LOW POWER ANALOG CMOS CIRCUITS FOR AMPLIFICATION AND ADAPTIVE DETECTION OF NEURAL SPIKES

Shashank Dwivedi

DESIGN OF LOW POWER LFP AMPLIFIER AND ADAPTIVE NEURAL SPIKE DETECTION CIRCUITS

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

SHASHANK DWIVEDI

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

JULY 2019

Certificate

This is to certify that the thesis entitled "**DESIGN OF LOW POWER LFP AMPLIFIER AND ADAPTIVE NEURAL SPIKE DETECTION CIRCUITS**", submitted by **Shashank Dwivedi** (10610211), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under our supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:
Guwahati.

Prof. A. K. Gogoi
Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

In memory of my father

Late Sh. Yadvendra Dwivedi

Acknowledgements

First and foremost, I feel it as a great privilege in expressing my deepest and most sincere gratitude to my supervisor Professor Anup K. Gogoi, for his excellent guidance throughout my study. His kindness and gentlemanliness have been a great inspiration to me. My heartfelt thanks to you Sir for the unlimited support and patience shown to me.

I am forever indebted to my course work instructor and doctoral committee chairman Professor Roy Paily Palathinkal. It is due to him that I found the interest in Microelectronics. His wisdom and expertise has been essential to my work. I am also very thankful to my doctoral committee members Prof. Shaik Rafi Ahamed and Dr. Sonali Chouhan for sparing their precious time to evaluate the progress of my work. I express my heartfelt thanks to Dr. Mahima Arrawatia for providing valuable suggestions on this thesis. I am also grateful to all the members of the research and technical staff of the department without whose help I could not have completed this thesis. My special thanks to Josphine Ma'am, Utpal Sarma Sir and Rakesh Singha for providing components for in-house testing and various resources useful for the research work.

I convey my sincere thanks to the group head Mr. H. S. Jatana from Semi-Conductor Laboratory (SCL), Chandigarh, senior scientists Mr. Uday Khembette, Mr. Asim, and Mr. Deep Sehgal for all the support and providing me with the great learning opportunity. My Ph.D. work definitely would not be complete without their technical expertise that allowed me to comprehend my work even better. My special thanks goes out to my friends Jitendra Kumar and Rajan Singh at the Centre of Nano Technology (CNT) Laboratory for setting up of equipments and successfully carrying out further testing work. I am extremely thankful to my senior colleague Sumit Agarwal for extending his support, both technically and personally, during my stay at IITG. I have to acknowledge his positivity and creativity which has inspired and helped me in several ways. In addition to this, I would also like to thank all my friends I have had during these years, who have been involved in this process. I cannot mention you all by name, but I will always remember your patience and help. Thank you.

My deepest gratitude goes to my father for keeping me motivated throughout my studies till his last breath. I do not find enough words with which I can express thanks to my mother for her undying moral support. Their unlimited sacrifices are the reasons where I am and what I have accomplished so far. Lastly, I want to thank my wife for her patience, love and companionship during my work.

Abstract

The scope of this thesis work is motivated by implementation challenges of designing low-power analog circuits for front-end amplifiers and area-efficient spike detector for neural recording microsystems. The key challenge in the design of neural amplifier is large input DC offsets associated with the electrodes. Neural signals superimposed on these offsets saturates the amplifier's front-end. One of the solution is to improve input linear range in order to accommodate offset voltages of up to few tens of mV. First part of the thesis proposes a sub-theshold OTA topology for the development of a nano-power local field potential (LFP) recording amplifier with a high dynamic range specification. This work also addresses the design criteria of power dissipation, area consumption, and the noise performance.

The second part of the thesis presents a low-power neural recording front-end amplifier with a band programmability feature for separation of extra-cellular neural spike and LFP signal. This allows for the preconditioning of neural signals for delivery to subsequent spike detection system. An improved version of neural recording amplifier configured as LFP amplifier is also presented. The amplifier design focusses on minimizing low cut-off frequency for filtering out large DC offset and low frequency noise. The performance of amplifier is validated by implementing the LFP recording amplifier integrated circuit in a commercially available 1.5 V 180 nm CMOS process. The measured results shows the amplifier to be programmable for different mid-band gains while achieving a sub-10 mHz cut-off frequency for low noise neural data processing.

Third part of this thesis proposes a novel spike detection algorithm capable of estimating the instantaneous neural background noise. The objective is to develop a robust spike detection system which yields a reliable performance under rapid variations of spike amplitude and firing rates. We describe an area efficient, low power subthreshold analog spike detector circuit capable of adaptively discriminating neural spikes from background noise

in real-time. The circuit implementation focusses on achieving an optimal performance in terms of energy-efficiency and compactness. Simulations are conducted which establishes the robustness of the proposed technique in terms of detection performance when compared to state-of-the-art. The algorithm implemented in analog CMOS circuit leads to a low power consumption and occupies an area that fits well into a size of acceptable standards.

The major contributions of this thesis are as follows:

- Linearity improvement in a 1 V transconductor with a nonlinearity error of less than 0.5 % for the differential input signals of up to 1.2 V (peak-to-peak)
- A 0.8 V single-ended LFP amplifier with a THD of 1%@20.2 mV_{pp} having layout area of 0.098 mm^2 while consuming 68.4 nW of power.
- A 1.5 V programmable neural recording amplifier configured for LFP with a cut-off frequency of 3 mHz having layout area of 0.086 mm^2 and power dissipation of less then 10 μW.
- A 0.8 V automatic real-time spike detector with an adaptive threshold estimation with power dissipation of 4.2 μW and a layout area of 0.016 mm^2.

Keywords: Linearity, operational transconductance amplifier, Dynamic range, Action potential, Local field potential, ultra-low power circuit design, neural amplifier, weak inversion, Spike detection, nonlinear energy operator, Extracellular neuronal recordings, Spike train, Threshold selection.

Contents

List of Figures

List of Tables

List of Acronyms

ADC	Analog to Digital Converter
AFE	Analog Front End
AC	Alternating Current
AP	Action Potential
AT	Amplitude Threshold
BMI	Brain Machine Interface
BJT	Bipolar Junction Transistor
BW	Bandwidth
CCIA	Capacitor Coupled Instrumentation Amplifier
CMOS	Complementary Metal-Oxide-Semiconductor
CMFB	Common Mode Feedback
DC	Direct Current
DCCA	Double-ended Capacitively Coupled Amplifier
EEG	Electro Encephalography
ECoG	Electro Corticography
ED	Energy of Derivative
EMG	Electro Myogram
HD_3	3^{rd} Harmonic Distortion
LFP	Local Field Potential
MEMS	Micro Electro Mechanical System
MOSFET	Metal-Oxide-Semiconductor Field-Effect Transistor
MEA	Micro Electrode Array
NEF	Noise Efficiency Factor
NEO	Non-linear Energy Operator

OTA	Operational Transconductor Amplifier
PDF	Probability Density Function
PCB	Printed Circuit Board
PSD	Power Spectral Density
RMS	Root Mean Square
SCL	Semi Conductor Laboratory
SNR	Signal to Noise Ratio
SUA	Single Unit Activity
SCCA	Single-ended Capacitively Coupled Amplifier
SET	Spike Elimination Technique
SRAM	Static Random Access Memory
SFDR	Spurious Free Dynamic Range
THD	total harmonic distortion
VLSI	Very Large Scale Integration

1

Introduction

Contents

1.1 Introduction to Neural Signal Acquisition

In recent decades, a considerable research effort has been put into developing an implantable neural recording front-end system yet several challenges remain due to severe design constraints on power and area efficiency. The simulations on the brain functionality leads to more sophisticated methods for testing new technologies like drugs and neural implants. Over the last couple of years, neural implants have become more significant in understanding human brain and in treating and monitoring various neurological disorders. The implanted system should be able to record the neural activities at a high spatial and temporal resolution for accurate and efficient processing of neural information. Brain-machine interface (BMI) are devices that decode neural activity to provide control signals for external devices, computers or neural prostheses. It enables communication between the human brain and a machine with an ultimate goal of restoring full motor function for people suffering from conditions such as spinal cord injuries or loss of limbs. One of the most important part of brain-machine interface is the development of fully implantable micro-electrode arrays (MEAs) for monitoring extracellular neural activities. Nordhausen [2] used high-density microelectrode array featuring a sensing area of 4×4 mm^2 that can record up to 100 extracellular neural signals simultaneously. The MEAs are ideally implanted underneath the skull, reads out bio-potential signals from the brain and transmits them via a wireless telemetry to a receiver outside the skull. It allows neuro-scientists to monitor and manipulate neural activity and enables the control of neuro-prosthetic devices such as robotic arms.

Acquisition and processing of these bio-potential signals are an important task in biomedical systems. Action potentials (AP), or "spikes", generated electrochemically by individual neurons are the fundamental neuronal measure for neurobiological experiments. These electrophysiological activities are recorded by placing an electrode in proximity ($\sim$10 μm away) of an active neuron. Such voltages are generated in the extra-cellular space in the current field outside the cell. Extra-cellular action potentials look very much similar to the action potentials that are recorded intracellularly, but the signals are much smaller (typically about 100 μV). For recording single-unit neural activity, the tip of a single electrode must be small enough to be able to resolve the size of a neuron (≤ 50 μm) [3]. Further, for capturing an action potential with a high signal-to-noise ratio, it should be able to attenuate weak signals emanating from nearby neurons especially in a densely populated neural environment (neural somas of 5$-$15 μm diameter) [4]. A sharp-tipped micro-electrode of 50 μm has been reported [2] that is capable of isolating these single units increasing signal-to-noise ratio of the recorded neural responses. The micro-electrode recording surface of 100 μm diameter may cover an approximate spherical distance of 50−350 μm [5]. BMIs utilize these electrodes to record extra-cellular activities in order to control prosthetic devices with many degrees of freedom [6]. Temporal occurrences of spikes are known to represent important information which allow BMIs to decode neural recordings into control commands [7]. Neural recordings contain both the spikes and the background noise. In chronic neural recordings, signal-to-noise ratio may vary drastically [8]. A technique usually employed is a spike detection method to distinguish spikes from background noise. Spike detectors also reduce the data rates which reduces bandwidth requirement for ADCs and bio-telemetry applications. Spikes are very low in amplitude and contain large spectrum of neural information therefore these signals need to be preconditioned before any further processing.

Adding on-chip CMOS circuitry to the passive micro-probe improves overall signal quality and permits single-unit recording at a higher spatial resolution [9, 10]. Therefore, it is desired to realize neural front-end interface for signal acquisition. Fully-integrated signal acquisition systems typically include a low noise amplifier for amplification of weak neural signal, a band-limiting filter to filter either the spike or LFP bands for signal conditioning, a analog sample-and-hold and an ADC or spike detector for signal processing on the same substrate as recording electrode. These designs have relied heavily on analog techniques to implement front-end interface which mainly includes the neural amplifier and bandpass filter [11,12]. The requirements on the design of such an interface are extremely stringent, especially with respect to noise, integration area, and power consumption.

1.2 Analog Front-end Interface (AFE)

To realize an implantable neural data acquisition system with low-power wireless system and high-speed data transfer, a feasible solution is to integrate analog front-end circuitry with micro electrodes which can perform telemetry on bio-potential signals to a remote computer. Analog front-end circuits for measuring bio-potential signal are implemented in CMOS technology using integrated circuit (IC) since it offers low current consumption and dense integration. However, the nature of bio-potential signal and the electrode-tissue interface poses several circuit design challenges on the front-end micro-electronics. Weak neural signal measured by each electrode must be boosted first before passing signal to data compression circuitry. Evidently, the signal processing block must be preceded by a neural recording amplifier in the front-end.

The amplified signal are transferred for further signal processing where they are converted to digital domain in the signal processing block. The temporal feature of neural signal such as time instants of spike occurrence encode important neural information [7,13] which is critical in interpreting motor commands for BMI applications. Since the spike occurrence is sparse with firing rates typically between 10−100 Hz [11], data rate can be reduced by transmitting only the AP waveforms and their respective time stamps. This compression of data is essential for integrated neural recording devices since it decreases the data rate that must be handled by the subsequent low-power digital circuitry (data converters) and permits increase in number of channels. Data compression is performed with the help of spike detectors. A high performing spike detection system should be able to distinguish spikes from the background noise. Correct detection of spikes largely depends up on the quality of extra-cellular recording which ultimately depends on amplitudes of spikes relative to the background noise.

Figure 1.1 shows a typical architecture of a neural recording microsystem for a single channel. It consists of three major blocks: (1) a low noise amplifier, (2) a band-pass filter, and (3) a signal processing unit. Low noise amplifiers (LNA) are directly interfacing with the electrodes. The neural signal in LNA is sensed differentially with respect to a reference electrode. The amplified signal is band-pass filtered which defines upper and lower cut-off frequencies. The signal processing unit typically consists of spike detector for achieving some level of data compression and subsequent reduction in bandwidth requirement. Later on, the amplified and digitized data is relayed via a wireless telemetry unit.

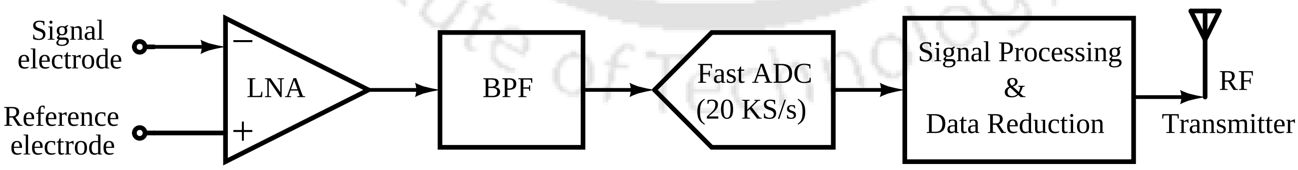


Figure 1.1: A generic block representation of a single-channel neural recording microsystem.

We report a novel data reduction technique and discuss hardware efficient architecture for implementing spike detection in real-time. This research work focusses on the design aspects of neural recording front-end integrated circuit capable of acquiring neural activity over large number of channels. We also describe low-power analog circuits that can be applied to many brain-machine interfaces.

1.2.1 Design Challenges of Data Management

A rapid development of highly integrated multi-channel recordings enables monitoring from hundreds to thousands of neurons simultaneously. The raw data rates with these many numbers of recording units are prohibitive in high volume. In addition, these systems have severe size and power constraints which are limited in the amount and computation complexity of information that can be processed [14]. A data reduction technique should be applied before transmitting neural recording over a wireless channel [11,15]. Since neuronal spikes contain the relevant information therefore only spike events needs to be communicated instead of entire raw data. With the spike firing rate of 50 times per second and taking the spike duration to be 2 ms, a data rate reduction of only 10 times can be achieved. Data can be further reduced by indication of only the presence or absence of neuronal firing activity. Data transfer rates and power consumption can be reduced with different data management approaches in neural recording microsystems.

In these schemes, on-chip circuitry performs the identification and the extraction of the incoming neural waveforms, so that the system only focuses on active portions of recorded signal. Detection of the presence of the spikes is demonstrated in [16]. A simple scheme which consists of defining an amplitude threshold for locating the spikes in real time have been demonstrated [17,18]. Another data compression approach employing wavelet transform is suggested in [19]. Other data reduction methods consists of the software based spike sorting algorithms which indicate mere presence of spikes [20]. In a very recent work on data compression, spike detection through amplitude thresholding is implemented in analog domain. The energy and area-efficient successive approximation ADC is implemented and converts only the detected spikes, decreasing the power dissipation and the amount of neural data [21]. Another publication on spike sorting performs conditional spike detection, alignment, adaptive feature extraction, and online clustering with sorting threshold self-tuning capability [22]. Various data reduction schemes can be broadly classified into three types of approaches depicted in Figure 1.2. A pre-amplifier (LNA) stage is essentially used in all cases for amplifying μV level signal from the electrodes.

