating the inter-patient variability in the neural signals. The system is designed to be programmable for each user. This has been incorporated by making it possible to combine the detection by each algorithm using Boolean logical operators. Apart from the algorithm itself, a built-in self test (BIST) mechanism has also been implemented. Using the BIST, each block of the processor can be tested internally for correct functionality in terms of computational integrity. Due to lower hardware footprint of the individual algorithm, the multi-algorithm processor is implemented with a low power consumption. The choice of individual algorithms or their combinations is patient specific. This enables the usage of the developed epileptic seizure detection processor over a wider range of patients.

3.1 System Overview

In Chapter 2, the DWT-QA algorithm was developed and implemented successfully. It was shown to be highly efficacious. However, the efficacy varied over the various test subjects. Thus, the inter-patient variability of the neural LFP signals was highlighted and the importance of programmability in any seizure detection algorithm was emphasized. Due to this variable nature of LFP data, the seizure detection efficacy for each user varies over different algorithms. A high efficacy algorithm for a particular user may not be necessarily efficacious enough for another user. Hence, one of the ways to improve the user scalability of a seizure detection processing unit is by having a provision for multiple algorithms. Depending on the requirements of the
user, the correct algorithm can be selected in order to achieve the best results. This is the basic principle of the multiple algorithms processor presented in this chapter. Apart from that, based on the study by Raghunathan et. al. [36], it is verified that, by combining the detection outputs of the implemented algorithms using simplistic Boolean logic functions, it is possible to increase the detection efficacy significantly. This comes at a marginal cost of increased hardware. The entire processor developed in this chapter is operated at a $V_{DD}$ supply near the threshold-voltage of the transistor (400 mV in the implemented TSMC 65 nm bulk Si) resulting in low-power consumption.

The implemented system has an algorithm bank consisting of four algorithms. Each of these algorithms is based on a mathematical parameter or metric, which can be used to extract the seizure feature from the input LFP signal. These algorithms are selected based on their ability to extract the seizure onset feature and proven functionality for seizure detection application. The seizure features are used to demarcate the seizure from the baseline. The algorithms operate on real-time input data (LFP) and produce processed signals. The processed signals are then compared with prefixed threshold values to cause a detection of the onset of seizures. In this system, there are four detection signals corresponding to four algorithms. Each detection can be used individually to detect the onset of seizure. However, in order to increase the efficacy of detection, the detection from the four algorithms can be logically combined according to the needs of the patient. Such programmability for high efficacy operation enables the use of the processor on a wider range of patients. The thresholds for each of the algorithms and the Boolean logical combination to be used are decided in the training phase. The following section describes the algorithms, the Boolean combination technique and the training phase in detail.
3.1.1 Algorithms

As mentioned before, the system consists of four algorithms in the algorithm bank. These are the Energy, Coastline, Non-Linear Energy and Hjorth parameter based algorithms. The choice of these algorithms is based on their abilities to isolate the ‘onset’ feature. The Energy and Coastline are simple arithmetic parameters which measure the energy content and the trace-length of the LFP recordings respectively. The Hjorth variance parameter and Non-Linear Energy are more statistical in nature. They are also capable of capturing the non-linear features in the recorded LFP. All of these algorithms operate in the time domain on real-time data stream. The recorded LFP data is digitized into 10-bit 2’s complement format and streamed into each of the four algorithms. The definition of the four parameters and their operation is explained in following section.

Energy

The “Energy” parameter calculates the energy content of the signal in a prefixed window. It is a well known observation that as the onset of seizure approaches, there is gradual and continuous surge in the amplitude of the signal in specific frequency bands. This can be sensed in the form of an increase in energy content of the signal. The Energy parameter operates on adjacent windows of the streaming data. As the energy of these adjacent windows is computed, a gradual increase in the energy value implies that the signal is fast approaching a seizure onset. The size of the window is taken as 1024 samples in this case. By appropriately comparing the energy value in the window with a prefixed threshold, a detection of the onset can be made. The mathematical description of the energy metric is given in Eq. (3.1). Here \( E \) is the energy parameter, \( x \) is the input data, \( N \) is the size of the window and \( n \) is the window index.

\[
E_{AVG}[n] = \frac{1}{N} \sum_{i=1}^{N} E(i + (n - 1) \times N) \tag{3.1}
\]
where,
\[ E(i) = x^2(i) \]

**Coastline**

The “Coastline” parameter is also known as ‘line-length’ or ‘trace-length’ algorithm. Its definition has been adapted from [38]. This feature measures the actual physical length of the trace of the signal over a certain period of time or window size. This is computed by measuring the absolute distance between adjacent data points \( x \) in the input LFP signal and accumulating them over a window of width \( N \). In the baseline section of the LFP recording, this parameter remains more or less constant. However as one approaches the onset of seizure, due to increase in amplitude and frequency of spikes in the signal, the distance between the neighboring data points increases. Hence the accumulated trace length value also increases and can be compared with a threshold to raise a detection flag. The threshold for this parameter can be adjusted accordingly in the training phase. The Neuropace responsive stimulation device by Medtronic Inc. utilizes the “coastline” or “line-length” as one of the parameters for seizure detection [8], [36]. The mathematical expression for computing the coastline (CL) for \( k^{th} \) window is given in Eq. (3.2).

\[
CL(k) = \sum_{i=1}^{N} |x[i + (k - 1) \times N] - x[i - 1 + (k - 1) \times N]| \quad (3.2)
\]

**Non-linear Energy**

The “Non-linear Energy” parameter is based on Teager’s algorithm [39]. In the conventional energy parameter, all the frequency components in the signal were weighed equally. In comparison, the non-linear energy operator weighs the components at different frequencies non-uniformly. Using square-law weighting technique, it emphasizes higher frequencies over lower frequencies. Since it is known that there is a surge in amplitude at higher frequencies in the seizure phase of the LFP recordings,
it expends greater energy during the seizure or ictal phase. This non-uniformity is compensated in the Teager’s algorithm. It is computed by accumulating the difference between the square of current sample and the product of the neighboring samples. Approaching the onset, due to higher occurrence of pike-like signals, this difference is high and hence shows an increase in the non-linear energy parameter. It has been shown that Teager’s algorithm emphasizes with a higher contrast, the distinction between the energy content in the baseline and the ictal phase [40]. This results in better seizure detection. The non-linear energy $\Psi$ for $k^{th}$ window of size $N$ for data $x$ is given by Eq. (3.3).

$$\Psi(k) = \frac{1}{N} \sum_{i=1}^{N} x^2[i + (k - 1) \times N] - x[i + (k - 1) \times N - 1] \times x[i + (k - 1) \times N + 1]$$ (3.3)

### Hjorth Variance Parameter

“Hjorth Variance” parameter has been widely used for statistical analysis of EEG [41], [42]. There are numerous Hjorth variance parameters that can be calculated. The first of these parameters also termed as “activity” corresponds to the variance of the signal amplitude. It calculates the variance of a window of $N$ samples. This variance is then averaged over adjacent windows in order to obtain the mean variance. The mean of the $k^{th}$ window, $\mu_k$, is also calculated to obtain the first Hjorth parameter. The mathematical expression for computing the Hjorth Variance parameter is given in Eq. (3.4).

$$Var(k) = \frac{1}{N} \sum_{i=1}^{N} x^2[i + (k - 1) \times N] - \mu_k^2$$ (3.4)

where,

$$\mu_k = \frac{1}{N} \sum_{i=1}^{N} x[i + (k - 1) \times N]$$
3.1.2 Boolean Logical Combination

The four algorithms described above are used to process the input LFP data and independently detect the onset of the seizure. However, in some cases these algorithms are unable to provide the necessary efficacy individually. This could be due to noisy recordings or due to simplistic nature of the metrics themselves. In such cases, the multi-algorithm seizure detection system can be programmed to provide a Boolean combination of the individual detection of any two of these algorithms. By using simple Boolean ‘AND’ or ‘OR’ operation on the individual detection, it is shown that the detection efficacy of the system is substantially improved in lieu of a marginal increment in hardware cost. For instance, it can be seen in Fig. 3.1
that the false positives detected by the coastline algorithm are eliminated by using the Boolean ‘ANDing’ of the coastline and the Hjorth parameter detections. It can also be seen that this might lead to increased delay of detection. Depending upon the nature of training data and neural LFP recording for each patient the best possible combination of algorithms for the patient can be identified. The detection could be independent using individual algorithm or by the Boolean combinations of multiple algorithms. This ability of programming enables the use of the processor over wide range of patients, thereby increasing the scope of its application.

3.1.3 Training Phase

The four algorithms described in previous section generate processed values from the input according to Eq. (3.1–3.4). These values have to be compared with thresholds in order to classify the data as a baseline or seizure. These thresholds are prefixed in the training phase. In order to train the algorithms, a set of data, ‘training data’, is needed. The training data is randomly selected from the LFP recordings consisting of an equal duration of baseline and seizure signals. This data is subsequently processed using each of the four described algorithms and the threshold is adjusted in order to have the highest efficacy of detection. The definition and method for calculation of efficacy will be discussed in section 3.3. The training phase is also used to decide the Boolean logic to be used for the combination of multiple algorithms. This is because a patient may not respond positively to individual algorithm but may show a significant improvement if a combination of algorithms is used. Subsequent to fixing of threshold and Boolean logic to be used, the algorithm and the supporting circuitry can be implemented into a power efficient hardware.

3.2 Hardware Implementation

The programmable multi-algorithm seizure detection processor is implemented as shown in the block diagram in Fig. 3.2. The primary blocks of the system are the
Algorithm Bank, Threshold Bank, Boolean logic selector and the Built-in self-test (BIST) blocks. The system can be operated in two functional modes viz. ‘operate mode’ and the ‘test mode’. The operate mode is used to put the system in the seizure detection mode. On the other hand, the test mode activates the BIST block functionality and performs a self test on the system to verify its integrity. The individual implementation and function of each of these blocks is described in the following section.

3.2.1 Algorithm Bank

The algorithm bank consists of four parameter based algorithms as described in the section 3.1 viz. Energy, Coastline, Non-linear energy and Hjorth Variance parameter. The hardware implementation is based on the Eq. (3.1–3.4) respectively. The block diagram for each of these algorithms is shown in Fig. 3.3.
The Energy parameter uses the multiplier to square the 10-bit input data and an adder to accumulate it over a window. The window width is 1024 and is controlled by the counter. The counter generates a signal at the end of the window which resets the registers and loads the computed average energy to the output register. Since the window width chosen is a power of two, the average energy is calculated by simply dropping the last 10 bits of the accumulated sum. This avoids a power hungry divider for computing the average. The accumulated sum is passed on to the threshold register bank for comparison and detection.

The implementation for coastline parameter is shown in Fig. 3.3(b). This is in accordance to Eq. 3.2. The registers at the input store the two adjacent values of the signal. The absolute difference is then accumulated by the adder. The width of the window is controlled using a counter which generates a signal for passing the accumulated difference to the threshold bank.

The Non-linear energy parameter is computed using the sample preceding and succeeding each data point as seen in Eq. (3.3). This is implemented in hardware.
by introducing a delay of one clock cycle. The first value is computed only after 3 samples of data are streamed in the system. Thereafter the operation is pipe-lined. The counter is used to control the window size and pass the output to the threshold bank.

The Hjorth parameter is calculated by finding the difference between the accumulated squared data and the squared mean in a window of 1024 samples. The counter controls the window size. The Hjorth first parameter is compared to threshold in the threshold bank.

In the operate mode, all the algorithms are simultaneously processing on the 10-bit input data for detecting the seizure (Fig. 3.4). However in the test mode, the algorithms are provided with previously stored data vectors from the BIST block. The intermediate processed values indicate the correctness of computation in each of the algorithms. It should be noted that, the algorithm bank can be scaled to encompass more algorithms. In this research, four algorithms have been implemented in order to test the concept.

### 3.2.2 Threshold Bank

The processed signals from the algorithm bank are compared to thresholds stored in the threshold bank. The threshold bank consists of four 20-bit registers. These registers hold the threshold values for each of the algorithms. The values in the threshold registers are decided in the training phase as described in the previous section. In the operate mode, the threshold bank is loaded with these values by serially streaming in the bits using a dedicated clock. The processed signals are compared with the thresholds to detect the onset of the seizure. In the test mode, the clock can be used to serially stream out the stored value to verify its integrity. The system has a dedicated clock to operate with threshold registers. This is independent of the system clock and can be turned off after the loading operation of the threshold registers is complete.
3.2.3 Boolean Logic Selector

The Boolean logic selector is the programmable block which can be used to select the type of output needed for seizure detection. Depending on the select lines of this block, the output can be either the detection by individual algorithms or their Boolean ‘AND/OR’ combination. Since the system is designed to be efficacious over a range of patients, the type of logic used for seizure detection is decided in the training phase. In the test mode, by appropriately applying the select values the Boolean logic selector can be tested for its functionality of multiplexing the correct data to the output.

3.2.4 Built-in Self Test (BIST)

The seizure detection system implemented on the ASIC includes a built in self testing block (BIST). The BIST block is provided to have an on chip testing circuitry.
This enables verification of the functional integrity of each block of the system. The BIST block is brought into operation by pulling the test / operate signal low. It provides a set of prefixed, hard-coded data vectors to the algorithm bank to test the functionality of the algorithms. The intermediate outputs of each of the algorithms are compared to stored values to verify their correctness. The comparisons generate signals which are given to the Boolean logic selector. The Boolean logic selector, which was used for logically combining the detection outputs of each algorithm is used for selecting the functionality verification flags in the ‘test’ mode. By using appropriate select signals the output of the system can be used to check the functionality of each of the algorithms. The BIST is also used to test the threshold bank registers by streaming out the stored threshold values through a serial output port by a dedicated threshold clock. The Boolean logic selector block is tested by internally/externally providing logical inputs to the block and testing the selectivity of the block. The BIST covers all the significant sections of the system to maintain sustained and correct functionality.

3.3 Efficacy and Efficiency Results

Based on the algorithm and hardware implementation described in the previous sections, the multi-algorithm seizure detection processor was implemented in 65 nm bulk-Si TSMC technology. Fig. 3.5 shows the die photograph of the system on TSMC 65 nm technology. Fig. 3.6 shows the test setup used for testing the ASIC using LFP test data. The functional integrity of the system was tested using the internal BIST block. The algorithm was designed using Matlab and the efficacy obtained was also verified by comparing the hardware simulation with the MATLAB simulation result. The LFP neural data was recorded in-vivo from large animal studies (rats) at a sampling rate of 1.526 KHz along with corresponding video. As mentioned before in Chapter 2, the seizures were induced in the animals using periodically administrated kainic acid injections. The LFP data, electrographically classified into seizure and
Fig. 3.5. Die Photo of Seizure Detection Processor

Fig. 3.6. System Test Setup
baseline by medical team as mentioned in Chapter 2, was used to test the efficacy. The recorded LFP data was streamed into the system input in a digitized 10 bit 2’s complement form. The system was enabled in the “operate mode” and the output was observed for detection flag indicating the onset of seizure. The Boolean logic selector is programmed according to the training phase to select the best configuration of Boolean logic suited for each set of LFP test data. Subsequently, the efficacy was calculated based on the Specificity, Sensitivity and Average Detection Rate (ADR) as defined by Eq. (2.4–2.6).

The Sensitivity represents the ability of the algorithm to detect any seizure-like event in the input signal. These events could be due to onset of seizures or motion artifacts such as wet-dog shake. A higher sensitivity ensures that the algorithm will not miss any seizure-like activity. However, it also implies an increased FP. On the other hand, Specificity represents the ability of the algorithm to reject a FN. Since these two metrics are mutually dependent, by appropriately optimizing for a good ADR, the algorithm efficacy can be maximized. Using the above mentioned efficacy metrics, the efficacy results are summarized for the four implemented algorithms in Fig. 3.7. It can be seen that the efficacy of the algorithms vary over different animals due to inter-subject variability of LFP data. On an average, the Hjorth variance parameter algorithm performs better than the others individually. However, it has a higher hardware cost. The coastline parameter has a much smaller hardware cost but also a reduced efficacy. However, if we use a Boolean combination of two of these algorithms, the efficacy can be increased for a marginal increase in hardware. This is shown in Fig. 3.8. In this figure, the average efficacy of the four algorithms is plotted with respect to normalized hardware cost. The hardware cost is calculated by combining the normalized total power consumption and the area of the hardware implementation of the algorithm. The area and power for the algorithm hardware implementation is calculated using Synopsys and Cadence tools. As seen in Fig. 3.8, the coastline and hjorth parameter are combined using a single Boolean ‘OR’ gate.
Fig. 3.7. Efficacy of the Algorithms for Large Animal Study

Fig. 3.8. Average Efficacy vs. Normalized Hardware Cost
This configuration increases the effective efficacy. By specifically configuring the system for each patient, the best possible efficacy can be achieved.

The seizure detection processor was synthesized using Synopsys Design Compiler. The seizure detection processor ASIC was operated with a clock of 10 KHz. Such low frequency of operation allowed for aggressive voltage scaling in order to reduce the energy consumption of the system. The supply voltage was scaled to 400 mV which is very close to the transistor threshold voltage of 310 mV for the TSMC 65 nm technology. This is the most energy efficient region of operation. Due to quadratic dependence on the supply voltage ($V_{DD}$) the dynamic power consumption is significantly reduced. The processor functionality was verified. The power consumption of the system is listed in Table 3.1. Along with scaling the system $V_{DD}$ to 400 mV, the I/O voltage was scaled to 0.8 V. The four algorithms are operational at all times, irrespective of the selection for Boolean combination, and consume about 12.8 µW as leakage power. The dynamic power is 360 nW. This is almost 80% lower than the wavelet based algorithm (DWT-QA as described in Chapter 2) [31]. Although the parameter based algorithms do not show higher efficacy than the wavelet based algorithm, the programmability feature can be utilized to improve the efficacy with marginal change in power consumption. It should be noted that, in this implementa-

<table>
<thead>
<tr>
<th>Technology</th>
<th>65 nm bulk-Si (TSMC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension</td>
<td>200 µm X 340 µm</td>
</tr>
<tr>
<td>I/O supply</td>
<td>0.8 V</td>
</tr>
<tr>
<td>Core supply</td>
<td>400 mV</td>
</tr>
<tr>
<td>$f_{clk}$</td>
<td>10 KHz</td>
</tr>
<tr>
<td>$P_{leakage}$</td>
<td>12.8 µW</td>
</tr>
<tr>
<td>$P_{dynamic}$</td>
<td>360 nW</td>
</tr>
</tbody>
</table>
tion, the power consumption reported is high as the system is implemented without the utilization of any leakage control techniques such as power gating, clock gating etc. However, programmable logic and near-threshold voltage operation make this seizure detection processor efficient as well as practical for implantable prosthesis.