(i) Approach A: sample and quantize the raw data and only transmit a clip of the waveform around the spike (few approaches send features needed for spike sorting);

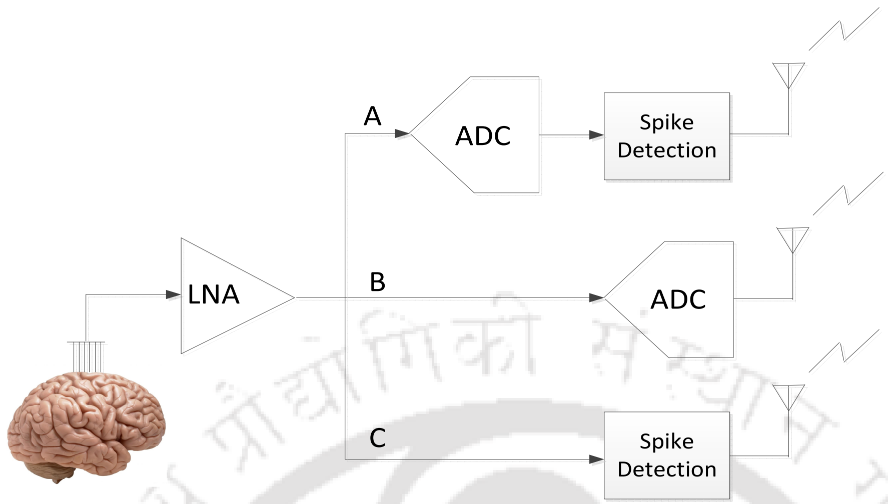


Figure 1.2: Block diagram for three degrees of data reduction.

(ii) Approach B: sample and quantize the raw data using conventional techniques;

(iii) Approach C: only transmit the spike times.

The reduction techniques of approach A offers a moderate data reduction and is performed in the digital domain alongside approach B. The processing blocks in approach A can also be interchanged by parsing of data around spike first and then digitize the information. In approach B the wideband amplified neural signal directly passes through ADC and is a most straight forward technique for digitization. These neural recording systems typically use sampling rates of 30 kS/s with resolution of 15 bits producing data rates of over 45 Mb/s from a 100-electrode array [23]. Transmitting data at such high rates over a wireless transcutaneous link requires more bandwidth and power which is not feasible in a small-sized, implanted system. Approach C of spike detection offers highest data reduction and is mostly performed in the analog domain, which we adopt in our present work. Analog spike detection is considered as more power-efficient since the ADC would only need to run when there are spikes, whereas in digital detection the ADC must constantly be running, since detection occurs only after sampling.

When a neural information is processed in implantable systems, the primary objective is to establish the real-time continuous streaming data across brain-machine interface while preserving variable wave shape features so that the extracted bio-potential can be further utilized in subsequent processing steps. For example, the time of occurrence, the maximum amplitude value, and the minimum amplitude value are some of the essential waveform features that are important information in order to associate a biopotential signal with a specific neuron. For this reason, amplitude thresholding is specially attractive for performing real-time detection due to its relative simplicity and implementation cost. Therefore, in approach C, spike detection can be done by simple thresholding and, due to the limited duration of the action potentials, transmitting data recorded for a few milliseconds combined with a time-stamp is sufficient. This form of spike detection scheme can achieve ultimate data compression for neural recordings, where only spike times or binned spike counts are transmitted [24].

It is customary to perform data management after the signal is passed through a neural amplifier so as to amplify the signal for minimum detectable amplitude and to make detection less sensitive to the background noise. Since the detection performance depend up on pre-processing of neural data hence, the design considerations of a suitable front-end neural amplifier is crucial to a high-performing neural data management system.

1.2.2 Design Challenges of Neural amplifier

The neural recording amplifiers are an indispensable part of the integrated CMOS circuit of micro-electrode arrays and performs amplification of bio-potential signal through electrode-tissue interfaces in the front-end of a neural recording system. Bio-potential signal recordings from extracellular microelectrode arrays consist of two components superimposed on each other: action potentials from single and multi unit activity (spikes or AP), and slow varying field potentials (LFP).

Noise-Power-Area Trade-off: Extracellular AP and LFP measurements deal with weak signals although the latter is highly invasive compared to the former. Because the neural signal to be recorded has a very low amplitude ($\sim 10\ \mu$V), it must be amplified before processing any further. Hence it is imperative that the amplifier has to operate with ultra-low power while occupying a minimal implantable area in a neural recording system. Simultaneously, because of very small amplitude of neural spikes, the amplifier should also have a low input-referred noise (below 5 V_{rms}) with large DC gain (greater than 30 dB). Lower power consumption cannot be attained without sacrificing the noise or gain of the amplifier and the area efficiency. To access the trade-off between the power consumption and the input-referred noise level in the design of a circuit, a noise efficiency factor (NEF) is defined in [25] and is widely used to compare neural-amplifier designs:

$$NEF = v_{ni,rms}\sqrt{\frac{2I_{tot}}{\pi.U_T.4kT.BW}} \tag{1.1}$$

where $v_{ni,rms}$, is the total input-referred noise of the amplifier, I_{tot} is the total supply current, U_T

is the thermal voltage KT/q, and BW is the -3 dB bandwidth of the amplifier respectively. NEF is a function of noise, bandwidth, and supply current. With a trade-off between noise and power the best achievable NEF = 1. Since for attaining low noise, the amplifier requires more current supply, the device dimensions needs to be increased. Moreover, capacitors occupy a dominant area of the layout. It can be shown that the total input referred noise of OTA consisting of two differential pair input transistors is inversely proportional to the load capacitance C_L. The input-referred noise per unit bandwidth of the OTA is $\overline{v_{ni}^2} \approx 2 \times 2kT/(\kappa g_m) = 4kT/(\kappa g_m)$ where g_m is the transconductance of a single differential-pair transistor. For minimum input-referred noise, the transistors should run in subthreshold, such that we have $g_m = \kappa I_D/U_T$. The parameter κ, is the subthreshold gate coupling coefficient of a MOSFET which has a typical value of 0.6−0.7. Assuming a first order roll-off of the frequency response, the input-referred noise of the ideal OTA is expressed as

$$\overline{v_{ni}} \approx \sqrt{\frac{4kT.U_T}{\kappa^2 I_D}.BW} \tag{1.2}$$

Assuming upper cut-off frequency f_U to be very large than lower cut-off, f_L, $BW = f_U - f_L \approx \pi/2.f_U = g_m/4C_L A_{m1}$, where A_{m1} is the mid-band gain of OTA (refer equation (4.2)). The input referred noise integrated over bandwidth BW is expressed as

$$V_{ni,rms} \approx \sqrt{\frac{4kT}{3C_L A_{m1}}} \tag{1.3}$$

Therefore, the challenge in the design of implantable neural recording amplifier also lies in the tradeoff between noise and layout area. Equation (1.1) and (1.3) suggests that the noise can not be reduced without a lower limit since it has a trade-off with power dissipation as well as silicon area.

Electrode DC Offset: Another design challenge of neural recording amplifier arise from the variation in the input DC offset voltage due to the electro-chemical interaction at the interface of electrodes and living tissues. These offset voltages are superimposed on the actual voltage signal to be measured. It can range up to several mV which can saturate the output of neural amplifier. Since the time constants associated with neural front-end are large therefore for a large input fluctuation, the amplifier may stop amplifying and it takes a long time to return to its normal bias voltage. The front end of neural recording systems are not well equipped to handle large DC offset or the "baseline drift". A capacitor in series with the amplifier input terminal can overcome the problem by absorbing any dc fluctuation at the input and increasing the dc gain while also decreasing the input referred noise. For these many reasons, the conventional amplifiers employ a large capacitor or the bank of capacitors placed at the input of recording channel. However, these integrated capacitors cannot have very large values as it would increase chip area which decreases number of recording channels. It may also decrease input impedance degrading the common-mode rejection ratio. Further the capacitors cannot be scaled with technology. Therefore, a circuit solution is highly needed that is able to grant the necessary signal swing for the application and cater to the restrictions of noise and size specifications.

It is required for such amplifier design to be robust to large signal fluctuations with a high linear range and simultaneously the ability to filter DC offset with lower input referred noise. To counter this situation, we can either increase the DC gain or increase the dynamic range of amplifier. With a constant demand to reduce device size and weight without effecting recording quality, the challenge to meet these design specifications in such amplifiers continues to grow. In order to meet these challenging specifications, a topological configuration has been proposed which finds suitability for the present application. In this thesis work, the signal processing approach presented is predominantly analog in nature. The amplifier and processing blocks are realized with analog circuits. Analog circuits are desirable for the neural front-end design as it provides more flexibility of choosing design variables and allows the input-output behavior of CMOS transistors to be better utilized for an optimum performance of power and area computations.

1.3 Motivation and Problem Statements

The scope of this research is motivated by the implementation challenges of neural recording system which includes the design of a spike detector circuit and a suitable neural pre-amplifier. We identified the following shortcomings/problems in the existing amplitude threshold based spike detection systems.

(i) The presence of spike in the neural data stream biases the estimate of the background noise even when the firing rate is moderately high ($\sim$ 50 Hz). The firing rate severely effects the overall accuracy of those measurements and subsequently the spike detection performance.

(ii) The background noise estimation are sensitive to the morphological features of spike and their temporal occurrences as a result the performance of these detectors shows partial or total dependence up on the voltage magnitude, inter-spike interval and variability in wave shape of incoming spike.

(iii) The adaptive threshold of the conventional NEO-based algorithms [1, 26, 27] is derived from a low-pass filter. These algorithms achieve best performance only within a specific range of filter cut-off frequencies. The detection circuit performance is also sensitive to process and temperature variations.

Further, for proper functioning of spike detectors, low-power neural pre-amplifiers are very crucial, since one such amplifier is required before a spike detector circuit for each electrode. Hence the amplifier design should adhere to the following implementation challenges:

(a) Since the detectable spikes have small amplitude ($\sim$ 1-10 μV) therefore it is essential to amplify them up to an adequate level ($>$ 1 mV) over the frequency band of interest (from sub-1 Hz to 10 KHz) while keeping low distortion and a low input referred noise.

(b) The area of on-chip neural recording system should be kept to minimum since it is limited by the pitch of MEAs whose width generally varies between $200-400$ μm [28].

(c) The overall power dissipation of a multi-electrode neural recording implant must be restricted (80 mW/cm^2) to prevent excessive heating of brain tissues [29].

(d) The neural recording amplifier should have adequate dynamic range to convey neural spikes / LFPs with offset voltages as large as 10 mV in amplitude.

(e) Random DC offset and flicker noise have to be reduced to prevent saturation of the bioamplifier.

1.4 Thesis Contribution

This thesis aims at the design of low voltage, area efficient analog spike detection circuit and a low voltage, low noise neural amplifying device for data acquisition for a fully implantable neural front-end. The proposed algorithm improve the detection performance by providing more accurate background noise estimation. As in the existing amplitude threshold based spike detection algorithms [1,17,18,26,27], we assume in this thesis that: (1) there are no overlapping spikes (2) High frequency artifacts have been eliminated (3) Baseline interference have been filtered out. The major contributions of the thesis are presented in the following five chapters.

1.4.1 Chapter - 2

In this chapter we propose a 1 V transconductor in strong inversion region with a nonlinearity error of less than 0.5 % for the differential input signals of up to 1.2 V (peak-to-peak). we talk about improvement of input linear range to cope with offset voltages of up to 10 mV for biomedical instrumentation. We propose an operation transconductance amplifier (OTA) suitable for neural amplifier and discuss design principles which caters to our requirement. The performance of OTA is also validated by realizing a neural amplifier with the proposed OTA.

1.4.2 Chapter - 3

In Chapter 3, we propose an ultra low power, low noise amplifier for recording local field potential signal. The OTA that is introduced in chapter 2 is used as a core structure to develop low noise amplifier for a bandwidth of 2 to 200 Hz. The chapter presents the design technique which optimizes gain with respect to size and power consumption. It achieves a gain of 31.7 dB of gain while consuming 68.35 nW of power with a dynamic range of 68 dB. It also presents a very low noise resulting in noise efficiency factor near its theoretic limit and well above the background noise from electrode site (15 -20 μV_{rms}).

1.4.3 Chapter - 4

The focus of chapter 4 is on the interface between the neural tissue and the amplification of neural signals prior to detection. Hence, we incorporate band programmability feature to the existing amplifier so as to record either low frequency LFP (1 Hz $-$ 300 Hz) or extra-cellular high frequency spikes (300 Hz $-$ 6 KHz) without sacrificing noise or gain requirement. A $g_m - C$ band pass filter stage is used to tune the corner frequencies to pass only the signal band of interest and to make the noise independent of the supply current. The gain is then boosted with the help of a high gain amplifier stage to obtain a gain of more than 70 dB. Both the stages contribute minimal noise so they can be neglected in overall noise calculations.

The later part of this chapter describes a modified scheme of the first-stage amplifier design used for LFP recording so that the low-frequency noise can be suppressed. The main design target therefore lies in minimization of the achieved high-pass corner frequency so that the cut-off frequency can be brought down into sub-Hertz range. For validating the design concept, the post layout simulations are presented to verify the feasibility of proposed design. A proof of concept is given by fabricating two prototype amplifier microchips in 0.18 μm CMOS process. The second prototype configured for LFP has a low cut-off frequency of 3 mHz having layout area of 0.086 mm^2 and power dissipation of less then 10 μW. The amplifier is suitable for applications ranging from neuro-prosthetics to seizure monitors.

1.4.4 Chapter - 5 and 6

A neural spike detection technique is discussed in chapter 5. A spike detector ideally converts spike events into digital domain and allows for the reduction in data rate by transmitting only the relevant information. The objective is to develop a robust spike detection system which yields a reliable performance under rapid variations of spike amplitude and firing rates and can be area efficient. An adaptive algorithm is proposed which is capable of detecting spikes in real-time with the high firing rate of up to 100 Hz.

Finally, in Chapter 6, the performance is evaluated using realistic simulations of extracellular recordings which establishes the robustness of the proposed technique compared to state-of-the-art. The algorithm is implemented in analog CMOS circuit. Analog design techniques are discussed which leads to a 0.8 V automatic real-time spike detector with an adaptive threshold estimation with power dissipation of 4.2 μW and a layout area of 0.016 mm^2.

1.5 Organization of the Thesis

This thesis is organized as follows.

Chapter 1 states the problems studied in this thesis and the motivation behind these problems. This chapter highlights the data management schemes for data reduction and design challenges of existing neural recording amplifier. This chapter also summarizes the thesis contributions and provides a brief outline of the thesis organization.

Chapter 2 proposes the design of an operational transconductance amplifier (OTA) in weak inversion region for improvement of linearity and investigates the implementation of a typical single-ended neural amplifier with respect to the size and dynamic range features.

Chapter 3 describes an improved design of an ultra low power, low noise amplifier for recording of local field potentials (LFP) and provides a better noise-power trade-off capability.

Chapter 4 introduces band programmability into a low power analog front-end interface with an additional feature to record neural spikes and LFP signal separately. This chapter also describes two versions of LFP amplifier with a very low cut-off frequencies and presents proof of concept by fabricating microchips using 180 nm CMOS process.

Chapter 5 and 6 describes a real-time neural spike detection scheme by proposing an area-efficient algorithm and its low-power analog design implementation for better performance.

Chapter 7 concludes this thesis with a summary of the work done and includes some suggestions that may be investigated in future research.