3.4 Conclusion

The implemented “near-threshold” seizure detection processor is low in power consumption and has adequate efficacy [37]. Furthermore, the provision of programming the combination of algorithms makes it adaptable to wider range of patients by adjusting the efficacy dependent on severity of the epilepsy. The seizure detection processor successfully addresses two major concerns in development of implantable epilepsy prosthesis viz. low power consumption and accurate functionality over inter-patient variation of LFP neural data. The processor is operational at near threshold voltage supply. This reduces the power consumption significantly and hence increases the longevity of the battery. Leakage control techniques such as powering off the unused algorithms can be applied to further reduce the power. In the next chapter, a multi-stage seizure detection technique employing the power-gating principle will be presented. It will be shown that the use of such low power leakage control techniques help in further reducing the energy consumption of the system.
4. DETECTION OF EPILEPTIC SEIZURE USING MULTIPLE(2) STAGES

In this chapter, a multi-stage circuit level technique is presented to detect epileptic seizure. This uses two algorithms of varying complexity in order to monitor and subsequently detect the seizure onset. The first stage, which is a low complexity algorithm, monitors the LFP input data for any seizure like event. The second stage, consisting of a high efficacy algorithm is powered ‘ON’ only if the first stage detects a seizure like event. Such a combination of monitoring and detection operation helps in reducing the average energy consumption per computation of the system. It will be subsequently shown in this chapter that the multi-stage technique also results in improved efficacy of detection as compared to individual algorithms. In this system, the first stage has been implemented using the “Coastline parameter” algorithm and the second stage has been implemented using the “DWT-QA algorithm” as described in the previous chapters.

4.1 Overview

In Chapter 2, a novel algorithm utilizing the wavelet transform and QA was described. It was also shown that the DWT-QA algorithm is not only low power as compared to other wavelet based algorithms but also has a high efficacy of detection. However, for a majority of the time in the period of operation, the seizure detection algorithms, typically, operate on neural signals which are devoid of any seizures. The occurrence of seizures is an intermittent event. Hence, during the time period when the neural signals are in the baseline phase, it becomes unnecessary to perform complex signal processing on the LFP data. Such processing results in redundant over-computations. Hence the highly efficacious and low-power consuming DWT-QA
algorithm described in the Chapter 2 is wasteful in terms of energy consumption over a long period of time. This is addressed by dividing the seizure detection operation into multiple stages.

The primary focus of the two-stage epileptic seizure detection system is to reduce the energy consumption of the system without affecting the efficacy of seizure detection. This is achieved at both algorithmic and circuit levels. At the algorithmic level, the redundant computations in a high-efficacy high-power algorithm are avoided by introducing a lower complexity monitoring stage using the “coastline” parameter algorithm. At the circuit level, the power hungry wavelet stage is conditionally powered down during the monitoring operation.

In the DWT-QA algorithm, it was shown that by selective choice of wavelet coefficients and weights for quasi averaging, it was possible to reduce the power consumption of the system [31]. However, during the baseline period of LFP recording, the DWT-QA is over-computing for sensing the “absence” of a seizure resulting in wasteful consumption of power. Such computational redundancy is overcome in the two-stage algorithm. At the circuit level, the system is operated at scaled power supply voltage which is close to the threshold voltage of the transistor. Such voltage scaling also helps in significantly reducing the energy consumption of the system. The methodology for detection, training of the algorithm and the near-threshold operation is explained below.

4.1.1 Algorithm

The two stage algorithm divides the seizure detection system into two stages viz.“Monitoring Stage” and “Detection Stage”. The monitoring stage consists of an algorithm of lower complexity and lower efficacy than the detection stage. The monitoring stage is also having a higher sensitivity. In this system, the “coastline parameter” based algorithm (described in Chapter 3) is used as the monitoring stage [36]. On the other hand, the detection stage is the DWT-QA algorithm (described in
Chapter 2) which detects the onset of the seizure with a high efficacy and consumes higher energy per computation [31].

Monitoring Stage

The monitoring stage consists of the coastline parameter based algorithm. The coastline parameter (CL) measures the physical trace length of the input signal. It computes the distance between adjacent peaks in the signals and accumulates this distance over a prefixed window of data [36] [38]. The mathematical representation of coastline parameter is given in Chapter 3 (Eq. (3.2)). The CL is computed over adjacent windows. If the windows consist of baseline signal, then the CL value will be on an average invariant as there would be no spiking feature in the signal. However, as the window approaches the onset of seizure CL value gradually increases due to increased spiking activity. The value of CL parameter for each window can be compared to the threshold prefixed for monitoring stage to raise a flag. This can be termed as a ‘warning flag’. The threshold is fixed in the training phase dependent on the required latency of detection and the frequency of occurrence of seizure in the training data. The warning flag is subsequently used to activate the detection stage which analyzes the signal by further processing it and classifies it as a seizure or otherwise. The monitoring stage is trained for high sensitivity. However it can also filter out a large number of seizure-like features which would, otherwise, be FP for the detection stage. Such FPs would result in wastage of energy in case of a single stage algorithm. However, in a two-stage algorithm, if an FP does get detected in monitoring stage, the highly efficacious detection stage can classify it correctly by processing the data in detail.

Detection Stage

The detection stage in the multi-stage algorithm consists of the DWT-QA algorithm as presented in Chapter 2 [31]. The warning flag generated in the monitoring
stage is used to activate the detection stage for further processing of the LFP data. The DWT-QA algorithm takes the digitized input data and decomposes it into narrow frequency bands in terms of wavelet coefficients. As mentioned in Chapter 2, the significant coefficients are identified in training phase for specific patient. These are then quasi-averaged and weighted over prefixed window size. The weighted quasi-averages are added to produce the detection signal which is compared with a prefixed threshold to generate the detect flag. The detection flag is used to classify the signal as seizure or baseline. It should be noted that the detection stage comes into operation only if the monitoring stage raises a warning flag over any seizure-like patterns. Hence for a majority of the time when the input LFP neural signal consists of baseline phase, the detection stage is inactive and the monitoring stage is operational. This helps avoid the redundant computations in the complex and power hungry DWT-QA algorithm during the baseline and hence saves the limited available energy. The algorithmic flow for the operation of the two stage seizure detection system is shown in the Fig. 4.1. The thresholds for both the monitoring and the detection stages as well as the coefficients and weights for the DWT-QA algorithm are fixed in the training phase.

4.1.2 Training

As described in previous chapters every seizure detection algorithm has user-specific parameters which need to be set by using a training data. A random set of LFP recordings is selected as the training data. It consists of four seizures and baseline in it. In order to have an equal probability of seizure and baseline, an equal duration of seizures and baseline is selected. The inclusion of baseline in the training data is essential to include motion artifacts and seizure like neural activity, which might trigger a false detection. The data is pre-classified into ictal (seizure) and baseline sections as described in Chapter 2. The training for the two stage algorithm is carried out stage by stage and not as a whole. This is because; it is desirable that
the first stage (monitoring) should be comparatively more sensitive to seizure like onsets. This is in order to prevent any FN and miss any seizures. The training is performed offline using Matlab simulations. In order to train the monitoring stage, the selected training data set is digitized and streamed in to the coastline parameter algorithm. The CL is calculated for each adjacent window of prefixed size (1024 in this research). In iterative procedure, the threshold is varied until such a threshold
is selected where all seizures are detected with minimum delay of detection. This might result in slightly increased FP for the monitoring stage due to high sensitivity. However this is essential in order to provide a time margin for the detection stage to compute and detect with high efficacy and low detection latency. It should be noted that the window size for calculation of the coastline parameter can be varied depending on the severity of the disorder in different patients, thereby making the algorithm user-specific.

The training for the detection stage is performed as explained in detail in Chapter 2 [31]. The training data used for training the monitoring stage and detection stage is identical to maintain the correlation between the monitoring and detection stage. The DWT-QA algorithm processes the training data in accordance to the methodology described in Chapter 2. The threshold for the detection stage is set by iteratively varying it to achieve maximum efficacy. The definition of the efficacy metric for the two stage algorithm is explained in detail in section 4.3. Apart from weights, wavelet coefficients and thresholds fixed in training phase, it is also necessary to fix the time interval for which the detection stage should operate. This is because once the monitoring stage has raised the warning flag, it is necessary to analyze the input data subsequently for a fixed period of time. This time is decided based on the average duration of seizure in the training data. Based on this, the operational time for the detection stage is chosen as ten seconds. This implies that once the warning flag is raised, the detection stage will analyze ten seconds of subsequent data to confirm or deny the onset of seizure. After this time interval, the detection stage will deactivate and the monitoring stage will be activated again to continue its operation. It should also be noted that these prefixed parameters are variable between subjects depending on the characteristics of LFP neural signals. For instance, if the average seizure duration for a subject is small, the detection stage has to be turned on for shorter time interval. This will also lead to energy savings as the detection stage consumes more energy as compared to the monitoring stage due to presence of DWT block.
4.1.3 Near-Threshold Voltage Operation

The two-stage algorithm for epileptic seizure detection is proposed to provide a practical solution, which not only maintains a high efficacy but also reduces the total energy consumption of the system. The total energy consumption of the system for a clock cycle is the sum of the dynamic and the leakage energy components as depicted in Eq. (4.1). Here \( C \) is the total switching capacitance and \( \alpha \) is the switching activity factor.

\[
\text{Energy/Cycle} = \alpha CV_{DD}^2 + V_{DD}I_{\text{Leakage}}T_{\text{Clock}}
\]  

(4.1)

where,

\[
T_{\text{Clock}} \sim \frac{CV_{DD}}{I_{ON}}
\]

The first term, representing the dynamic energy consumed, has a quadratic dependence on supply voltage \( (V_{DD}) \). It is the energy consumed when the system is performing the computational task. The second term represents the energy due to leakage current, which is exponentially dependent on \( V_{DD} \). This is the energy consumed when the system is on a standby and awaiting inputs. Based on these definitions, the energy consumption of the system can be significantly reduced by aggressively scaling \( V_{DD} \). Fig. 4.2 shows a typical plot of energy dissipation with respect to supply voltage \( (V_{DD}) \). As can be seen, the optimal voltage of operation would be at the point of minimum energy. This is very close to the threshold voltage of the transistor. At supply voltage below the optimal point, the energy is dominated by leakage current (delay increases exponentially) while the energy is dominated by the dynamic component when the supply is above the optimal point (linear delay improvement). The system is successfully operated at scaled \( V_{DD} \) as close to the threshold voltage of the technology as possible. This is practical in the case of seizure detection system because of very low frequency of operation. Due to reduced on-state current \( (I_{ON}) \) at scaled \( V_{DD} \), it is possible to achieve correct functionality by lowering the clock frequency (higher \( T_{\text{Clock}} \)). The neural LFP recording is sampled at a low frequency of 1.526 KHz, thereby making aggressive \( V_{DD} \) scaling practical.
4.2 Hardware Implementation

The block diagram for the hardware implementation of the two stage algorithm is shown in Fig. 4.3. The main components of the system are the monitoring stage (coastline parameter), the detection stage (DWT-QA) and the controller. The hardware description of these blocks and circuit techniques used to optimize them are explained in following section.

4.2.1 Monitoring Stage-Coastline Parameter Algorithm

The coastline parameter is calculated as given in the Eq. (3.2) [36]. The block diagram for the coastline parameter algorithm is shown in Fig. 3.3 (b). The functioning of this algorithm was described in detail in Chapter 3. The absolute magnitude
of the difference in neighboring data points is accumulated over the prefixed window size. This measures the trace length of the signal. A counter controls the windowing operation of the algorithm. After the required count is reached it resets the internal registers to start a fresh calculation for the subsequent window. It also generates the synchronizing pulse to output the accumulated value to the comparator input at the end of each window. The prefixed threshold from the training phase is used to threshold this value and raise the warning flag. The warning flag is passed on to the controller. Using the warning flag, the controller performs the power switching operation to activate the detection stage. The clock to this monitoring stage is provided through the controller. When the monitoring stage is deactivated, the corresponding clock is also gated. Due to use of clock gating, there is no switching at the input of the monitoring stage when it is deactivated. This is essential in power efficient implementation.
4.2.2 Detection Stage-DWT-QA Algorithm

The block diagram for DWT-QA algorithm is shown in Fig. 2.9. The functioning of this algorithm was explained in detail in Chapter 2. CSHM and CSE based implementation is used in the detection stage of the two-stage algorithm. As mentioned in Chapter 2, the user-specific parameters such as the wavelet coefficients, weights and thresholds are fixed in the training. The processed signal when compared to the prefixed threshold will give the final detection. The clock to the DWT-QA algorithm in the detection stage is also gated by the controller when the stage is deactivated. The final detection from the DWT-QA is sent to the controller, which controls the algorithm output of the system depending upon the stage that is activated.

4.2.3 Controller

The controller is the block which is used to activate and deactivate the two stages according to the multi-stage algorithm explained in the previous section. In this research, coarse power gating is used, wherein entire block representing the detection stage is completely turned ‘OFF’. In order to achieve that, power gating footer transistors are used. Appropriately sized NMOS transistors \(G_1\) and \(G_2\) are introduced between each of the stages and the ground terminal (Fig. 4.3). By applying control signals to the gate terminal of the footer transistors, the two stages can be turned ‘ON’ or ‘OFF’. The state of \(G_1\) and \(G_2\) is controlled by the power gating signals \(PG_1\) and \(PG_2\) which are generated by the controller. The warning flag generated by the monitoring stage is latched by the controller. This flag is used internally to generate the power gating signal \(PG_2\) for the detection stage. This turns on the \(G_2\) and completes the \(V_{DD} - V_{SS}\) path. The detection stage now starts computing on subsequent samples of input LFP data. Simultaneously, \(PG_1\) is pulled down thereby turning ‘OFF’ \(G_1\) and hence turns ‘OFF’ the monitoring stage. These gating signals are also used to generate control signals for clocks for corresponding stages. This ensures that each stage gets the clock signal only when it is activated and not otherwise. This is
to prevent switching signals at the inputs of each of the stages. The controller also
generates two reset pulses for each of the stages. These are used to reset the stage
just after it is activated. Such resetting operation ensures that no spurious outputs
occur from either stages. The timer load port in the controller is used to keep the
system programmable in terms of the time for which the detection stage should be
‘ON’. The timer is implemented using a counter which is activated along with the
detection stage and counts till the value loaded at the timer load port is reached. This
count represents the time interval for the operation of detection stage. At the end of
the time interval, it generates an internal signal which deactivates the detection stage
by powering it ‘OFF’ and restarts the monitoring stage by resetting it.

4.3 Efficacy and Efficiency Metrics

In the Chapters 2 and 3, efficacy of an algorithm has been calculated based on the
definitions of Sensitivity, Specificity and Average Detection Rate (ADR)(Eq. (2.4–2.6)
[31]. These parameters capture the occurrence of FP and FN accurately. Depending
on the value of ADR, an adequate intuition can be obtained about the performance
of an algorithm for a set of data. However, the definitions do not take into account
the latency of detection which is an important factor in seizure detection. A lower
latency of detection ensures that the subsequent therapeutic action is administrated
at an optimal time instant. Hence it is essential to optimize the system for low latency
of detection too. In this chapter, the multi- stage algorithm highlights the importance
of delay of detection. The monitoring stage is trained such that it has minimal delay
of detection (in the form of high sensitivity) and the detection stage is trained for
minimal FP and FN (in the form of high ADR). However, the efficacy of the combined
system is more meaningful if the total latency between the onset of the seizure and
the final detection of seizure is taken into account and optimized. Hence, the efficacy
is computed using modified definitions which include the delay of detection as a part
of FP and FN. According to these definitions, the FP and FN are measured in terms
Fig. 4.4. Detection of Epileptic Seizure with marked FP FN and Seizure

of percentage time of the neural signal instead of counting their instances. FP are false detection, which may occur because of seizure like events, are calculated as the percentage of the baseline time spent under FP when the detection flag is high. This is shown in the Fig. 4.4. In the multi-stage algorithm, the monitoring stage computes and filters out a significant number of seizure-like events which could have caused a false detection in the detection stage. This is due to high sensitivity of the monitoring stage. However, if the detection stage is activated due to the warning flag generated from the monitoring stage, then the possibility of a false positive exists depending on the training and efficacy of the detection stage. On the other hand, FN are recorded in the case of missing a seizure. FN are highly undesirable in terms of functionality due to the critical nature of operation and should be kept to a minimum. According to the modified definition, FN is calculated in terms of percentage of the time in seizure phase of the LFP signal, that the seizure is not detected. As can be seen in
Fig. 4.4, the delay in detection of seizure after its onset, is included as time spent under FN. A larger delay implies a higher percentage of time spent in FN which can eventually be 100% in case of a missed seizure. Ideally, the detection flag should be raised up at the same instant that the onset begins. Eq. (4.2–4.6) summarize the method of computing efficacy taking the latency of detection into consideration.

\[
\%FP_{\text{time}} = \frac{\sum t_{FP}}{t_{\text{baseline}}} \tag{4.2}
\]

\[
\%FN_{\text{time}} = \frac{\sum t_{\text{delay}}}{t_{\text{seizure}}} \tag{4.3}
\]

\[
\text{Sensitivity} = 1 - (\%FP_{\text{time}}) \tag{4.4}
\]

\[
\text{Specificity} = 1 - (\%FN_{\text{time}}) \tag{4.5}
\]

\[
\text{Efficacy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{4.6}
\]

Based on the algorithms described in Chapters 2 and 3, it is known that the goal of any seizure detection system is to minimize FN as well as FP. However, these are closely correlated to each other especially in thresholded systems. An increase in threshold may reduce FP but increase FN and vice versa. Hence in a two stage algorithm, the monitoring stage is trained such that, it is highly sensitive to seizure like event. This results in very low FN and hence, delay of detection. Individually, the monitoring stage will have an increased FP. In contrast, the detection stage is trained for a very high efficacy in order to reject majority of the false detections from the monitoring stage. The efficacy of the system is calculated as the average of the Sensitivity and Specificity, which is similar to the definition of the ADR. This metric captures the performance of the system and can be used to compare with other algorithms for justifying its usage. Due to implantable nature of the system, apart from the efficacy, it is also important to measure the hardware efficiency of the system. The total power consumption of the hardware used to implement the algorithm is to be kept at a minimum. However, the total power consumption may not give a complete picture of the implementation. This is because, the seizure detection system is a highly data dependent application. Hence, the energy efficiency of the
system should be measured in terms of average energy consumed per computation. This metric can be subsequently compared in the baseline as well as the seizure part of the LFP recording. Based on this comparison, a fair judgment can be made about the algorithm and its implementation in terms of both functionality and practicality.