2

A Linear CMOS Transconductor For Neural Recording Amplifier

Contents

2.1 OTA Design Techniques

Operational transconductance amplifier (OTA) is an important building block in integrated continuous-time filters. As the device sizes scale down, conventional saturation based OTAs face design challenges in terms of linearity and output impedance. OTA design techniques such as non-linearity cancelation [30] and adaptive biasing techniques [31] are commonly used to improve the linearity performance. However, it suffers from significant second-order harmonics partly from the saturation operation of MOSFETs and partly because of device mismatch due to process variation. For circuit to operate in weak inversion, bump linearization technique [32] can be used to extend the linear range. This attenuation technique achieves linear range extension by attenuating the input signals (i.e., via bulk driven transistors or two transistors in series). However, this attenuation technique increases the input referred noise, which disqualifies this technique from being applied to biomedical circuits such as electrocardiography, cochlear implants or neural signal recording applications where the noise performance is of prime importance. We presented an adaptive biasing scheme for linearization of CMOS transconductor in [33]. Such linearization techniques are more suitable for transistors working in strong inversion. Furthermore, the class of input signals which can be processed is limited since the requirement of fully balanced signals is needed for the squaring circuit to function properly hence this technique may not be useful for application of OTA requiring single-ended operation.

The work reported here focusses on low voltage, low noise OTA design which can be used as a basic amplifier for invasive extracellular neural recording microdevices designed to record neural activities within the cerebral cortex. Figure 2.1 illustrates the block diagram of an implantable neural recording microsystem comprising of several recording probes attached to a signal preconditioning module, a neural processing unit which is a part of analog preprocessing block, an A/D converter, and a wireless interfacing module. In this work we propose a very low voltage symmetrical OTA implemented with two differential pairs in parallel asymmetric multi-tanh configuration [34]. The proposed structure utilizes a DC shifting technique which exhibits linearity for a wide common-mode input range. Later on the OTA has been utilized to implement a neural recording preamplifier used in medical diagnostic method like electro-encephalography (EEG) and magnetic resonance imaging (MRI). The detailed noise analysis is also presented for the OTA followed by simulation performance and comparison of our neural preamplifier with other similar biopotential amplifier designs.

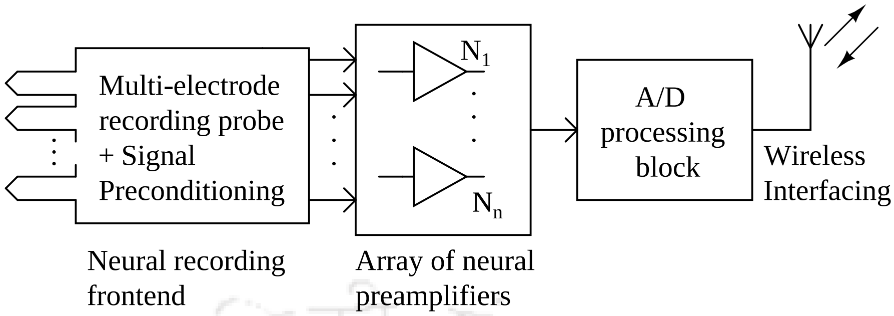


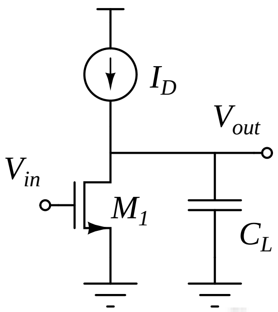
Figure 2.1: Analog front-end of implantable neural recording site.

2.2 Low power Circuit Design

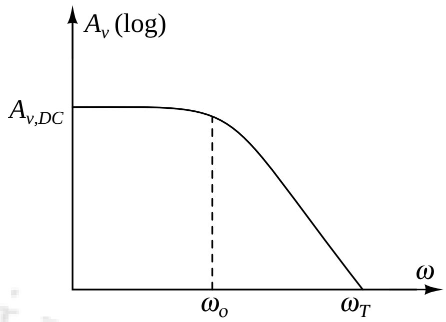
The device sizing rules result from designer specifications for a specified function and its performance validation e.g. regarding mismatch or channel length modulation. These specifications refer to transistor geometry parameters (width, length, area) and electrical transistor quantities (e.g. transistor drain-source voltage). Traditional analog design methodologies typically require iteration for example Square Law technique employed for conventional OTA design have a limited scope since as we move towards submicron devices the design equations become more inaccurate and cumbersome because of difficulty to achieve an optimum performance level in terms of power consumption. A more sensible approach to interpreting the MOS transistor can be based on the level of inversion (or inversion coefficient, IC). The inversion coefficient can he expressed in simplified form as:

$$IC = \frac{I_D}{I_S\,(W/L)} \tag{2.1}$$

where I_S is a process dependent current given by the expression given in (2.4). The best method of interpreting the inversion coefficient is via the efficiency of the transconductance, g_m/I_D. By evaluating g_m/I_D against I_D a lot of insight can be gained into MOS transistor where MOS transistor behavior are contained in two asymptotes - the weak inversion asymptote, where g_m/I_D approaches the thermal voltage limit of transconductance efficiency, and the strong inversion square law asymptotic limit, where the slope in this plot is $-1/2$. In the region where these two asymptotic limits cross, the interpolation region of moderate inversion exists between these two limits. g_m/I_D based design is more reliable technique since it employs design charts (or spreadsheets) measuring device performance for all operating regions to accurately size transistors. It can thus very well link the design variables (g_m, f_T, I_D, etc.) to specification like gain, bandwidth, and power consumption.



(a) Intrinsic gain-stage



(b) The frequency response

Figure 2.2: Intrinsic Gain Stage with M_1 saturated

A single-MOS circuit, shown in Figure 2.2(a), consists of a saturated common source MOS loaded by a capacitor, C_L. The objective is to find gate width, W and drain current, I_D so as to achieve a prescribed gain-bandwidth product ω_T. Let's consider a small signal equivalent where a constant current source $g_m V_{in}$ is in parallel with the output conductance g_d and the load capacitor C_L. The g_m represents the transconductance of M_1. At high frequencies, C_L acts like a short circuit path and all the current delivered by the current source flows through the capacitor. The output is given by $v_{out} = (-1/j\omega C_L) \times g_m v_{in}$. Hence AC gain can be given as

$$A_{v,AC} = -g_m/j\omega C_L \tag{2.2}$$

At low frequencies, the capacitor C_L is practically open circuited so that the current flows through the output conductance g_d . As a result, $v_{out} = -g_m v_{in}/g_d$. Hence DC gain is expressed as $A_{v,DC} = -g_m/g_d$. Figure 2.2(b) shows the frequency response of the intrinsic gain stage. The cut-off angular frequency is, $\omega_o = g_d/C_L$ while the transition angular frequency is, $\omega_T = g_m/C_L$. In order to achieve a desired transition frequency f_T one has to fix the drain current and W/L ratio. g_m is already fixed by the expression $g_m = \omega_T C_L$. To connect the drain current I_D and the W/L ratio to the transconductance g_m one has to use square law method in strong inversion where $\beta = \mu Cox\,(W/L)$ and g_m is expressed as $g_m = \sqrt{\frac{2\beta I_D}{n}}$. Thus W/L can be expressed as

$$\frac{W}{L} = \frac{n g_m^2}{2\mu Cox}.\frac{1}{I_D} \tag{2.3}$$

The above equation suggests that gain-bandwidth product can be held fixed by proper choice of (W/L) & I_D product. In weak inversion, the drain current can be represented by means of an exponential expression [35]:

$$I_D = I_S \exp\left(\frac{V_{GS}}{nU_T}\right) \tag{2.4}$$

where, $I_S = 2n\mu_0 Cox {U_T}^2$. The transconductance is given by

$$g_m = \frac{I_D}{nU_T} \tag{2.5}$$

where U_T stands for kT/q and k for the Boltzmann constant. To attain the desired ω_T, the minimum drain current must be equal to $I_D = nU_T g_m$. This equation shows that the drain current in weak inversion alone fixes the gain-bandwidth product while the aspect ratio has no influence at all. It is noticeable that equation (2.2) is not valid for lowest drain currents while equation (2.4) is only valid for low drain currents.

2.2.1 g_m/I_D Design Methodology

The foregoing discussion suggests that the g_m/I_D ratio does not depend on the gate width, W [36] either in strong or in weak inversion region. Therefore, the drain currents achieving a prescribed gain-bandwidth product can be derived from the expression given in (2.6). The numerator consists of the transconductance given by $g_m = 2\pi f_T C_L$ while the denominator has the transistor's g_m/I_D ratio. This ratio is derived from a device having similar electrical characteristics and whose gate width W^* and gate length L^* are known:

$$I_D = \frac{g_m}{(g_m/I_D)^*} \tag{2.6}$$

Once the the drain current is known, width can be calculated from the proportionality:

$$W = (W)^* \frac{I_D}{(I_D)^*} \tag{2.7}$$

Equations (2.6) and (2.7) together form a set of parametric equations that determines I_D and W which helps in achieving the gain-bandwidth product which is fixed by g_m. In few circuit design methods, gain-bandwidth product is fixed by I_D. In this case g_m and W can be evaluated from the above two equations. A general g_m/I_D design rule starts by fixing the specifications to optimize for example: gain and transition frequency in order to determine the unknowns that are MOS devices sizes W and L. The universal g_m/I_D as a function of $V_{EFF} = V_{GS} - V_T$ characteristic of the CMOS technology under consideration (0.18 μm) is considered here [37]. Having the following data assumed to be provided by the designer: gain-bandwidth product, f_T, slew rate, overall gain A_v the design

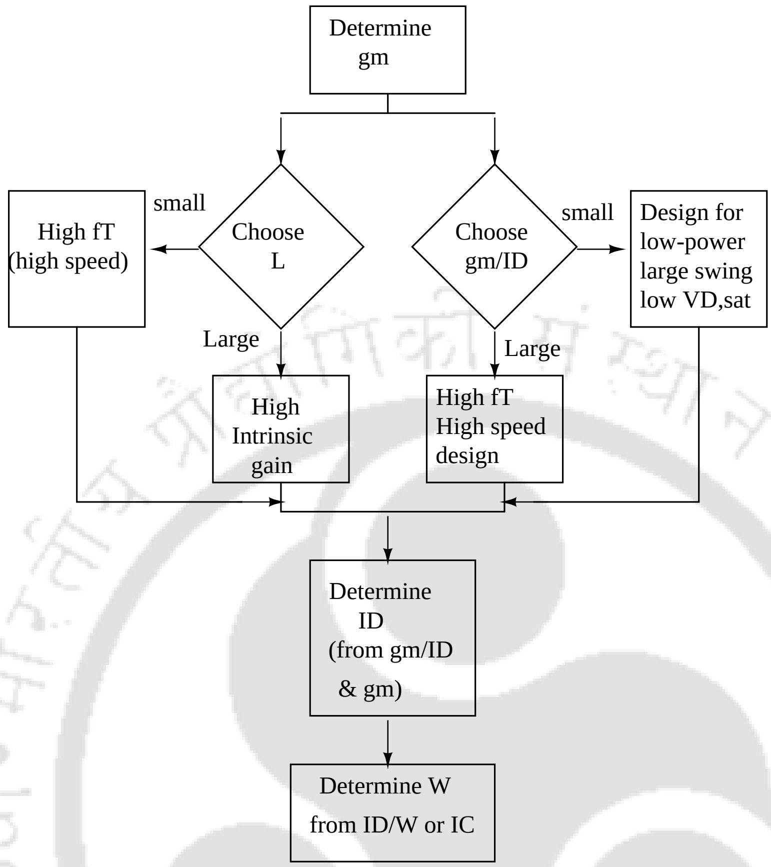


Figure 2.3: A generic design flow.

methodology is as follows:

(i) The drain current of each transistor of the OTA is determined from the specified total supply current I_S along with the current mirror multiplier ratio m. Choice of m depends upon the stability of the OTA since it also directly multiplies f_T and slew-rate as well.

(ii) Choosing the values for g_m/I_D, V_{EFF} is determined for each transistor from the experimental g_m/I_D versus V_{EFF} curves.

(iii) Then, with the V_{EFF} value found in point 1, the W/L of each transistor is found. The intended values of g_m/I_D are chosen accordingly to their effect on the OTA performance.

The maximum value of g_m/I_D we may choose is limited on one hand by the weak inversion maximum value of the technology (about 25 V^{-1} in bulk MOS transistors) and on the other hand by the stability requirements because as we increase g_m/I_D, with fixed current, the transistor sizes and parasitic capacitances are increased and the phase margin is reduced. A generic design flow is illustrated in Figure 2.3 in the form of a flow chart.

2.2.2 Parameter Extraction Methodology

The accuracy of device characteristics and prediction of the performance of a circuit depends not only on the device model but also on the parameter values being used [38], [39], [40]. Hence, the procedures employed to extract the device model parameters are of prime importance. The important part in (2.6) is to setup the parameter, g_m/I_D which enables to sweep the transistor through all modes of operation. It can be derived experimentally from (2.8) by storing the drain currents measured from real transistors with W^* and L^* known *a priori* in look-up tables.

$$\left(\frac{g_m}{I_D}\right)^* = \frac{1}{{I_D}^*}{g_m}^* = \frac{1}{{I_D}^*}.\frac{d{I_D}^*}{dV_G} = \frac{d}{dV_G}\log\left({I_D}^*\right) \tag{2.8}$$

Other method is to derive drain currents from advanced MOS models such as BSIM [1] which allow reconstructing drain currents very close to real values. In the present work, BSIM3v3 device parameter sets were extracted from a series of device $I-V$ characteristics for 0.18 μm process technology. The parameters are determined sequentially in the region of operation where they have there most significant influence on the MOSFET $I-V$ characteristics. In the first step to determine the level of inversion we note down the process parameters like geometry range, voltage range, and temperature specifications from model files of UMC 0.18 μm CMOS process.

– Geometry range: $\begin{cases} 0.18\mu m \leq L_{DES} \leq 50\mu m \\ 0.24\mu m \leq W_{DES} \leq 100\mu m \end{cases}$

– Voltage range: $\begin{cases} 0V \leq V_{GS,DS} \leq 1.8V(\pm 10\%) \\ -1.8V \leq V_{BS} \leq 0V \end{cases}$

– Temperature range: $-55°C \sim +125°C$

The MOST semi-empirical model description used in this work comprises a look-up table of the following data:

(i) g_m/I_D as function of the effective voltage V_{EFF}. The dependency of g_m/I_D with W, V_{DS} is slight and in a first approximation it can be neglected if narrow devices are not used.

[1]BSIM (Berkeley Short-channel IGFET Model) [41] refers to a family of MOSFET transistor models widely used in integrated circuit design available in public domain

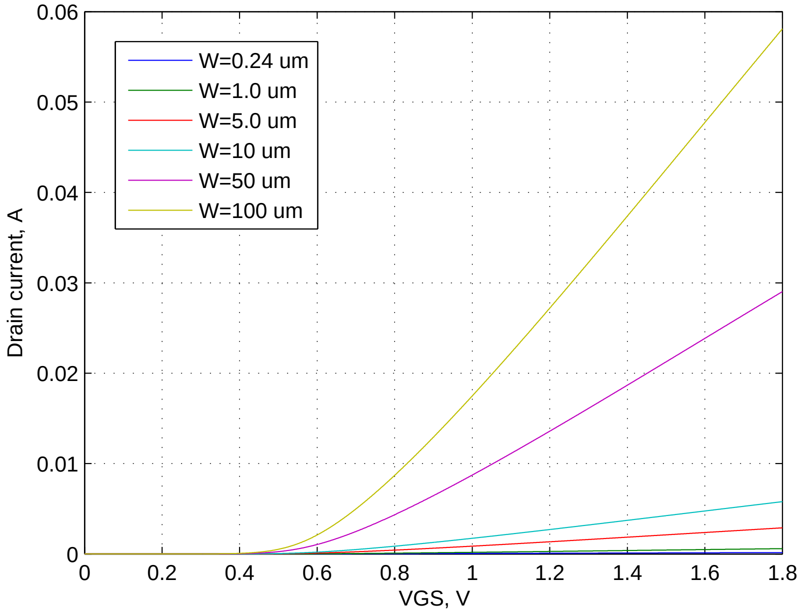


Figure 2.4: Simulated $I_D - V_{GS}$ plot for different gate widths.