4.4 Results

The multi-stage algorithm for detecting epileptic seizure was designed using two stages and implemented using 65 nm bulk-Si (TSMC) library. The operation of the algorithm was verified by comparing the software implementation and the hardware simulation using classified data as in Chapter 2 and Chapter 3. Based on the new definitions of FP and FN as described in the previous section, the system was evaluated on two aspects viz. Algorithmic Efficacy and Hardware Efficiency.

4.4.1 Algorithm Efficacy with minimal detection latency

The efficacy results are plotted in Fig. 4.5 for five different LAS subjects. The efficacy shown for the individual Coastline and DWT-QA algorithm is the highest achievable using the identical selected training data. In Fig. 4.5, the proposed methodology of two stages is compared with the individual stages in term of efficacy. In addition, a third case is also included for efficacy comparison wherein the two stages are logically ANDed (Boolean combination proposed in Chapter 2. This implies that the seizure is detected only if both the algorithms have classified it as a seizure. As was mentioned in Chapter 2, Boolean operation can be used to combine the detection from two individual algorithms to increase the overall efficacy of detection. It can be observed that the efficacy of the individual algorithm varies over different subjects as expected. However, the two stage algorithm increases the efficacy on an average by 12 % as compared to the DWT-QA based algorithm. This improvement is in conjunction with taking into account, a minimal delay in detection. This is due to the modified definitions of FP and FN for optimizing the system for user-specific param-
Fig. 4.5. Algorithmic Efficacy for the Two-Stage Algorithm and its comparison with individual stages and logical combination

eters. In some of the LAS subjects it can be seen that the DWT-QA algorithm may have a comparable efficacy as compared to the two stage algorithm. In spite of that, it is preferable to use the two stage algorithm because of the significant advantage in terms of energy consumed as compared to the individual DWT-QA algorithm.

4.4.2 Hardware Energy Efficiency

Hardware Efficiency of the two stage system is listed in Table 4.1. A typical data set used for testing the system is also tabulated. Based on that, a typical test data set consists of 12 seizure events, each one about 180 seconds in duration. The multi-stage seizure detection algorithm is implemented using VHDL and synthesized using standard synthesis tool such as Synopsys Design Compiler. The power consumption of
Table 4.1.
Hardware Efficiency for Typical Test Data

<table>
<thead>
<tr>
<th></th>
<th>Typical Test Data</th>
<th>Power and Energy Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( V_{DD} = 500 \text{ mV}, f = 100 \text{ KHz} )</td>
</tr>
<tr>
<td>Total Duration</td>
<td>3274 sec</td>
<td>Monitoring Stage (Coastline) 3.275( \mu \text{W} )</td>
</tr>
<tr>
<td>Baseline Duration</td>
<td>1040 sec</td>
<td>Detection Stage (DWT-QA) 69.78( \mu \text{W} )</td>
</tr>
<tr>
<td>Marked Seizures</td>
<td>2194 sec</td>
<td>Controller 0.01( \mu \text{W} )</td>
</tr>
<tr>
<td>Avg. Seizure Duration</td>
<td>180 sec</td>
<td>System Power 79.37( \mu \text{W} )</td>
</tr>
<tr>
<td>False Detection</td>
<td>40 sec</td>
<td>Energy Consumed (Baseline) 2.15 nJ/comp</td>
</tr>
<tr>
<td>No. of Marked Seizures</td>
<td>12</td>
<td>Energy Consumed (Seizure) 45.73 nJ/comp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average Energy Consumed 31.2 nJ/comp</td>
</tr>
</tbody>
</table>

the system is simulated using Nanosim simulator for circuit level simulation. During the baseline period, the monitoring stage is powered ‘ON’ and the detection stage is powered ‘OFF’. Thus the energy consumed from the battery is used only for computations in the low-power coastline parameter algorithm. This energy is simulated at 2.15 nJ per computation. Whenever there is a flag raised by the monitoring stage due to any seizure-like event, the detection stage comprising of the DWT-QA algorithm is powered ‘ON’ and the monitoring stage is powered ‘OFF’. The energy consumption now increases to 45.73 nJ per computation as expected. This is because of the complex DWT computation in the in the DWT-QA algorithm. The detection stage operates for a predetermined time, which is ten seconds in this case. Subsequent to this time period, the system returns to monitoring operation and the energy consumption reduces again to that corresponding to the coastline parameter algorithm. In comparison the if a single stage DWT-QA algorithm was used (as in Chapter 2), it would be computing at all times. This would effectively consume 45.73 nJ per computation taking into account the minimal delay of detection. The average energy consumed by the two-stage system is data-dependent on the number of seizures and
the duration of baseline. For the tabulated sample data, the average energy was simulated at about 31.77% less than that consumed by DWT-QA stage alone. The reduced energy consumption was also assisted by the reduced power supply. The system was operated at 500 mV and 100 KHZ. Voltage scaling along significantly reduces the energy consumption due to quadratic dependence of power on the supply voltage. The low frequency of operation of the system further supports the aggressive voltage scaling involved. Apart from voltage scaling, power gating greatly reduces the energy consumption of the system as seen from the hardware efficiency results.

4.5 Conclusion

In this chapter, a power-efficient and efficacious methodology to detect the onset of an epileptic seizure was developed and implemented in hardware [43]. The proposed two stage method helps in significant reduction of energy consumption per unit computation because of leakage control techniques such power and clock gating for each individual stages. Aggressive voltage scaling (500 mV) results in low power operation. It also helps in achieving energy optimality, thereby significantly prolonging the battery life of the implant. The multi-stage algorithm is programmable to individual user and hence can be used over a wide range of inter-subject variability. It should also be noted that the proposed design methodology is also applicable using other seizure detection algorithms in the two stages. This is possible as long as the underlying principle for algorithm selection is maintained. The improvement in hardware efficiency will be dependent on the fact that the monitoring stage has lower energy consumption than the more elaborate detection stage. It is also possible to enhance the programmability by providing multiple algorithms to choose from in each of the stages. By combining the methodologies described in previous chapters, the multi-stage methodology is a significant step towards development of practical implantable epilepsy prosthesis.
5. LOW-POWER SYSTEM FOR DETECTING SYMPTOMATIC PATTERNS IN NON-SPEECH ACOUSTIC SIGNALS

The field of wearable electronics necessitates a very good energy efficiency in terms of consumption of limited battery power. In this chapter, the algorithm-circuit co-design strategy is extended to be applied to systems which are usable in wearable health monitoring products. A low-power, efficacious and scalable system for detection of symptomatic patterns in biological audio signals is developed. The digitized audio signals denoting various health related symptoms such as cough, sneeze etc. are frequency analyzed using DWT. Subsequently, processing techniques such as mathematical metrics described in Chapter 3 are also used to process the signal. Furthermore, Mel frequency cepstrum based analysis is applied to distinguish between signals which have closely matching frequency response. The system is designed to be power efficient, efficacious and scalable by application of algorithm-circuit co-design methodology.

5.1 Overview

In Chapters 2–4, using algorithm-circuit co-design, power efficient and efficacious systems were developed and illustrated for an implantable biomedical application. Various algorithms and circuit level techniques were used for enabling a solution for alleviating an intractable disorder like epilepsy. The signal processing technique of DWT was used to filter out unwanted frequencies and resolve the seizure signal into its constituent frequency components. In Chapter 3 various mathematical and statistical metrics were described which could detect the seizure with high degree of accuracy. In this chapter, these techniques are used in a wearable system design. Additionally, another signal processing tool, viz. Mel frequency cepstrum is also
used in conjunction. These signal processing methods are used together and applied
for detection of symptomatic acoustic patterns in human non-speech audio signals.
Apart from the DWT and mathematical metrics described in previous chapters, the
Mel-frequency cepstrum is also used in order to analyze the acoustic signals. The
Mel frequency cepstrum based detection of acoustic patterns is based on the Mel
scale which is a perceptual scale of pitches proposed in 1937 [45]. Depending on
the type of acoustic signal and the characteristics of the symptomatic pattern to
be detected, the correct signal processing tool is used. The selection of the signal
processing technique also is optimized with respect to the power consumption of the
hardware implementation. In the subsequent sections the algorithm methodology and
the hardware implementation is discussed in detail.