(ii) Early voltage, V_A and $V_{DS,sat}$, intrinsic voltage gain A_v, and transconductance g_m as function of V_{EFF} . The variation with W is very slight and it is not considered here.

(iii) Normalized capacitances C_{ij}, with $ij = gs, gd, gb$ versus V_{EFF} and the variation of transition frequency, f_T due to intrinsic capacitances. The spread with W and V_{DS} is reasonably small and it is not included in a first approximation.

(iv) Thermal noise parameters as function of V_{EFF}. The variation of noise parameters with W can be neglected in the first approximation.

(v) Flicker noise parameters versus V_{EFF}, at the RF working frequency. The dependency of K_F with W and V_{DS} is very low and hence not considered here.

The parameter extraction starts by varying input voltage, V_{EFF} and the transistor width W such that the channel current I_D remains constant and $V_{EFF} - W$ parameters will be extracted from graphical representation of Figure 2.4. By holding a drain current to a fixed value allows a designer to choose a range the inversion coefficient $IC = I_D/[I_S\,(W/L)]$ (or V_{EFF} which ever is convenient) in which the OTA performance can be observed in different operating regions and suiting to his design needs.

2.2.3 Design Leverages in Moderate Inversion

One major limitation of the traditional symmetrical OTA operating in weak or moderate inversion is its narrow differential linear input range. A low voltage, low power operation of MOSFET is achievable by exploiting the subthreshold regime or weak inversion region. Nevertheless, designing CMOS circuits with devices in weak inversion have to compromise on the device shape factor resulting in large gate capacitance, low bandwidth, and high DC leakage [42], [39]. Moderate inversion offers a compromise of high (g_m/I_D) ratio, lower $V_{DS,sat}$ voltage and moderately high bandwidth necessary for power efficient, low voltage design. A relatively simple MOSFET model valid in all regions of operation: weak, moderate, and strong inversion is approximated from the EKV MOS model [43] by

$$I_D = I_S S \left[\ln \left(1 + \exp \left\{ \frac{\kappa \left(V_G - V_T \right) - V_S}{2U_T} \right\} \right) \right]^2 \tag{2.9}$$

where $U_T = kT/q$ is the thermal voltage, $S \equiv \frac{W}{L}$ the width to length ratio, κ $(= 1/n)$ is the gate coupling coefficient and represents the coupling of the gate to the surface potential. n is the constant parameter depending on the technology (typically 1.5 near weak inversion), while I_S is the subthreshold current-scaling parameter strongly dependent on the temperature given by $I_S = 2n\mu C_{ox} U_T^2$. C_{ox} is the gate-oxide capacitance per unit area, μ is channel carrier mobility. The MOS transistor biased at weak side of moderate inversion must comply with the following requirement: -72 mV $< V_{EFF} (= V_{GS} - V_T) < 40$ mV, $V_{DS} > 3U_T$ for weak saturation while I_D can be evaluated from the condition $I_D < 2nK_n{}' \frac{W}{L} U_T{}^2$, where $K_n{}' = \mu_n C_{ox}$ is a transconductance parameter of an n-channel MOS transistor.

Although operation below the strong inversion region leads to a lower bandwidth, there are several advantages when frequency is not a major consideration. If the differential input stage uses wide channel devices, it is easy to bias these devices into the moderate or weak inversion region and achieve several additional advantages.

(i) The higher voltage gain results from operation below the strong inversion region [44].

(ii) The low device power dissipation results from the low value of quiescent drain current. This feature has been used in several early low-power designs [45, 46].

(iii) The harmonic distortion decreases. It is appropriate to minimize distortion introduced by input stages since succeeding stages will amplify any first stage distortion.

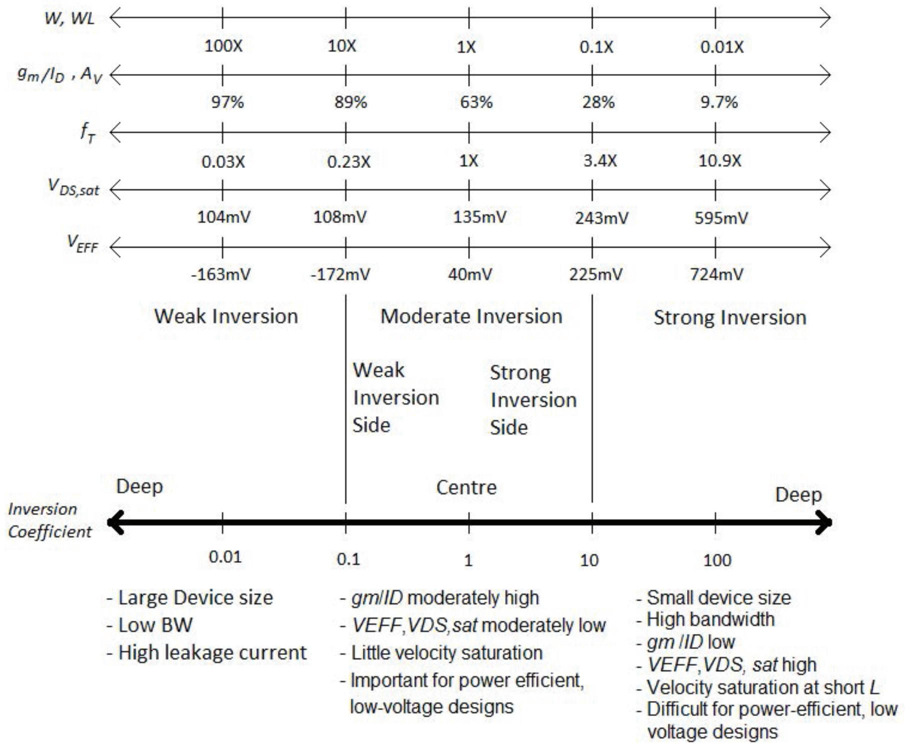
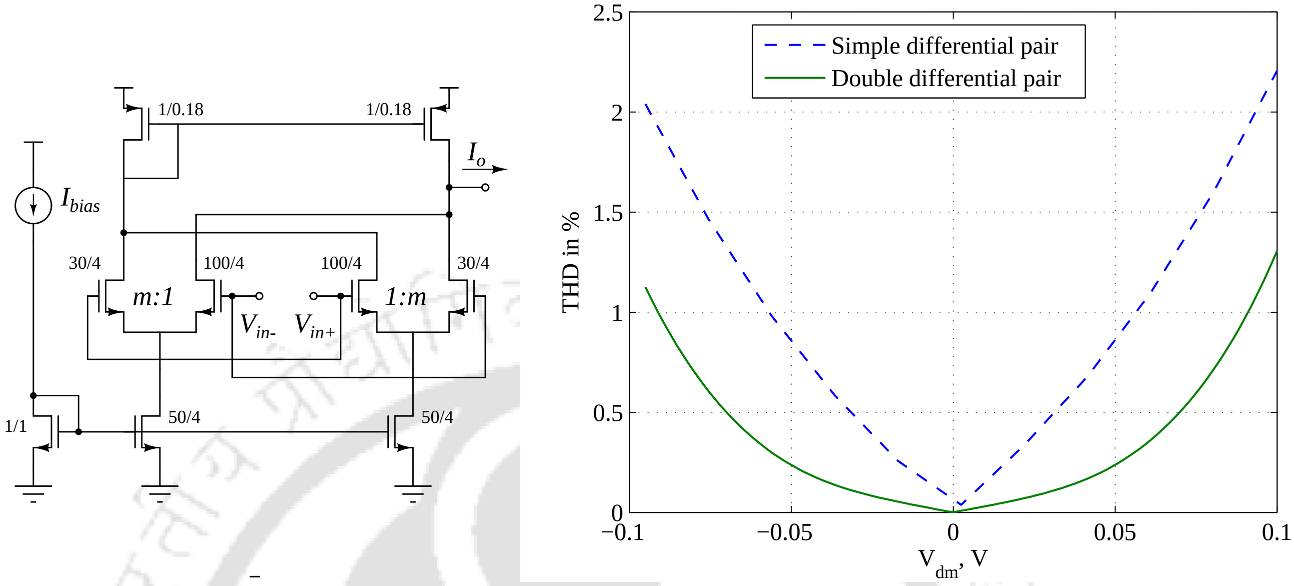


Figure 2.5: The inversion coefficient presented as a number line.

(iv) The output resistance of the devices at the input stage increases which results from the low drain currents. This further limits the bandwidth and allows the compensating capacitor to have a smaller value. While the low bias current of the first stage decreases the slew rate of the amplifier, the smaller value of compensating capacitor offsets this effect.

Examining performance tradeoffs over design choices of the inversion coefficient provides guidance on the optimal region and level of inversion for a given application. Figure 2.5 introduces the inversion coefficient and identifies the weak-, moderate-, and strong-inversion regions and subregions along with the corresponding V_{EFF} values. It presents design choices for the inversion coefficient along a number line and summarizes values of W, WL (gate area), g_m/I_D, A_{Vi}, f_{Ti}, $V_{DS,sat}$, and V_{EFF} for n held fixed at $n = 1.4$ and $U_T = 259$mV at $T = 300$K. The performance tradeoffs considered are for native devices having no resistive source degeneration. Source degeneration lowers or degenerates g_m, drain-referred thermal- and flicker-noise current, and drain mismatch current, while raising the drain output resistance. Additionally, source degeneration increases gate-referred thermal-noise voltage due to the presence of thermal noise in the source resistance and raises the drain voltage necessary for operation



(a) Double differential transconductor.

(b) Total Harmonic Distortion versus V_{dm}.

Figure 2.6: Topology of a double differential amplifier and comparison of total harmonic distortion with a single differential amplifier

in saturation due to the voltage drop across the source resistance. Device performance predictions must then be modified when source degeneration is present [47].

2.3 Transconductor For Bio-potential Amplifier

Recently, portable medical devices for health care management are required. This necessitates the design of a low-power and compact medical devices. For processing of biopotential signals, ultra low-power and fully-integrated active filters too are required. Since the biological signals are in the range of 1μV - 100 mV, while frequencies may range from 10 mHz up to 100 Hz, a large time constant is required for low frequency active filters in order to process the biological signal. For limitation of power and size, the operational transconductance amplifier used in active filters, is required to have a very low transconductance. If an active low-pass filter is designed whose cutoff frequency is lower than 10 Hz, then the OTA transconductance needs to be lower than 1 nA/V. Additionally, the transconductor should have a high linearity to avoid distortion of the biological signals.

OTAs with differential pairs biased in lower inversion regions such as weak or moderate inversion regions are preferred in biomedical applications. It provides low power consumption, a low g_m transconductance, a high (g_m/I_D) ratio with a reasonable performance. Because of these features, it finds suitable use in efficient implementation in low frequency continuous time filters, for example from sub-10 Hz frequency range up to 10 KHz. Continuous-time linear filtering for applications such as bionic ears, is one class of analogue circuits for which subthreshold CMOS design is more challenging. Traditional differential pairs operating in weak inversion offers a linear range of few tens of mV. Several techniques for extending the linear range of differential pairs in weak inversion such as source degeneration via resistor/diodes, attenuation and nonlinear cancelation have been suggested in the literature [48–50].

A two-transistor differential amplifier has a very limited input range when used as a linear transconductor in low power applications. When input differential voltage v_{dm} becomes greater than around 55 mV, the total harmonic distortion of the output current exceeds 1 percent as shown in Figure 2.6(b). The transconductor chosen for the design of the neural amplifier is based on the double differential pair topology [51]. It consists of two differential pairs having unequal sized transistors connected in parallel. The output is connected to a single current mirror with current bias, I_{bias} as shown in Figure 2.6(a). The basic idea of linearization is to give each differential pair operating in parallel an appropriate input offset voltage and a tail current in such a way that the combined input/output characteristic of those two pair has a flatter slope. An unequal-sized MOSFET transistors with different aspect ratios also behave the same as that of an equal-sized pair with a DC offset applied. This difference in size is defined as the scaling factor m which is used to increase linearity of the circuit. For a minimum distortion at zero differential-mode voltage $V_{dm} = V_{in}^{+} - V_{in}^{-}$, the scaling factor M is chosen as approximately four [51]. In this case the two drain current curves partially linearize each other when they are added together. In distortion analysis, it is observed that THD is a monotonically increasing function of differential input voltage V_{dm} which is shown in Figure 2.6(b). As compared to a single differential amplifier, the dynamic range of double differential structure is enhanced, allowing an increase of input differential voltage swing to 90 mV for achieving same distortion performance. The same technique can be extended to more number of input stages to further increase the linear input range.

2.3.1 Proposed scheme

Figure 2.7 shows the basic schematic of the proposed OTA operated in subthreshold. The proposed OTA consists of transconductor with two asymmetric differential pairs implemented by transistors M_{1a}-M_{1b} and M_{2a}-M_{2b} connected in parallel. The effect of unequal sizing of the transistor pairs is to

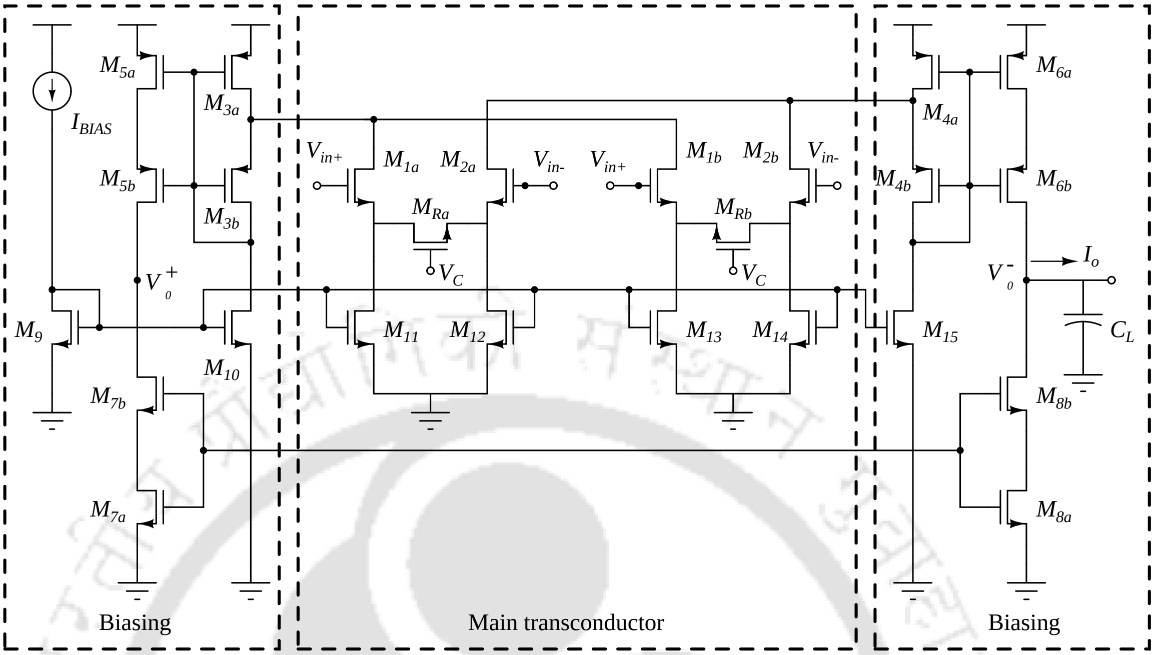


Figure 2.7: Subthreshold OTA with symmetric multi-stage balanced operation.

create an intentional voltage offset [34]. To further increase the input linear range, single diffusor is used as a means of source degeneration with each differential pair. The conductivity of the diffusor is determined by the respective W/L ratio and the applied gate potential, V_C. The proposed circuit is thus a large-signal transconductor, where transconductance g_m can be tuned through the control voltage V_C. The differential current I_0 can be obtained in terms of differential input voltage, $V_{dm} = V_{in}^{+} - V_{in}^{-}$ as

$$I_0 = I_{BIAS} \tanh\left(\frac{V_{dm}}{2nU_T} + \frac{1}{2}\ln m\right) + I_{BIAS} \tanh\left(\frac{V_{dm}}{2nU_T} - \frac{1}{2}\ln m\right) \tag{2.10}$$

where m is the relative W/L ratio of the transistor pairs. A possible criteria for optimizing the linear range is maximal flatness [30]. With one degree of freedom, we can set the first nonzero derivative of g_m equal to zero. Now setting second derivative equal to zero, we find that the only positive root occurs at $m = 2 + \sqrt{3}$. Approximating $m = 4$ introduces equal and opposite input differential voltage offset resulting in maximum nonlinearity cancelation in transistor pairs $M_{1a,b}$ and $M_{2a,b}$.

Without additional hardware, it can be observed with a single supply differential amplifier circuit that the active load is operable only when the common-mode (CM) voltage is near to the positive supply rail while it starts to switch off as CM voltage is close to the ground voltage. This results in a reduced input common mode range and causes signal distortion. To avoid distortion and to get the maximum signal swing requires a DC shifting of the input signal which can be done by either inserting a battery at input of the active load or by adding complementary symmetry transistors where the signal can be worked with rail-to-rail operation. The first method would require an additional circuitry simulating a battery while the second option may increase hardware complexity in order to avoid any dead zone condition.

A viable option is to use a two-transistor self-cascoded structure with a common gate configuration which can be treated as a single composite transistor. The composite transistor formed with two series connected n-type or p-type MOS transistor having the same well is an important configuration for subthreshold circuits [52] as depicted in Figure 2.8(a). Noting that $V_{DSa} = V_{GSa} - V_{GSb}$, equating the drain currents, I_{DSa} and I_{DSb} with voltage V_{DSa} applied so that M_b is in weak saturation, we obtain the following relationship:

$$V_{DSa} = U_T \ln\left[1 + \frac{(W/L)_b}{(W/L)_a}\right] \tag{2.11}$$

Note that the drain-source voltage of saturate transistor is independent of gate-source voltage. Moreover V_{DSa} is not effected by a changing V_{DSb}. The composite transistor structure is implemented in Figure 2.7 where pMOS device pairs M_{3a}-M_{3b}, M_{5a}-M_{5b} and M_{4a}-M_{4b}, M_{6a}-M_{6b} form a composite transistors while common gate transistors M_{3b} and M_{4b} function as level shifters. This forces drain voltage of differential pairs to be equal since drain voltages of M_{3a} and M_{4a} are equal and constant thus ensuring an optimal match for each of differential stages M_{1a}-M_{2a} and M_{1b}-M_{2b}. As the active load and the differential amplifiers are biased at the same potential therefore the voltage V_{DS3a} is

$$V_{DS3a} = U_T \ln\left[1 + 2\frac{(W/L)_{3b}}{(W/L)_{3a}}\right] \tag{2.12}$$

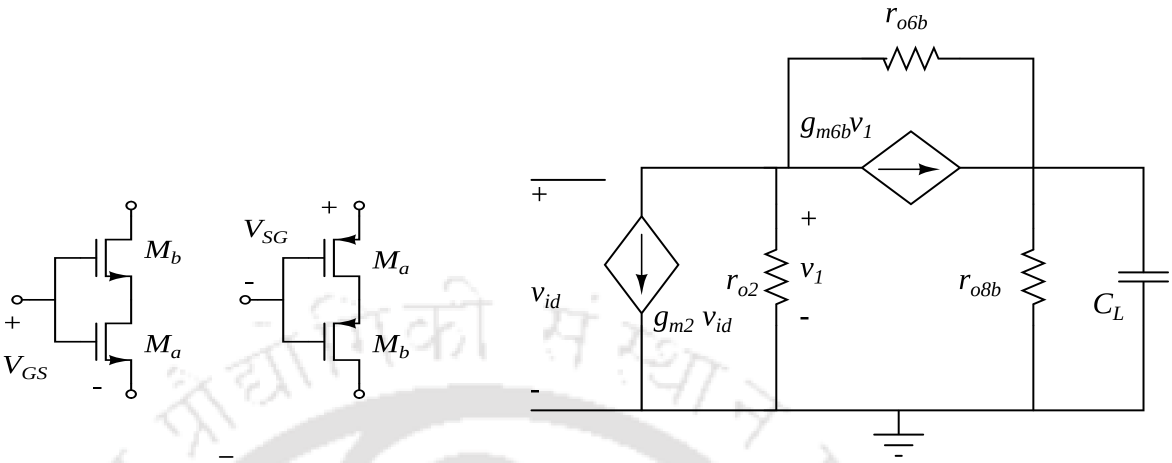
An analogous expression can also be obtained for V_{DS4a}. The expression for the third order harmonic distortion for $M_{1,2}$ can be derived by expanding equation (2.10) through Taylor series:

$$I_0 = \frac{8m}{(m+1)^2}\left[\frac{V_{dm}}{2nU_T} - \left(\frac{4m}{(m+1)^2} - \frac{2}{3}\right)\left(\frac{V_{dm}}{2nU_T}\right)^3\right] \tag{2.13}$$

Since HD_3 is defined as the ratio between coefficients of 3^{rd}-order harmonic and the fundamental

$$HD_3 = \frac{1}{4}\left|\frac{4m}{(m+1)^2} - \frac{2}{3}\right|\left(\frac{V_{dm}}{2nU_T}\right)^2 \tag{2.14}$$

By choosing $m = 4$, the maximum input differential signal can be computed for a given HD_3 specification. Thus the multi-tanh scheme gives better improvement in linearity as compared to a conventional



(a) n/p-type composite transistors.

(b) Small signal model of the proposed OTA.

Figure 2.8: Circuit schematic for composite transistors and small signal analysis of neural amplifier differential amplifier design.

2.3.2 Noise Analysis of G_M OTA

The input gate-referred thermal noise voltage density is mainly dominated by the self-cascode current mirror devices $M_{7,8}$ and the current drivers $M_{10,15}$ since one or more of these transistors are operating in strong inversion for the same drain current flowing through these devices. The input transistor pair, $M_{1,2}$ and the composite transistor pairs M_{3-6} which lie along the signal path also contribute to thermal noise. Noise from the tail current sources M_{11-14} and its external reference device is canceled-out since they are largely common-moded. The OTA input-referred thermal noise voltage, power spectral density (PSD) is given by equation (1.10).

$$\overline{v^2_{OTA,th}} = 2(4kT)\left[\frac{2(n\Gamma)_1 g_{m1}}{G_M^2} + \frac{(n\Gamma)_3 g_{m3}}{G_M^2} + \frac{(n\Gamma)_5 g_{m5}}{G_M^2} + \frac{(n\Gamma)_7 g_{m7}}{G_M^2} + \frac{(n\Gamma)_{10} g_{m10}}{G_M^2}\right] \tag{2.15}$$

Here kT is the product of Boltzmann's constant and absolute temperature. $(n\Gamma)_1$, $(n\Gamma)_3$, $(n\Gamma)_5$, $(n\Gamma)_7$, and $(n\Gamma)_{10}$ are the products of the substrate factor, n, and thermal noise factor Γ for respective devices where Γ is expressed as

$$\Gamma = \frac{\frac{1}{2} + \frac{2}{3} IC}{1 + IC} \tag{2.18}$$

Γ is expressed in terms of inversion coefficient factor IC of the EKV model [53], given by

$$IC = \frac{I_D}{2n\mu C_{ox} {U_T}^2 \,(W/L)} = \frac{I_D}{I_S\,(W/L)} \tag{2.19}$$

with symbols having their usual meaning. To obtain the noise level the input referred thermal noise density can be integrated over the bandwidth $\Delta f = f_2 - f_1$. Equation (2.15) integrated in the desired band of frequency, namely f_1 to f_2 gives

$$\overline{V^2_{OTA,t,th}} = \int_{f_1}^{f_2} \overline{v^2_{OTA,th}} df = \frac{8KT\Delta f}{G_M^2} \sum_i (n\Gamma)_i g_{mi} \tag{2.20}$$

where i=1,2,3,5,7,10. The input referred flicker noise voltage PSD for the OTA is given by

$$\overline{v^2_{OTA,f}} = \frac{2}{C_{ox}} \frac{1}{f} \left[\frac{2K_{f1}g^2_{m1}}{(WL)_1 G_M^2} + \frac{K_{f3}g^2_{m3}}{(WL)_3 G_M^2} + \frac{K_{f5}g^2_{m5}}{(WL)_5 G_M^2} + \frac{K_{f7}g^2_{m7}}{(WL)_7 G_M^2} + \frac{K_{f10}g^2_{m10}}{(WL)_{10} G_M^2} \right] \tag{2.21}$$

In the given equation K_{fi} is the flicker noise factor and is a process-dependent parameter, f is the operating frequency, and G_M is the output transconductance of the OTA given as:

$$G_M = \frac{g_{m2}}{1 + \frac{1}{r_{o2}\left(g_{m6b} + g_{go6b}\right)}} \tag{2.22}$$

Integrating flicker noise density over the desired bandwidth gives

$$\overline{V^2_{OTA,t,f}} = \int_{f_1}^{f_2} \overline{v^2_{OTA,f}} df = \frac{2}{C_{ox} G_M^2} \ln\left(\frac{f_2}{f_1}\right) \sum_i \frac{K_{fi} g^2_{mi}}{(WL)_i} \tag{2.23}$$

The total input referred noise of the OTA consisting of both thermal and flicker noise power spectral density integrated over the bandwidth defined by $\Delta f = f_2 - f_1$ can be computed as

$$\overline{V^2_{n,OTA}} = \sum_i \left[\frac{8KT\Delta f}{G_M^2} (n\Gamma)_i g_{mi} + \frac{2\ln(f_2/f_1)}{C_{ox} G_M^2} \frac{K_{fi} g^2_{mi}}{(WL)_i} \right] \tag{2.24}$$

As seen from the above equation, for low frequency applications the flicker noise component dominates the noise spectrum. The total input referred noise of the proposed G_M OTA is minimized by operating input devices $M_{1,2}$ in weak side of moderate inversion for high transconductance efficiency $g_{m1,2}/I_{D1,2}$ and high transconductance gain $g_{m1,2}$, while operating non-input devices, $M_{3-10,15}$ in strong inversion side of moderate inversion for low g_{m8}/I_{D8} and g_{m8}.

2.3.3 Common Mode Feedback for the G_M OTA

A CMFB circuit should ideally be designed to have a very small impedance for the common-mode signals but transparent for the differential signals. Taking care of this feature helps in reducing harmonic distortion components and improve THD and HD_3 response of the circuit. In a generic neural recording system, differential amplifiers are used to measure the potential of each signal electrode with respect to a large, low-impedance reference electrode. Such a differential amplifier should have a high common-mode rejection ratio to minimize interference from common-mode 60 Hz power line noise that is nearly equal at all locations on the body. Also a large common-mode input operating range is deleterious to low-power operation because it makes a large power-supply voltage necessary. Therefore as the amplifier is not intended for very large voltage swing the common-mode feedback is chosen to be implemented by differential-difference amplifier structure [54] which uses four identical transistors to average and compare the common-mode voltages.

The complete CMFB-OTA having a single supply voltage of V_{DD} = 0.8 V is demonstrated in Figure 2.9. The transconductance gain of CMFB is kept optimally higher in order to minimize the common-mode noise voltage of OTA. Nevertheless, transconductance gain can be reduced for input transistors $M_{16a,b,c,d}$ thereby reducing (W/L) ratios of the MOS transistors and by increasing the bias current source, I_b thus securing linearity to some extent. The difference between the common mode voltage and V_{ref} is amplified by the error amplifier M_{18} and output current is fed back to the gate of OTA source transistors. The non-dominant poles caused due to the internal nodes of CMFB appear at much higher frequencies than the dominant pole at f_d and can be neglected. Even though the topological structure of OTA circuit is a standard sub 1 V design suitable for low power biomedical signal processing, the sizing of the devices used in circuit plays crucial role for achieving low noise and low power simultaneously. The overall circuit of CMFB-OTA gives a core current consumption of 96.4 μA, with input device drain currents of 7 μA and power consumption of 77.1 μW which excludes bias reference current and voltage sources.

2.3.4 Self-cascode structure

The symmetrical OTA output stage is implemented with $M_{7a,b}$ and $M_{8a,b}$ self-cascoded structure which are meant to increase the output resistance without having extra voltage headroom. Self-cascode is the new technique, which does not require high compliance voltages at output nodes [55]. It provides high output impedance to give high output gain and so it is useful in low-voltage design. By using such composite structure at the output stage, the circuit can have much larger effective channel length and much lower effective output conductance. The lower transistor M_a is equivalent to a resistor, whose value is input dependent. For optimal operation, the W/L ratio of upper transistor M_b is kept larger than that of lower transistor M_a. For the composite transistor to be in saturation region M_b have to be in saturation and M_a in linear region. In the Figure 2.9, the aspect ratios of M_{7b}

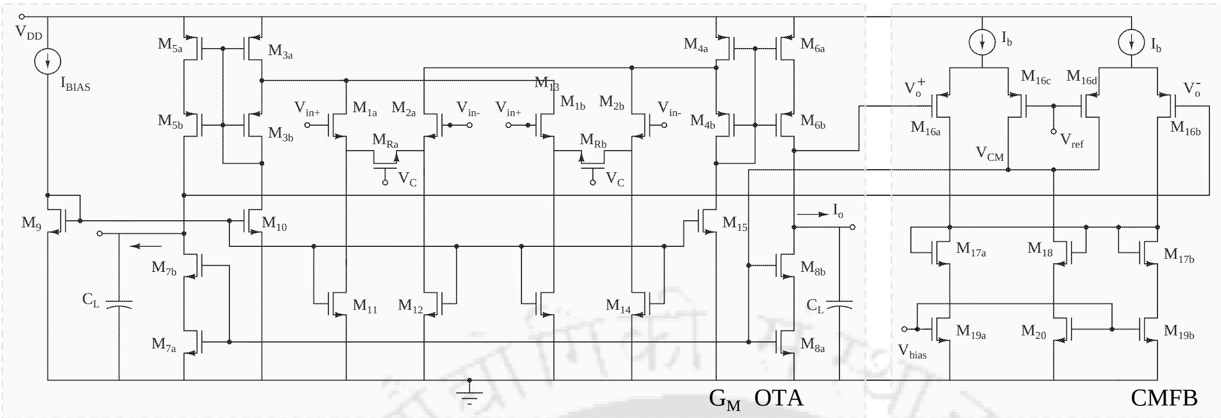


Figure 2.9: Proposed Common mode feedback symmetric OTA.

and M_{8b} cascode devices are kept much larger than that of M_{7a} and M_{8a} rail devices for obtaining high output impedance without sacrificing common-mode linear range operation suitable for low voltage application. From (2.22), if k_M is defined as $1/r_{o2}\,(g_{m6b} + g_{go6b})$, where $r_{o2} \gg 1/(g_{o6b} + g_{m6b})$, then $k_M \ll 1$. The output transconductance of OTA, G_M is considered nearly equal to the transconductance of the input pair devices because self-cascode structures have approximately unity current gain, hence $G_M \approx g_{m1} = g_{m2}$.