5.1.1 Algorithm Methodology

In this section, the algorithm used to detect multiple types of biological acoustic
symptoms in streaming non-speech human audio signals is described. The algorithm
methodology is depicted in Fig. 5.1. The input data used for this algorithm is the
digital audio recording of human audio symptoms such as cough, sneeze etc. This
data is streamed into the algorithm at its sampling frequency of 11.025 KHz. The
various steps of the algorithm are depicted in Fig. 5.2. It can be seen from Fig. 5.1 and
Fig. 5.2 that the input signal is processed using DWT in the first step of the algorithm.
The wavelet coefficients of interest are then identified. These are the coefficients
that show significant activity for their corresponding type of input pattern to be
detected. These coefficients are subsequently processed using metric based processing
or cepstrum based processing depending on the characteristic of the acoustic pattern
in the input audio signal. The processed signal from these stages are then compared
against prefixed thresholds in order to raise the detection flags. These stages of the
algorithm and the justification for their usage are explained in detail in this section.
Majority of the ailments affecting a large number of people worldwide are detectable using common acoustic symptoms. These audio signals are sufficient for detecting a multitude of disorders ranging such as respiratory disorders, digestive disorders etc. The acoustic signals identified in this research to be used for detection of symptoms of degrading human health are cough, sneeze, belch, wheeze and vomit. These five signals can be used for early detection of various health issues as shown in Fig. 5.3. These signals are sampled at 11.025 KHz and available in ‘.wav’ format. In order to test the algorithm, these signals can be streamed in digital 2’s comple-
ment format into the input. It can be noticed from Fig. 5.3 that each of the signal under consideration has a specific pattern. This corresponds to a distinct frequency response. Taking advantage of this characteristic of the signal, the input is resolved into its frequency spectrum. This is achieved using the wavelet transform.
Wavelet transform for spectral analysis

The input data corresponding to the acoustic signals mentioned in the previous section are digitized and processed using the wavelet transform. Using DWT the acoustic signal is subjected to spectral decomposition. As discussed in detail in Chapter 2, the multi-resolution property of the DWT accurately resolves the input signal into narrow frequency bands or wavelet coefficients $D_i$. The mother wavelet used for computing the wavelet transform is the Daubechies $4^{th}$ order wavelet (Fig. 2.4). The acoustic signals under analysis are best represented by this wavelet function due to their similarity of the coarseness and smoothness. The use of DWT is based on the observation that specific acoustic symptomatic patterns occur in specific frequency bands. For instance, the acoustic patterns corresponding to wheezing and vomiting are resolved in the $D_5$ and $D_6$ wavelet coefficients respectively. Fig. 5.4 shows the sample spectral decomposition for the ‘Wheeze’ signal. Similarly, the pattern consistent with burp/belching is found in multiple coefficients ($D_4$ and $D_5$). However, the cough and sneeze signals have a similar frequency spectrum and are resolved into single DWT coefficient ($D_3$). Subsequent to the signal decomposition, the spectral and the temporal information of the signal is used for further processing in the algorithm. Although, the symptomatic patterns are resolved into the wavelet coefficients, it is necessary to smooth out some of the noisy spikes which might result in false classification. Depending on the characteristic of the pattern to be detected, these DWT coefficients are subjected to various mathematical metric based computation and Mel- frequency cepstrum coefficient (MFCC) based computation.

Mathematical Metric Based Computation

The DWT coefficients from the previous steps are identified and marked for various symptomatic patterns. Based on the property of the pattern being detected, the energy, coastline and quasi-averaging mathematical metrics are applied for processing the DWT coefficients. These mathematical metrics were discussed in detail in Chapter
2 and Chapter 3. In this section, these metrics are revisited with justification for their use in the symptomatic acoustic pattern detection algorithm.

**Energy parameter for ‘Vomit’ pattern**

It is observed that in the case of the DWT coefficient $D_6$, which corresponds to a frequency band of 86 Hz–172 Hz, there is significant increase in activity in the case of acoustic signal input for vomiting sound. The nature of this signal is that there is a continuous and a substantial increase in the amplitude over a certain period of time as seen in Fig. 5.5. This property of the signal in a particular frequency bands makes it an ideal candidate for energy metric based detection. The definition of energy metric is mathematically represented as Eq. 3.1 in Chapter 2. The energy is calculated over adjacent windows of size N and is compared against a threshold value.
Fig. 5.5. Continuous increase in amplitude in ‘Vomit’ pattern

to mark a ‘vomit’ pattern detection. The window size and threshold are fixed in the training phase. Excessive occurrence of this symptom indicates an impending health disorder such as ear infection, diarrhea or whooping cough [26]. Early detection of this symptom can be very advantageous for a patient.

Coastline parameter for ‘Wheeze’ pattern

The DWT coefficient, $D_5$, corresponds to a frequency band of 172 Hz–344 Hz. It is observed that the pattern pertaining to wheezing is resolved into $D_5$. This is seen from Fig. 5.4. The signal is characterized by frequent repetition of a symptom specific pattern over a short period of time. The pattern has more or less a constant amplitude after the initial increase at the onset. The mathematical metric which is best suited to detect this type of pattern is found to be the coastline parameter. In Chapter 3, the coastline parameter was discussed in detail. The mathematical expression for the coastline parameter is shown in Eq. 3.2. The coastline parameter was also used in the multi-stage algorithm developed in Chapter 4. The window size
selected for the coastline parameter in this algorithm is equivalent to 1024 samples of input audio data. This corresponds to 32 samples of the $D_5$ coefficient. The window size is fixed in the training phase of the signal.

**Quasi-Average for ‘Belch/Burp’ pattern**

In the resolved acoustic signal, it is also observed that the artifacts corresponding to belching or burping symptoms are found in two DWT coefficients, viz. $D_4$ and $D_5$. In Chapter 2 quasi-average was used to detect a pattern when it was observed in multiple wavelet coefficients. This ensured that the frequency information in all sub-bands of interest are taken into account. Similarly, in this algorithm, the quasi-average parameter is used to detect the belch/burp pattern occurring in multiple wavelet coefficients. The definition of quasi-average and its comparison with the standard definition of average is discussed in Chapter 2. The mathematical equation is described in Eq. 2.3. The window size for each coefficient is equivalent to 1024 samples of input acoustic data. The quasi average of $D_4$ and $D_5$ coefficients are weighted in order to normalize them and subsequently added. The belching symptom which, in excess, is indicative of digestive health issues [26] can be detected by comparing the weighted quasi-average with a prefixed threshold. The weights and thresholds values are fixed in the training phase.

**Mel Frequency Cepstrum Coefficient Based Computation for ‘Cough’ and ‘Sneeze’ pattern**

Apart from the three types of symptoms described above, the DWT also resolves the ‘cough’ and ‘sneeze’ signals into their respective frequency bands. In the DWT output, it is observed that the patterns pertinent to the ‘cough’ and ‘sneeze’ symptoms occur in the same wavelet coefficient, viz $D_3$. This is because, the frequency spectrum of these two symptomatic signals are very similar. This is evident from Fig. 5.6. Due to this similarity, it is not possible to distinguish between these two
Fig. 5.6. Frequency spectrum for typical cough and sneeze signals

types of symptomatic patterns by using the mathematical metrics described previously. The mathematical metrics will operate on the physical aspect of the signal, viz. amplitude and will still render them indistinguishable. In order to address this issue in this algorithm, the $D_3$ coefficient is processed using an advanced signal processing technique based on Mel frequency cepstrum. The detail description of the Mel cepstrum algorithm and the necessary modification needed for the purpose of this application is discussed in the subsequent section.
Mel scale and Mel frequency cepstrum coefficient

The term ‘Mel’ stands for melody and is a perceptual scale for pitches [45]. It is based on the biology of the human auditory and speech emitting systems. The human auditory system is incapable of discerning between closely spaced frequencies, especially at higher frequencies (> 1 KHz). The Mel scale maps this non-linearity on the frequency scale. By fixing a reference point, where 1000 Mels is same as 1000 Hz, the frequency resolution is made almost linear below 1 KHz and logarithmic at higher frequencies. This implies that there is a linear transfer of power at lower frequencies and a non-linear transfer at higher frequencies. This can be seen in Fig. 5.7. The conversion between Mel scale and frequency scale (in Hz) is computed using Eq. 5.1 and Eq. 5.2 [45].

\[
M(f) = 1125 \times \ln\left(1 + \frac{f}{700}\right) \tag{5.1}
\]

\[
M^{-1}(m) = 700 \times \left(e^{\frac{m}{1125}} - 1\right) \tag{5.2}
\]

Also, the human audio signals are heavily modulated by the shape of the vocal tract including the tongue, teeth etc. This shape is encoded in the acoustic signal in the
form of power spectrum envelope over a short period of time. Mel frequency cepstrum coefficients (MFCC) are a set of coefficients extracted from an audio signal which represent this short-time power spectral envelope. MFCC is extensively used in speech or speaker recognition [46] [47]. MFCC have been shown to accurately represent the short-time power spectral envelope which encodes the shape of the vocal tract [48] [13]. The basic algorithm for computing MFCC is shown in Fig. 5.8. As can be seen from the Fig. 5.8, the first step is to compute the windowed power spectrum (FFT or STFT) of the signal to resolve into its component frequencies. The envelope bounding the power spectrum has information about the physical shape of the source of origin of the signal. In order to extract this information, the power spectrum is subjected to filtering operation using Mel filters. These filters are the primary component of the MFCC and form the Mel filter bank. The Mel filter bank consists of a set of overlapping band pass filters, whose center frequency is uniformly spread across the Mel scale as shown in Fig. 5.9 [48]. Just like the human auditory system, the Mel filter bank clump the frequencies that are closely spaced around the center frequency of each of the filters. The Mel filters with triangular frequency response are also shown in Fig. 5.9. Based on Eq. 5.1 and Eq. 5.2, the uniformly spread center frequencies of the Mel filters transform to a logarithmic spacing on the frequency scale. This transformation is coherent with the fact that the human cochlea cannot discern the difference between two closely spaced frequencies, especially at higher frequencies. The clumped frequencies in each of the Mel filters has a certain spectral power which corresponds to the shape of the power spectral envelope. In order to capture this in the MFCC, the spectral energy in each of the Mel filters is calculated. The Mel

Fig. 5.8. Block diagram for standard MFCC algorithm
filter energies are subsequently processed by a logarithm block for a non-linear normalization. This corresponds to the human auditory system which does not linearly amplifies the sounds of varying amplitudes. This implies louder sounds are amplified less than fainter sounds. Since the filters are overlapped, there is significant correlation between the spectral Mel filter energies. The discrete cosine transform (DCT) block helps in de-correlating the energies in these overlapping band-pass filters. The output coefficients of the DCT block are the Mel cepstrum coefficients. The term cepstrum denotes the operation that it is calculation of ‘spectrum of a spectrum’.