Maintaining constant bias currents and supply voltages permits optimization of each transistor of OTA to operate in weak side, center, or strong side of moderate inversion depending on their aspect ratios. For each device to work in their intended region of operation, the transistor sizing (W/L), saturation current (I_{sat}), channel length (L) and width (W) is determined from inversion coefficient factor (IC) described in [53]. The central part of OTA consists of a differential input pair implemented with a gate driven input topology. Gate-driven operation as compared to bulk-driven offers a higher transconductance and a low input referred noise which is desirable for high gain OTA circuits working in low voltage, low noise applications. Evaluating output resistance of the OTA in terms of k_M from AC model of OTA given in Figure 2.8(b) gives:

$$R_{out} = \frac{1 + k_M}{g_{08b}\,(1 + k_M) + g_{go6b} k_M} \tag{2.25}$$

Since the voltage gain, $A_{v,dc} = G_M \times R_{out}$, it can be deduced from equation (2.22) and (2.25) that the gain of the proposed OTA can be maximized by operating input pair $M_{1,2}$ devices towards weak inversion region (lower IC) while operating output devices M_{5-8} towards strong inversion region (higher IC).

2.4 Simulation Results of G_M OTA

The proposed symmetrical OTA terminated with a 7 pF of load capacitor was simulated in Cadence Spectre of standard UMC 0.18 μm CMOS process technology operated at $27°C$. The threshold voltages of nMOS and pMOS transistors are 315 mV and 498 mV respectively. The OTA design considers multi-tanh input transistor pairs $M_{1a,2a}$ and $M_{1b,2b}$ to be perfectly matched and working towards the weaker edge of moderate inversion region. Large transistor dimensions are maintained at the input (in multiple of $10\mu m/10\mu m$) to ensure minimal flicker noise pushing flicker noise corner frequency to sub-hertz range.

Table 2.1: Operating Point of devices in OTA

Devices	W/L (μm)	g_m/I_D (V^{-1})	IC
$M_{1a,2a}$	10/10	23.83	0.40
$M_{1b,2b}$	40/10	23.77	0.40
$M_{3a,4a,5a,6a}$	700/0.8	18.72	1.10
$M_{3b,4b,5b,6b}$	160/0.8	24.25	0.36
$M_{7a,8a}$	20/2	15.07	2.20
$M_{7b,8b}$	160/2	23.09	0.47
$M_{10,15}$	10/2	11.94	4.19

Measurement and characterization of MOS technology is provided by transconductance efficiency (g_m/I_D) and can be described as a quality factor for producing desired level of transconductance at a given level of drain current. It is maximum in weak inversion, decreases modestly in moderate inversion, continues dropping in strong inversion. (g_m/I_D) can be generalized in terms of inversion coefficient IC which is valid for all region of operation and is given by

$$\frac{g_m}{I_D} = \frac{1}{nU_T\left(\sqrt{IC + 0.25} + 0.5\right)} \tag{2.26}$$

The operating region can be defined with respect to the inversion coefficient; the device operating in weak inversion will have $IC < 0.1$, for device operating in moderate inversion, $0.1 < IC < 10$, while for the devices operating in strong inversion, $IC > 10$. Table 2.1 shows device dimensions and operating point of MOS devices in the proposed OTA circuit. Figure 2.10 shows the frequency response of

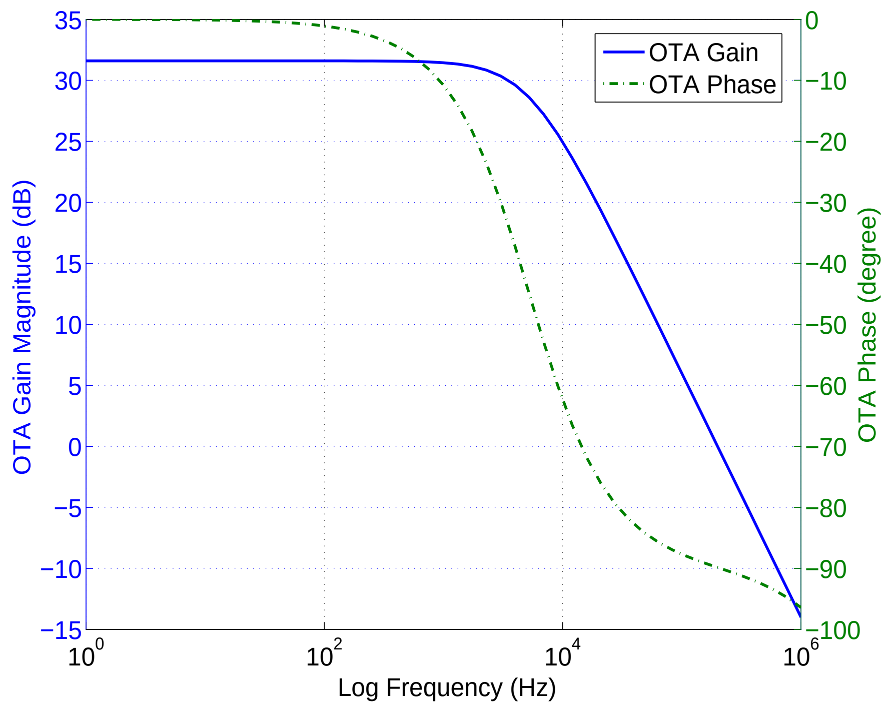


Figure 2.10: Simulation result of the gain and phase plot of symmetric G_M OTA.

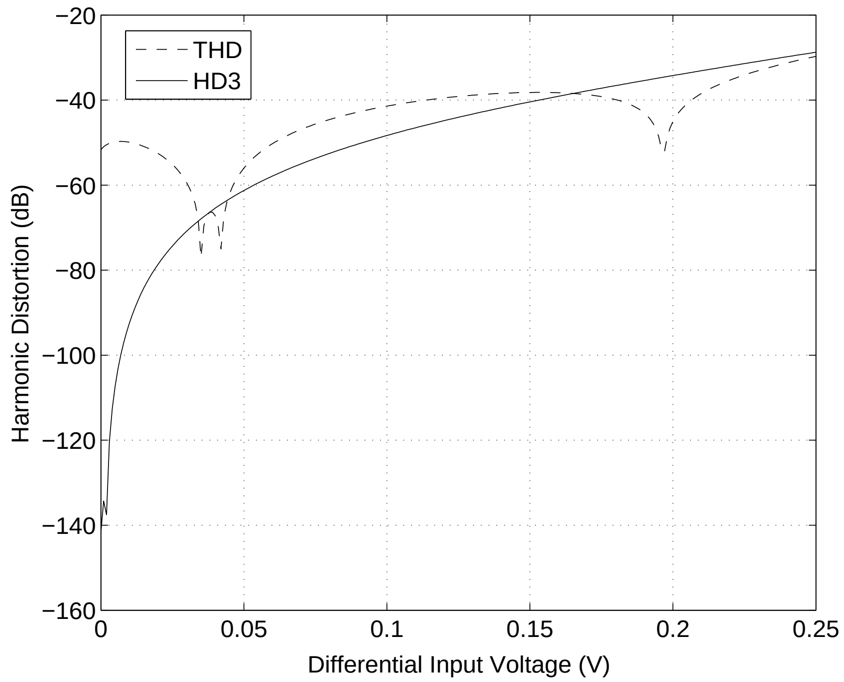


Figure 2.11: THD and HD3 simulation of symmetric OTA.

OTA-CMFB structure having the open loop DC gain as 31.6 dB and a unit gain bandwidth of 202.3 KHz with 90.18° of phase margin.

The OTA-CMFB circuit draws 96.38 μA of current operating with a single supply of +0.8V. The linear input range is extended by using DC shifting and current cancelation technique. The measured input linear range as shown in Figure 2.11 is 157 mV for $HD_3 \leq 1\%$ which is sufficient for neural recording and can be suited to other biomedical applications [56] [57]. The maximum signal swing was observed from 0 to 500 mV while the measured transconductance gain was 17.85 μA/V. The G_M value is relatively high because of the increased $I_{D1,2}$ since it is the product of g_m/I_D and I_D where g_m/I_D ratio for input devices is chosen high owing to the operation near weak inversion region.

2.5 A Neural Amplifier using proposed transconductor

Neural signal from extra-cellular recording are very weak in amplitude (varying from 10 μV to 500 μV). As a result amplification is needed before they can be further processed. Since neural recording system is a classical problem in neuro-scientific studies, many different neural amplifier topologies have been suggested in literature such as [58–60]. A low noise pre-amplifier is required for increasing the signal level from few μVs to more than a mV with minimal addition of noise. Such amplifiers are typically designed to be AC-coupled in order to remove DC electrode offset of the order of 10 mV depending on type of electrode used. These amplifiers have typical gains ranging from 50−200, with the bandwidths varying from 4−10kHz, and input-referred noise from 2−10 μV [23, 61].

2.5.1 Neural Amplifier Design

To measure the validity and feasibility of the proposed OTA topology, a preamplifier using our OTA was designed for neural prosthetic application. A neural amplifier is conventionally implemented either as a double-ended (DCCA) [62] or as a single-ended capacitively coupled (SCCA) preamplifier [63]. We have used the latter approach for our preamplifier design as depicted in Figure 2.12. It consists of a closed-loop gain along with a low-frequency pole using a capacitive feedback network and a combination of MOS-bipolar pseudo resistor M_{a-c} which is made highly resistive through proper gate voltage biasing. The pseudo resistor elements are used because they offer a very high incremental resistance of the order of 10 GΩ and occupy a very small silicon area which can be used to set a very low high-pass corner frequency ($\sim$ 1 Hz) suitable for bio-medical circuits. Here these pseudo resistors are replaced by the equivalent resistance r_{eq} whose value is set by the tuning voltage V_{tune}. Capacitor

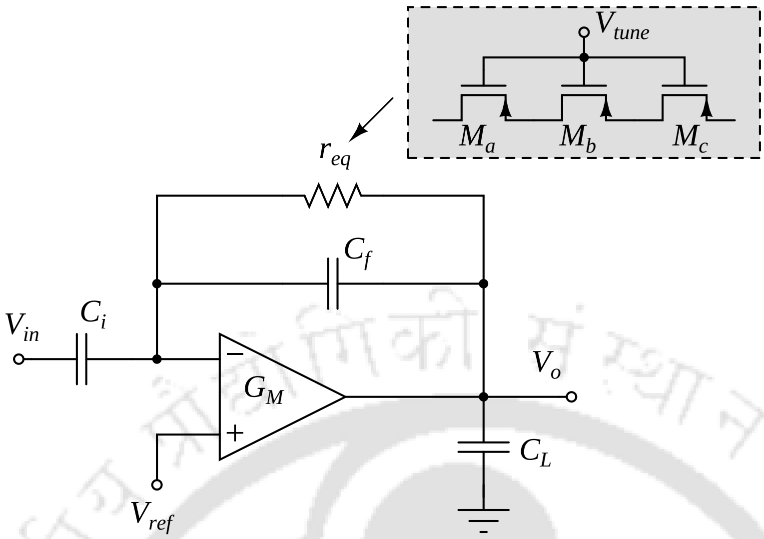


Figure 2.12: Circuit schematic of single-ended capacitively coupled neural amplifier with pseudo-resistive feedback resistor r_{eq}.

C_i is used to filter DC baseline interference and C_f sets the lower cut-off frequency. The closed-loop gain of the amplifier is $A_v = -Z_f/Z_i$, where $Z_i = 1/sC_i$ and $Z_f = r_{eq}/(1 + sr_{eq}C_f)$. Using Laplace transform the voltage gain can be expressed as

$$A_v(s) = -\frac{sr_{eq}C_i}{1 + sr_{eq}C_f} \tag{2.27}$$

Since $r_{eq} \gg 1/sC_i, 1/sC_f$, the transfer function of the neural amplifier can be approximated as

$$A_v(s) \approx -\frac{C_i}{C_f} \tag{2.28}$$

The lower cut-off frequency of bandwidth is given by

$$\omega_L = \frac{1}{\mathrm{r}_{eq}C_f} \tag{2.29}$$

Thus ω_L can be tuned using this gate voltage V_{tune} to control the impedance in the feedback and enabling selection or rejection of slow wave action potentials. The low-pass cutoff frequency due to the loading effect at the output of the G_M OTA is governed by the dominant pole frequency ω_U.

$$\omega_U = \frac{G_M}{A_vC_L} \tag{2.30}$$

where A_v is the mid-band gain.

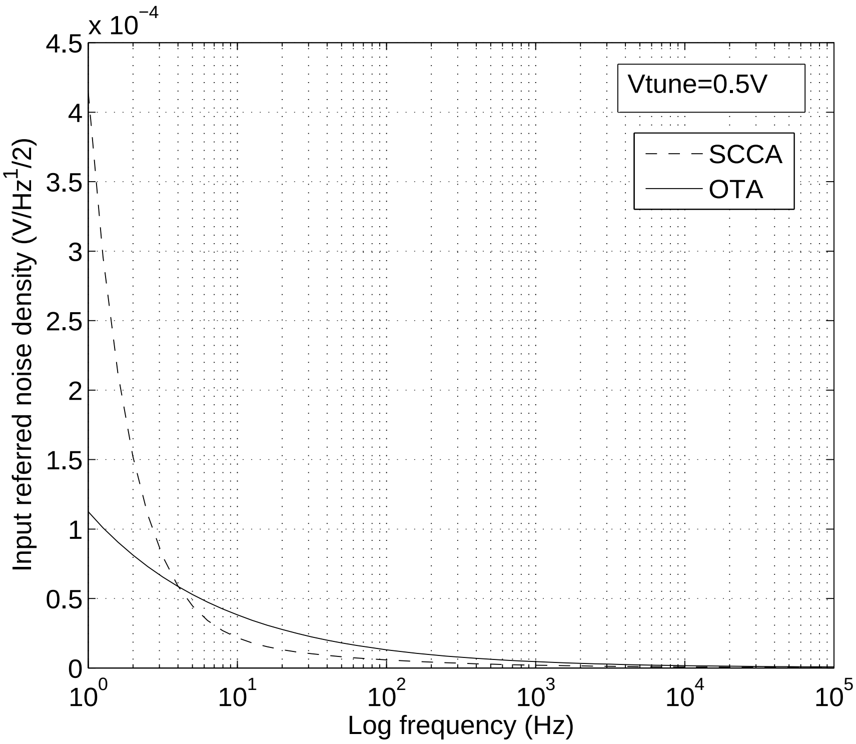


Figure 2.13: Input-referred noise density of OTA and SCCA.