Generally 26-40 filters are used in the Mel filter bank generating as many Mel coefficients. Depending on the complexity of the speech pattern to be detected the specific Mel coefficients can be analyzed. Generally, the first 10-12 MFCC are sufficient for speech recognition. However, in this research, we are interested in analyzing two distinct non-speech symptomatic patterns and hence need a reduced number of MFCC.

In the following section we will describe the modification made to the standard MFCC algorithm to be used in the acoustic pattern detection.
MFCC Based Computation

As described in the previous section, the MFCC algorithm is designed to decode speech patterns which have a high degree of variability. However, in the proposed algorithm, two types of patterns are to be classified. Hence the over computation in the standard MFCC algorithm can be reduced by modifying it as described in this section. The wavelet coefficient of interest, $D_3$, consisting of both the cough and sneeze signals is resolved by the DWT operation. Therefore, it is not necessary to compute the FFT as shown in Fig. 5.10. Traditionally the entire bandwidth of acoustic signal is divided into approximately 26 Mel filters. However, since $D_3$ represents a filtered version of the input acoustic signal, it is only necessary to have the Mel filters which overlap in the corresponding $D_3$ frequency band (689-1378Hz). This is illustrated in Fig. 5.11. This results in reduced number of Mel filters (3 in this case). The spectral energies of these filters are computed over a predetermined window size. The logarithm block is omitted in the modified MFCC algorithm (as shown in Fig. 5.8). This is because, the nonlinear normalization is useful, if the entire bandwidth of acoustic signals is to be analyzed for numerous patterns. However, since only 3 Mel filters are used to detect the 2 patterns (‘cough’ and ‘sneeze’), the logarithm block is not effective and hence can be removed. This manifests into a reduced power consumption in the hardware implementation. The filter energies are passed on to a DCT block which de-correlates them producing the modified cepstrum coefficients. Three coefficients are obtained from the DCT block. The first coefficient corresponds to the DC component and can be ignored. The second and the third coefficients of
the DCT correspond to the coefficients which can separate the cough pattern from the sneeze. The second coefficient corresponds to the ‘cough’ pattern and the third coefficient resolves the ‘sneeze’ pattern. These coefficients can be compared to a prefix threshold to raise the corresponding detection flag.

5.1.2 Threshold and Training

Thresholding operation is used to raise the detection flag for various symptomatic patterns. The processed acoustic data from the mathematical metric computations and the mel cepstrum computation is compared with prefixed threshold values to cause detection. Each symptom is independently detected. These threshold values are fixed based on the training data which represents a typical case for each of the type of symptoms that are being detected. Based on the above description of the algorithm methodology, it is evident that the proposed algorithm has a number of parameters which are user specific such as thresholds, weights, coefficients of interest etc. In this research, five types of acoustic patterns corresponding to the symptoms indicative of general health have been used for detection. However, it is possible to increase the number of symptomatic patterns that can be detected. The methodol-
ogy described above is generic in nature such that, it can be applied to other audio biological signals as well. This necessitates a proper training to select optimal wavelet coefficients of interest, set appropriate weights and thresholds in order to have efficacious functionality. A set of data containing various signal patterns to be detected is used as the training set. This data set is subjected to the algorithm described above. The wavelet coefficients are identified and depending on the nature of the signal, the corresponding mathematical metrics are selected to process the coefficients. The thresholds for each of the processed signals are set such that they give maximum efficacy in terms of accurate classification. The windowing operation in various blocks assumes the window size to be equivalent to 1024 samples of audio input data. In the next section, the above described algorithm, will be implemented into a hardware. Low-power methodologies are used to implement a power efficient system.

5.2 Hardware Implementation

In this section, the circuit level techniques that are used to implement the proposed algorithm into a power-efficient hardware are discussed. As explained in the previous section, certain design choices at algorithm level of design abstraction were made in order to facilitate the low-power implementation of hardware. Fig. 5.12 shows the block diagram of the system. These individual blocks are discussed in detail in the following section.

5.2.1 Wavelet Decomposition Block

The wavelet transform block is the most computationally intensive block in the system and consumes a significant amount of power. There are various methods available in literature to implement the DWT block [32]. As described in Chapter 2, the Mallat algorithm is the optimal choice for low power implementation of DWT block where the input signals are streaming in serially [29] [30]. The Mallat algorithm and its hardware implementation is discussed in detail in Chapter 2 and Fig. 2.3. As
explained in the algorithm methodology, for the purpose of this application, wavelet coefficients $D_4$ through $D_6$ are of interest. Hence, it is necessary to derive six wavelet coefficients. This necessitates six stages of wavelet transform, thereby, requiring six cascading stages of H and G (Fig. 2.3). Since the five acoustic patterns are to be detected, it is required to have five H filters and four G filters. All these filters are of 8th order due to implementation using the Daubechies 4th order mother wavelet. Since the standard implementation of these nine filters would be computationally intensive in terms of number of multiplications, multiplier-less technique of Computation Sharing Multiplier (CSHM) and Common Sub-expression Elimination (CSE) is utilized to reduce power consumption. This is the same hardware methodology used in epileptic seizure detection system in Chapter 2 [33] [34]. The down sampling operation is performed by halving the clock frequency for every successive stage. This is implemented
using a counter. The wavelet coefficients are amplified due to successive convolution operation and are normalized.

### 5.2.2 Mathematical Metric Blocks

The mathematical metrics used in this algorithm are similar to the ones used in the multi-algorithm seizure detection processor in Chapter 3. The block diagrams for the mathematical metric blocks are shown in Fig. 3.3 and Fig. 2.12. The energy parameter is computed according to the Eq. 3.1. The block diagram for the computation of energy is shown in Fig. 3.3 (a). It consists of a multiply and accumulate operation, which sums the squared value of the input, viz. $D_6$ coefficient. The $D_6$ window size is chosen in the training phase and corresponds to 1024 samples of the digitized input data. The average energy value is then compared against the threshold to detect acoustics pertaining to vomiting sound. The coastline parameter block diagram is shown in Fig. 3.3 (b). The $D_5$ coefficient is the input to the coastline block. The input is delayed by a clock cycle in order to calculate the difference between two adjacent samples. The magnitude of the difference is accumulated over a prefixed window in order to calculate the trace length of the signal. This accumulated value is then compared with the threshold for detecting wheezing. Since wheezing signal is periodic signal for time duration without any significant increase in amplitude, the coastline parameter captures this pattern accurately. The block diagram for the quasi-averaging circuit is shown in Fig. 2.12. In order to enable a memory-less implementation and a continuously moving average, the average calculated in the previous window is subtracted from the sum of the running window instead of the individual data sample. Since the window size is a power of two, the divider is implemented by discarding the appropriate least significant bits. The quasi average is calculated over two coefficients viz. $D_4$ and $D_5$. The weights are used to normalize the magnitudes of the two coefficients. The weighted sum of quasi averages is then compared with a pre-fixed threshold to detect occurrence of belching or burping.
pattern. Thus, three of the symptomatic pattern are detected using these detection flags.

5.2.3 MFCC Based Computation Block

In the MFCC based computation block, three overlapping band-pass filters are used in the Mel filter bank. These filters correspond to the bandwidth of the $D_3$ wavelet coefficient (689-1378 Hz). The Mel filters are designed for a triangular magnitude response around the center frequency. These filters are of 16th order so that the frequency response is closely matching the required triangular response. The coefficients of these filters are adjusted by reducing the number of 1s. This reduces the number of computations without adversely affecting the frequency response of the filter. These filters are also implemented using CSHM and CSE methodologies in order to reduce the power consumption of the filter. The coefficients of all the filters are successfully represented using three alphabets for pre-computation. The output is subsequently passed to the energy block to calculate the spectral energy in each of the Mel filters. The three filter energies in each accumulation window are passed to the discrete cosine transform (DCT) block. Due to the overlapping nature of Mel filters, the outputs are highly correlated. The DCT de-correlates these filter outputs and separates the spectral envelope into multiple MFCC based parameters. The DCT block is also implemented by modifying the coefficient matrix in order to reduce the number of 1s and facilitate the CSHM based implementation [33] [34]. The first output coefficient of DCT corresponds to the DC component and can be ignored. The second and third coefficient corresponds to cough and sneeze pattern respectively. These two coefficients can be compared to a threshold in order to raise the detection flag.
5.2.4 Threshold Block and Clocking Circuitry

The threshold block consists of registers which are loaded with the pre-fixed threshold values corresponding to each individual acoustic pattern to be detected. These threshold values are fixed in the training phase. Comparators in the threshold block are used to compare and raise an independent detection flag for each of the symptomatic pattern detected. The clock circuitry is used to synchronize all the operations in the system. The input is streamed in at 11.025 KHz. Each successive coefficient of the wavelet transform is computed at half the frequency as that of the previous coefficient. Each of the successive blocks operates at the same frequencies as the coefficients they have as the input. A 10-bit counter is used as a clock divider circuit in order to synchronize the operation of the system.

5.3 Results: Efficacy and Hardware Efficiency

Based on the algorithm described in section 5.1 and the hardware implementation in the section 5.2, the system for detecting symptomatic patterns in non-speech audio signal was simulated. A total of 74 recordings of various acoustic symptomatic patterns were used for testing the accuracy of detection. These recordings consisted of five types of patterns viz. cough, sneeze, belch, vomit and wheeze. The audio recordings were downloaded from readily available sound library [26] [49]. These recordings were in the ‘.wav’ format. Apart from the testing signals, a set of data, consisting of 30 recordings was used in training phase to determine various system specific parameters [49]. The algorithm was designed in software to verify its efficacy and functionality before translating it into hardware. The digital audio signals from the recordings were processed in Matlab, according to the algorithm described in Section 5.2. The functionality of the algorithm was verified and efficacy of the algorithm was calculated. The cough and the sneeze signals, which require the MFCC based computation, were successfully classified using the two coefficients. A sample result for MFCC based classification is shown in Fig. 5.13 and Fig. 5.14.
As can be seen the second MFCC based coefficient is sensitive to the cough pattern while the third coefficient is sensitive to the sneeze pattern. The first coefficient is ignored and hence is not shown in the figure. It can be seen from Fig. 5.13 and Fig. 5.14.
Table 5.1.
Classification accuracy for five acoustic symptomatic patterns

<table>
<thead>
<tr>
<th>Input</th>
<th>DWT Coeff.</th>
<th>Processing</th>
<th>% Classified As</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cough</td>
</tr>
<tr>
<td>Cough</td>
<td>$D_3$</td>
<td>MFCC based</td>
<td>90.3%</td>
</tr>
<tr>
<td>Sneeze</td>
<td>$D_3$</td>
<td>MFCC based</td>
<td>16.7%</td>
</tr>
<tr>
<td>Belch</td>
<td>$D_1, D_5$</td>
<td>Quasi-Average</td>
<td>0%</td>
</tr>
<tr>
<td>Wheeze</td>
<td>$D_5$</td>
<td>Coastline</td>
<td>0%</td>
</tr>
<tr>
<td>Vomit</td>
<td>$D_6$</td>
<td>Energy</td>
<td>12.5%</td>
</tr>
</tbody>
</table>

Fig. 5.14 that with the onset of the symptom, the MFCC coefficient corresponding to the symptom gradually increases in amplitude. Using an appropriate threshold, it is possible to detect both cough and sneeze symptom from the same wavelet coefficient. The classification accuracy for all the symptomatic pattern is calculated as the percentage of signals classified correctly. These results are summarized in Table 5.1. Since the acoustic intensity follows an inverse square law relationship with respect to any background noise, the recorded signal will not be disturbed significantly by the background noise [26]. The use of DWT as the first stage also helps in filtering out any unwanted frequencies before further processing of the signal for detection. It is evident from the results that the MFCC based processing results in 90% correct classification. The cough and the sneeze signals have a good rate of classification and hence the use of a MFCC based computation is justified. Other acoustic symptoms such as wheeze and burp/belch have perfect classification. In the acoustic pattern pertaining to vomiting, the accuracy is observed to be the lowest at 75%.

The hardware implementation of the system was described using VHDL and synthesized using Synopsys tools in TSMC 65 nm technology bulk-Si library. The system was optimized for 1.0V $V_{DD}$ and 100 KHz $f_{CLK}$. The extracted circuit was simulated using Nanosim with a sample test data to get the power consumption of the system.
Table 5.2.
Power and Area of Hardware Implementation

<table>
<thead>
<tr>
<th>Hardware Implementation and Simulated Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
</tr>
<tr>
<td>$V_{DD}$</td>
</tr>
<tr>
<td>$f_{CLK}$</td>
</tr>
<tr>
<td>Area ($\mu m^2$)</td>
</tr>
<tr>
<td>Dynamic Power</td>
</tr>
<tr>
<td>Leakage Power</td>
</tr>
</tbody>
</table>

The 10-bit digital data was streamed into the system and the output verified for correct operation. To lower the power consumption of the system, $V_{DD}$ was scaled to 700 mV. Due to the quadratic dependence of the dynamic power on the power supply, $V_{DD}$ scaling reduces the dynamic power significantly. The system power was observed to be leakage dominated. Leakage control techniques can be used to further reduce power consumption. The simulated power and area are tabulated in Table 5.2. The algorithm is designed to be scalable to other acoustic biological signals. Depending on the frequency spectrum of the signal of interest and the pattern to be detected, the wavelet coefficients can be calculated to even more than six stages. The MFCC based parameters can be used for detecting any signals which occur in the same wavelet coefficients. Algorithm-circuit co-design methodology can be utilized to optimize power consumption and maintain high efficacy.

5.4 Conclusion

In conclusion, a generic system based on wavelet transform, mathematical metrics and Mel cepstrum based analysis has been developed to detect symptomatic patterns in audio biological signals [50]. Modifications in the algorithm and use of low power methodologies to implement the algorithm into circuit, enable the design of a low
power system. The system can be scaled to include other health markers and also can be made user specific. The MFCC based processing which is generally used for speech or speaker recognition has been shown to successfully distinguish signals that share the frequency spectrum. The algorithm shows a high classification rate (≥ 75 %) with a low power implementation [26].
6. SUMMARY AND FUTURE DIRECTION

In this research, we have seen the application of algorithm-circuit co-design methodology for developing circuits for two main types of biomedical applications, viz. implantable systems and wearable systems. In Chapter 2 we discussed the importance of DWT and the reason why it is a better choice for processing the streaming LFP as compared to FFT or STFT. We also defined the ‘quasi-averaging’ as a modified definition of moving average. The combination of these two signal processing methods was used to develop a novel algorithm, which was shown to be the best fit for a low power implementation for detecting epileptic seizures. It was shown to be user-specific and high in efficacy as well as power efficient as compared to existing circuits. In Chapter 3, a multi-algorithm seizure detection processor was implemented in 65 nm silicon technology. The design was programmable, thereby, making it a user scalable system. It was also concluded that although the individual efficacy of the algorithms in the processor was not as high as the wavelet based system in Chapter 2, it was possible to increase the efficacy by simple boolean combination of the algorithm results. Power efficiency was achieved by operating the system at aggressively scaled $V_{DD}$ very near the threshold voltage for the transistor in that technology. In Chapter 4, the main issue of leakage was addressed by using low power circuit level methods such as power and clock gating. The algorithm was divided into two stages in order to reduce the power consumption. This was achieved by sharing the computational load between two algorithms of varying efficacy and energy efficiency. It was shown that, significant amount of energy savings are obtained along with an increased efficacy of detection, if leakage controlling techniques are used along with intelligent choices of the algorithms in two stages. These three techniques for seizure detection are promising and could be reference for designing other algorithms and circuits for implantable anti-epilepsy prosthesis. In the final Chapter 5, the algorithm circuit methodology
was successfully extended to a wearable biomedical application. A pattern detection system was developed which detects audio biological symptomatic patterns. Using the properties of DWT and metric based detection, the acoustic patterns related to various symptoms were classified with high degree of accuracy. A modified MFCC based approach was also developed in order to distinguish between signals which have very similar frequency spectrum.

Based on these results and analysis, we conclude in this dissertation that in order to develop systems for critical implantable biomedical application or a wearable health monitoring application, algorithm-circuit co-design is an optimal approach in order to achieve flawless functionality along with pragmatic implementation. The solutions developed in this research will hopefully, provide a basis for development of efficient systems for various biomedical applications.

6.1 Future Direction for Research

The findings in this dissertation show that the stringent design constraints for user-critical applications can be addressed by designing the algorithm and circuit in tandem. In the future, it would be interesting to see if the seizure ‘prediction’ algorithms could be designed by interpreting the information contained in the LFP, similar to brain-machine interfaces. Such systems would forewarn the patients of an impending seizure instead of detecting the onset itself. The human body and the brain in particular, by itself, has evolved into an extremely energy efficient system and has been extensively studied. The various signals generated within the human body contain a plethora of information which is processed by the brain using algorithms which are yet to be understood. However, these functions could be mimicked by using intelligent approximations and smarter processing. By designing the corresponding hardware simultaneously, high energy efficiency is also possible. One of the other primary areas of biomedical applications, where this research can be applied fruitfully is the area of image processing for restoring eyesight for the blind.
Such an implant would need extremely low power implementation for highly complex signal processing. The signal processing may be needed at the sensor side (eye) or the computational side (visual cortex in the brain). The ideas presented in the implementation of health monitoring system, could integrate a vast number of health parameters in the body to make more complicated diagnosis possible in advance. Apart from biomedical application, the co-design approach can be applied to mobile technology also where significant processing is required at minimal hardware cost. In all the possible applications mentioned, it should be noted that there is a certain degree of inherent error resiliency. The true measure of success of this research would be its application to solve intractable health problems and better the quality of human life.
REFERENCES
REFERENCES


VITA
VITA

Himanshu Suhas Markandeya was born in India. He did his schooling in Mumbai in the Atomic Energy School System. After completing his schooling in 1999, he commenced his undergraduate engineering studies at University of Mumbai. He received his Bachelor of Engineering in Electronics Engineering in 2003. Subsequently, he joined Tata Consultancy Services Ltd., Mumbai as Assistant Systems Engineer (trainee). In 2004, he commenced his graduate level studies in Illinois Institute of Technology, Chicago, Illinois. In the fall of 2005, he worked as a co-op student with Diehl Controls North America Ltd., Naperville, Illinois as a Hardware Design Engineer. He received his Master of Science in Electrical Engineering in 2006. Thereafter, until June 2007, he was working as Quality Analyst with Autodesk Inc. in San Rafael, California. In August 2007, he commenced his doctoral graduate studies at Purdue University, West Lafayette, Indiana. He was also working as Component Design Engineer on an internship position at Intel Corporation in the summer of 2013.