2.5.2 Simulation Results of Neural Amplifier

Thermal noise of MOS device is nearly unchanged as channel length increases since G_M has little effect on changing L but the flicker noise decreases as the square of increasing channel length. Thus optimization of input referred noise can be done by increasing gate area of input devices while maintaining $I_{D1,2}$ and IC (corresponding to $g_m/I_D > 23$) at a constant level. The noise spectral density at low frequency rolls-off as $1/f^2$ instead of $1/f$ if it is a flicker noise. An input-referred noise spectral density of our SCCA configured neural amplifier shows a significant amount of low-frequency noise as compared to that of G_M OTA as shown in Figure 2.13. The consideration of the noise contribution due to pseudo-resistor MOS transistor in Figure 2.12 is important in the design of neural amplifier as it appears at the front-end stage of the amplifier which is the most critical stage. To analyze its noise, let's calculate a part of the amplifier's input-referred noise that is contributed by pseudo-resistor MOS transistor, M_a. A circuit schematic illustrating this situation is shown in Figure 2.14. For simplicity, let's assume that M_a have the noise current density as $\overline{i_{nM}^2}$ and incremental resistance r_{eq} since they are biased at the same operating point. The input-referred noise contributed

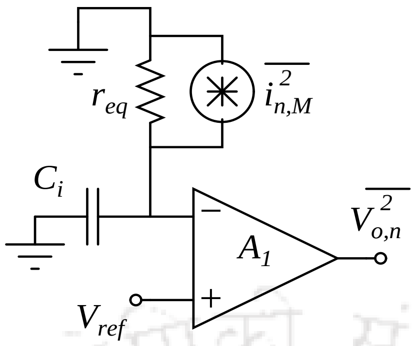


Figure 2.14: A schematic for calculating the input-referred noise from noise of M_a.

by M_a is calculated to be

$$\overline{v^2_{n,in}} = \overline{i^2_{n,M}} . \left(\frac{r_{eq}}{1 + sr_{eq}C_i} \right)^2 \tag{2.31}$$

The transistor M_a is biased in subthreshold such that the high-pass cutoff frequency of the amplifier is below 1 Hz. Therefore, the thermal noise in M_a can be approximated by [42]

$$\overline{i^2_{n,M}} = 2qI_D \left(1 + e^{-V_{ds1}/U_T}\right) \approx 4qI_D \tag{2.32}$$

since their drain source voltage V_{ds} is approximately zero and I_D is the channel current of M_a if it is in saturation. For the frequency range of $f \gg 1/(2\pi r_{eq}C_i)$, the input-referred noise in (2.31) can be approximated by

$$v^2_{n,in} = \frac{4qI_D}{(2\pi)^2 . C_i^2} . \frac{1}{f^2} \tag{2.33}$$

which agrees well with the $1/f^2$ rolloff at low frequency of the noise spectral density in Figure 2.13.

By performing noise analysis as shown previously it was evident that the low frequency noise that rolls-off as $1/f^2$ is due mainly to the filtering of the thermal noise in M_{a-c}. Therefore in order to reduce this low-frequency noise in the pass-band, the saturation current of M_{a-c} should be made very small, thus M_{a-c} should have as small gate-source voltage as possible. By setting the gate-source voltage of M_{a-c} such that the high-pass cutoff frequency of the amplifier happens at a very low frequency, the thermal noise of these pseudo-resistor MOS can be filtered out well before the frequency band of interest.

Simulation results of the input referred noise power spectral density is shown in Figure 2.15. With the positive gate voltage of $V_{tune} = 0.5$ V applied to transistors $M_{a,b,c}$ the parasitic BJT behavior sets in the lateral path along source, well and drain giving high incremental resistance of the order of peta-

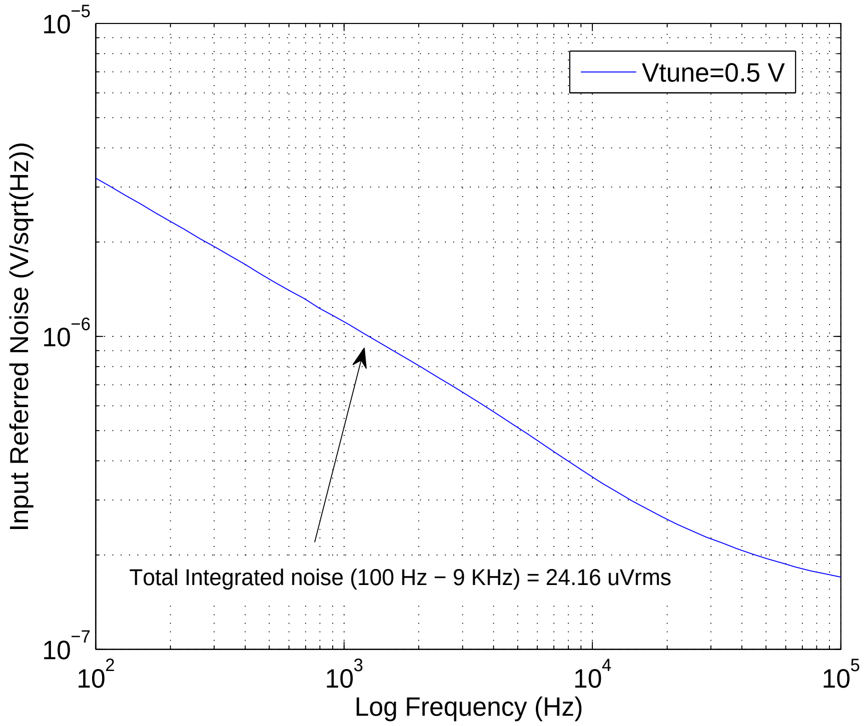


Figure 2.15: Input-referred noise density simulation by Cadence Spectre.

ohms. The noise calculation was done by using flicker noise parameters from the parameter extraction of nMOS and pMOS transistors using BSIM3v3 model and were evaluated as $K_f = 1.06 \times 10^{-25}$ V^{2}F for NMOS and $K_f = 9.07 \times 10^{-27}$ V^{2}F for PMOS. The integrated input referred noise was found to be 24.16 $\mu V/\sqrt{Hz}$ over (0.1 - 9.0) KHz bandwidth.

The integrated noise voltage can further be decreased by tuning input device drain currents to higher values since it is inversely proportional to the increasing drain current for a fixed inversion coefficient and channel length. However, thermal noise could not be lowered beyond a limit since the bias current was kept in control in order to limit overall power consumption within 77.1 μW. This meets the requirement of multichannel neural recording devices with modern MEMS arrays (like the one described in [64]) providing approximately 100 electrodes and a power dissipation limit of 10 mW, each channel must consume less than 100 μW, although it doesn't include shared resources on a chip such as analog-to-digital conversion, power regulation, control, and telemetry circuits.

Even though the design using proposed OTA topology achieves maximum power-noise trade-off, the OTA used was actually not very power efficient since large portion of total current was wasted in current mirrors (Appendix 3.3). This is why W/L ratios should be chosen with utmost care so as to

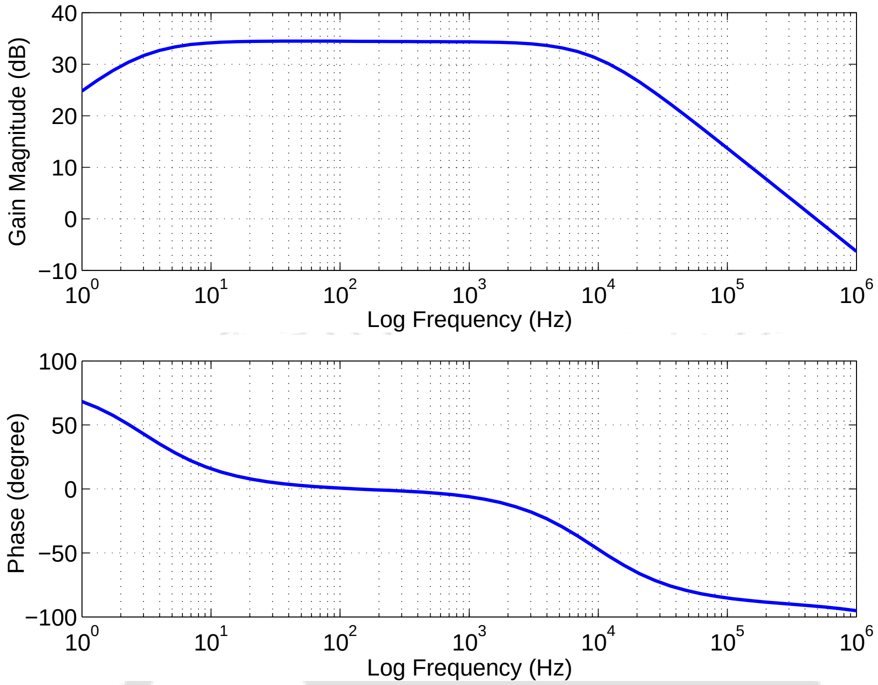


Figure 2.16: Gain and phase plot of the neural preamplifier.

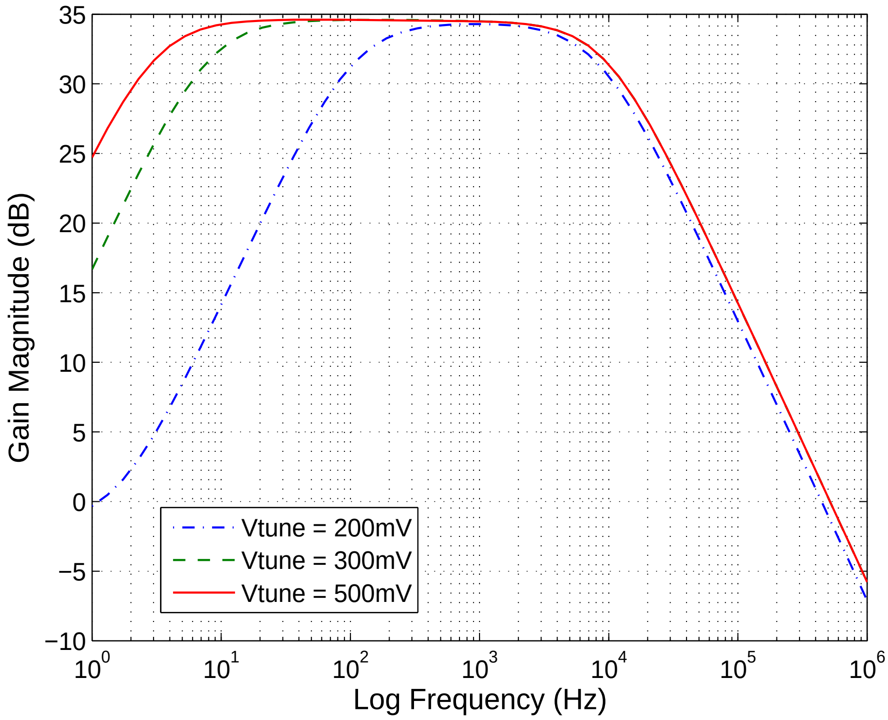


Figure 2.17: Frequency response of the amplifier with three different settings of the tuning voltage V_{tune}.

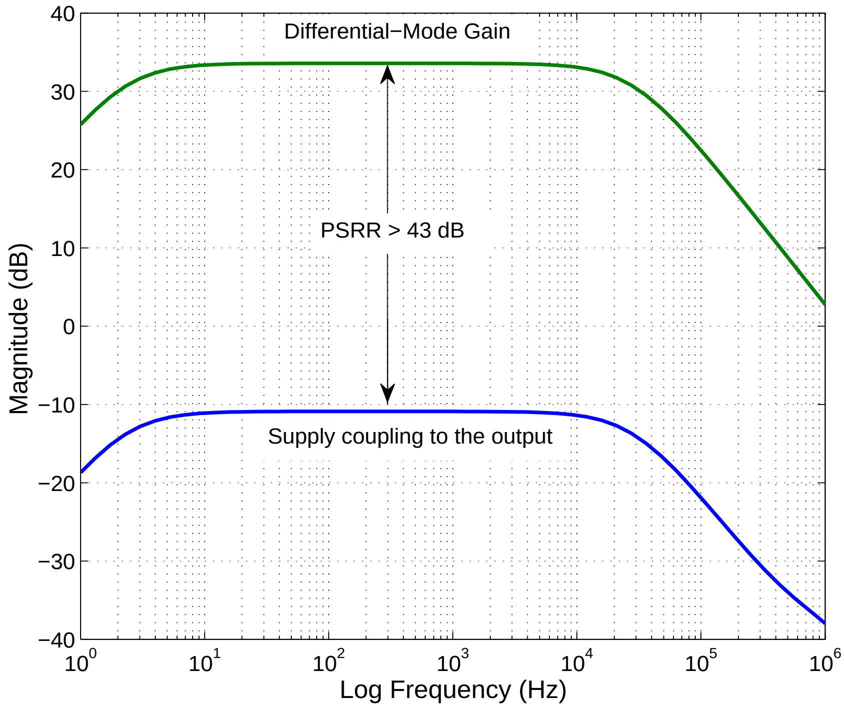


Figure 2.18: PSRR response of the neural preamplifier.

obtain minimal of noise without sacrificing much of supply current and hence the power consumption.

Simulation results of the frequency response of the preamplifier transfer function is shown in Figure 2.16. With the input capacitor $C_i = 6$ pF and the capacitor in feedback $C_f = 100$ fF, the mid band gain was found to be 34.5 dB over a bandwidth of 2.9 Hz to 9.2 KHz. The lower 3-dB frequency can be tuned via V_{tune} voltage by varying the effective channel conductance in the feedback loop. Figure 2.17 shows the transfer function of this amplifier measured at two gate voltage control settings of V_{tune}, resulting in high-pass poles at 3.0 Hz and 7.8 Hz while the low-pass corner frequency and the flat band gain remaining unaltered. The total harmonic distortion (THD) stays below 1% for differential input less than 9.11 mV. This level of input signal is larger than a typical action action potential ($<$ 500 μV). For larger input amplitude, THD can be increased by raising the supply voltage and trading with power consumption. The simulated result of power supply rejection ratio is calculated as the ratio of the differential-mode gain to the gain from power supply to the output. The common mode rejection as shown in Figure 2.18 was measured and was found to be better than 43 dB. Substituting the simulated values of total current, bandwidth, and input referred noise, the noise efficiency factor (NEF) of the neural amplifier was found to be 91.4. The layout of the amplifier occupies an estimated

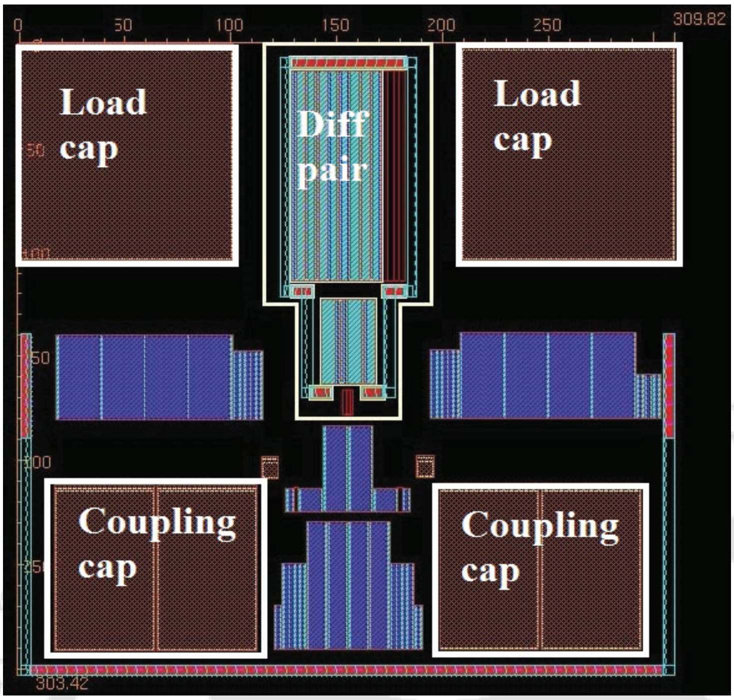


Figure 2.19: Layout of the designed neural preamplifier.

area of 0.094 mm^2 and is depicted in Figure 2.19 (pads and routing details not shown). A 100-channel system made from our amplifier will consume a silicon die area of less than 10 mm^2 in a 0.18-μm process and power of only 7.71 mW allowing for large scale implantable neural recording system and which is also at par with the existing VLSI bioamplifier design standards [62, 64, 65].

The comparative performance of neural recording amplifiers based on fully-integrated design is shown in Table 2.2. The amplifiers taken for comparison in Table 2.2 were designed in CMOS with the power consumption per channel in the range of microwatts with no off-chip components. The CMOS design in this work provides the lowest power supply design having comparable power level and the overall gain with very low distortion components.

2.6 Conclusion

A fully differential OTA with common-mode feedback was designed as a basic amplifier for neural pre-amplification. Current cancelation, source degeneration, and DC-shifting techniques were used for linearity improvement of the OTA while self cascode structure was employed at the output stage to further enhance the DC gain. A 0.8 V CMOS biosignal amplifier used in neural recording application was successfully designed and simulated. The single-ended capacitively-coupled neural preamplifier

Table 2.2: Summarized Performance and Comparison with CMOS Neural Amplifiers

Parameter	[66]	[67]	[62]	This work
CMOS Technology	1.5μm	0.35μm	1.5μm	0.18μm
Supply Voltage	1.5V	3.0V	±2.5V	0.8V
DC Gain	42.5dB	38.1dB	39.5dB	34.5dB
Operating frequency	22Hz-6.7KHz	1.4Hz-8.5KHz	25mHz-7.2KHz	3Hz-9.2KHz
Input referred Noise	20.6μV_{rms}	14.4μV_{rms}	2.2μV_{rms}	24.16μV_{rms}
Noise BW	10Hz-10KHz	1.4Hz- 8.5KHz	0.025Hz-7.2KHz	100Hz-9KHz
THD	–	1%@2.4mV_{pp}	1%@12.4mV_{pp}	1%@9.1mV_{pp}
Power Consumption	0.8μW	6μW	80μW	77.1μW

configured for recording action potentials gave an input-referred noise of 24.16 μV while consuming 77.1 μW of power and having flat-band gain of 34.5 dB over a frequency range of 2.9 Hz to 9.2 KHz. By maintaining proper g_m/I_D ratios, input MOS devices in the proposed design are operated at the edge of weak and moderate inversion region and non-input devices towards strong inversion region so as to achieve maximum power efficiency and an optimal flicker noise voltage PSD while operating on a limited power supply. The power supply can also be increased to more then 1 V in order to maintain large overdrive voltages in some above-threshold transistors to minimize their noise contributions. The proposed OTA topology can be employed for recording neural signal from 3 Hz to 9 KHz in low voltage, low frequency, and low THD neural preamplifiers.

3

A Nano-power Instrumentation amplifier for LFP Recording

Contents

3.1 Trends in Neural Sensors IC Design

Single-unit action potential or neural spikes supplements for the most direct information about a neuron offering the highest signal fidelity. This approach, however comes at the cost of increased safety risks owing to higher invasiveness as compared to other sensor modalities like electroencephalography (EEG) and electrocorticography (ECoG). With the same electrodes that used to measure single-unit spike activity (SUA), low-frequency signals in 1 Hz-200 Hz range known as local field potentials (LFP) are often persistent, even after single-unit activity has subsided, and has proved to be useful in some brain-machine interfaces (BMI), e.g, those used in paralysis prosthetics [68]. The intracranial population LFP signals are more robust for chronical recordings since they are less susceptible to tissue fibrosis around electrode tips which causes increase in impedance between neuron and electrode leading to signal degradation over a period of time. Few studies have also conjectured that LFPs can be measured even in the absence of spikes [69]. Unlike single-unit activity where spike-sorting is inevitable, LFPs can be decoded directly from the raw signal [70].

Field potentials recorded on dura mater or subdurally are localized within 250-400 μm in the vicinity of recording micro-electrode where >95% of LFP signal originates [71]. Therefore multi-electrode systems for recording LFPs placed entirely within the skull would have to incorporate a large number of neural amplifiers (of the order of 1000, with each signal electrode having a dedicated neural amplifier). Investigations show that while spikes require fewer channels in an implanted electrode array to encode direction of hand reaching movements, the behavioral information can by encoded by LFPs with fewer channels than that of spike data [68]. By simultaneously recording neuronal activity from a large number of neurons, the mapping of neural codes becomes more effective, resulting in a better decode performance thus allowing the neural prosthetic device to reach 3D targets more accurately. For implementing such high density multi-electrode systems, the heat dissipation in the brain should be minimized which makes ultra-low power and area-efficient amplifier design more challenging. Since LFP carries important information [72, 73], the LFP activity must be isolated from the neural signal before spike detection is performed to identify time of spike events of individual units. Hence LFP band can be segregated from neural spike for feature extraction at the front-end circuit by filtering in 10-100 Hz range.

The current trend in the design of a neural sensor integrated circuit is to simultaneously achieve a minimum noise efficiency factor and low power dissipation which is governed by the input referred noise

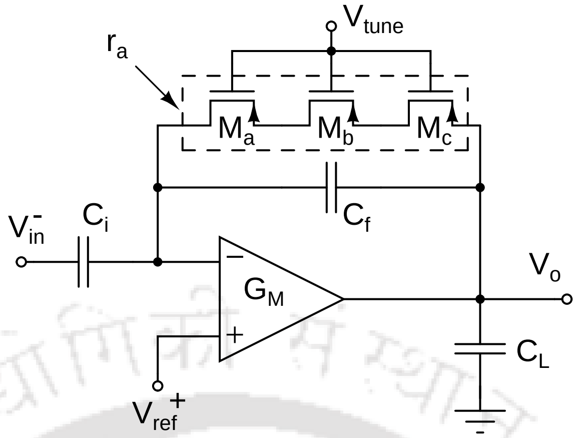


Figure 3.1: Schematic of a neural amplifier for LFP recording.

and quantization resolution of analog-to-digital converters (ADC). These constraints in the specifics of individual circuit blocks of the analog front-end ultimately restrict the number of recording channels in an implantable neural recording microsystem [29, 74, 75]. Motion artifacts and interferences can also cause electronics to saturate thus further posing challenge on front-end microelctronics. To overcome these limitations the neural amplifier system programmed to record either spikes or LFPs must exhibit a wide dynamic range. In this work the design of an ultra-low power neural recording amplifier configured for recording local field potential is presented. Under a specified power and area limitation the amplifier design is optimized to have an output dynamic range large enough to convey LFPs in the range of $\pm 1 - 10$ mV in amplitude and to meet the subsequent ADC quantization resolution requirement and also to make the amplifier system more robust to artifacts.

In the previous chapter, it was observed that despite increasing allowable signal swing, the disadvantage of using continuous-time CMFB circuit was that the common-mode detector reduced the DC gain. Moreover, transistors used in common mode circuit contribute to the thermal noise level in addition to the power consumption of the overall circuit. Since bio potential signals are weak in nature and indistinguishable from electrode noise, input-referred noise is more crucial to the first stage of neural pre-amplification. Also, because the bio-potential signal has a relatively small signal swing (of the order of millivolts) therefore by trading output linear range with the input-referred noise and power consumption, the proposed OTA can be redesigned with a single-ended topology while omitting the CMFB circuit. This would help achieving a better compromise between power, noise, and circuit area. In this chapter, a 0.8 V symmetrical operational transconductance amplifier (OTA) is designed which is implemented with two differential pairs in parallel asymmetric multi-tanh configuration. The OTA structure utilizes a DC shifting technique which exhibits linearity for a wide common-mode input range. The design methodology and noise analysis used in the OTA is also described which serves as the main building block of the bioamplifier for achieving an efficient power-noise-area tradeoff.

3.2 Low power Low noise Neural Amplifier

For recording bio-potential signals a capacitively coupled amplifier is commonly adopted for the neural front-end preamplification. Whereas a double-ended capacitively coupled input (DCCA) [62] is a typical implementation of neural amplifier we have resorted to single-ended capacitively coupled (SCCA) configuration [63] for our amplifier design as depicted in Figure 3.1. The capacitive feedback network consists of MOS pseudo resistor elements $M_{a,b,c}$ which is made highly resistive through a suitable gate voltage biasing creates a high pass corner frequency given by

$$\omega_L = 2\pi f_L = \frac{1}{\mathrm{r}_a C_f} \tag{3.1}$$

where C_f is the capacitor in feedback loop and r_a is the equivalent resistance of series connected bipolar-MOS elements. The variable gate voltage V_{tune} of MOS pseudo-resistors also provide band-tunability for recording cortical LFPs in different activity regions [68]. The neural recording signals are connected to the single-ended neural preamplifier via input capacitor C_i in order to decouple large DC polarization signals arising from electrode and tissue junction and to enable the neural signal of interest. This also removes any requirement of large off-chip components since high-resistance elements can be implemented in a smaller chip area. The mid-band gain is given by

$$A_v = \left|\frac{v_0}{v_{in}}\right| = \frac{C_i/C_f}{1 + \frac{1}{A_{v,dc}}\left(1 + \frac{C_i}{C_f}\right)} \approx \frac{C_i}{C_f} \tag{3.2}$$

where $A_{v,dc}$ is the DC gain of the transconductor used to realize the neural amplifier while the dominant pole ω_U due to load capacitor C_L governs the low-pass corner frequency.

$$\omega_U = \frac{G_M}{A_v C_L} \tag{3.3}$$

3.2.1 Neural Amplifier's OTA Topology

Figure 3.2 shows a schematic of the single-ended OTA operating in weak inversion region. The circuit topology consists of a transconductor implemented by two asymmetric transistor pairs MN_{1a}-

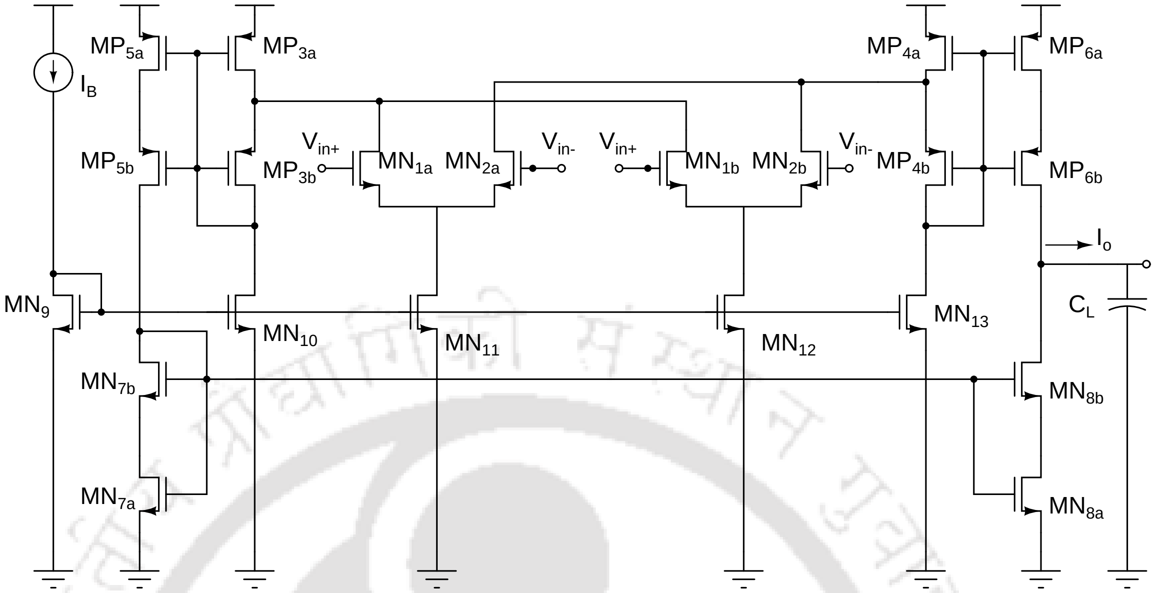


Figure 3.2: Schematic of the single-ended low-power OTA.

MN_{1b} and MN_{2a}-MN_{2b} connected in parallel with equal tail currents. The transistor pairs are unequally sized so as to create an intentional voltage offset resulting in a minimum distortion level. The differential current I_0 of the multi-tanh doublet in terms of differential input voltage, $V_{dm} = V_{in+} - V_{in-}$ is given as

$$I_0 = I_B \tanh\left(\frac{V_{dm}}{2nU_T} + \frac{1}{2}\ln m\right) + I_B \tanh\left(\frac{V_{dm}}{2nU_T} - \frac{1}{2}\ln m\right) \tag{3.4}$$

where the constant scalar m is the relative aspect ratios of the transistor pairs. By taking approximate value of $m = 4$ for maximally flatness condition introduces opposite input voltage offset resulting in maximum nonlinearity cancelation in transistor pairs $MN_{1a,b}$ and $MN_{2a,b}$ [30].

The composite transistor structure sharing the same well [52] is implemented in Figure 3.2 where pMOS device pairs MP_{3a}-MP_{3b}, MP_{5a}-MP_{5b} and MP_{4a}-MP_{4b}, MP_{6a}-MP_{6b} form a composite transistors while transistors MP_{3b} and MP_{4b} function as level shifters. The constant and equal drain voltages of MP_{3a} and MP_{4a} causes drain voltage of input differential pairs to be equal thus ensuring an optimal match for differential stages MN_{1a}-MN_{2a} and MN_{1b}-MN_{2b}. With the voltage V_{SD3a} of composite transistor applied such that MP_{3b} operates in weak saturation the voltage V_{SD3a} is

$$V_{SD3a} = U_T \ln\left[1 + 2\frac{(W/L)_{3b}}{(W/L)_{3a}}\right] \tag{3.5}$$

By employing composite transistor structure the need for common-mode circuit is also eliminated.

Also the offset bias scheme increases the linear range and dynamic range. It also minimizes the effective number of transistors that contribute noise. The output stage of OTA is implemented with $MN_{8a,b}$ self-cascoded structure. The W/L ratios of MN_{7b} and MN_{8b} cascode devices are kept much larger than MN_{7a} and MN_{8a} rail devices which reduces channel length modulation and enhances DC gain without losing on the output voltage swing suitable for low voltage application.

3.2.2 AC model

The AC model of the proposed OTA is represented with equivalent small signal model of all elements along signal path. By taking advantage of the symmetry, the proposed OTA of Figure 3.2 can be replaced with the π equivalent model as shown in Figure 3.3. G_M is defined as the output transconductance of OTA and is given as:

$$G_M = \frac{g_{m2}\left(g_{m6b} + g_{o6b}\right)}{g_{m6b} + g_{o6b} + g_{o2}} \approx g_{m2} \tag{3.6}$$

As seen from (3.6), G_M is nearly equal to the transconductance of the input pair devices because current mirrors consisting of self-cascode structures route signal currents from the input pair drains to the OTA output with nearly unity current gain. Evaluating output resistance of the OTA gives:

$$R_{out} = \frac{g_{m6b} + g_{o6b} + g_{o2}}{g_{m6b}g_{o8b} + g_{o2}g_{o6b} + g_{o8b}g_{o2} + g_{o6b}g_{o8b}} \tag{3.7}$$

It is clear from the voltage gain, $A_{v,dc}$, which is product of the G_M from (3.6) and R_{out} from (3.7) that $A_{v,dc}$ of the proposed OTA can be maximized by operating input pair devices at low inversion coefficients when biased more towards weak inversion region for high $g_{m1,2}/I_{D1,2}$ and $g_{m1,2}$, also combined with operating output devices M_{5-8} at long channel length for obtaining lower output conductance. From small signal analysis the input tranconductance is given by $g_{m2} = g_{m2a} + g_{m2b}$ while the output resistance due to channel length modulation is $r_{o2} = 1/g_{o2} = (r_{o2a} \parallel r_{o2b})$. The DC gain can be obtain by ignoring all capacitances in the circuit.

$$A_{v,dc} = \frac{g_{m2}}{g_{o2}\left(\frac{g_{o6b}+g_{o8b}}{g_{m6b}+g_{o6b}}\right) + g_{o8b}} \tag{3.8}$$

The OTA's dominant pole frequency is found to be

$$\omega_d = \frac{g_{m2}}{A_{v,dc}C_L\left(1 + \frac{g_{o2}}{g_{m6b}+g_{o6b}}\right)} \tag{3.9}$$

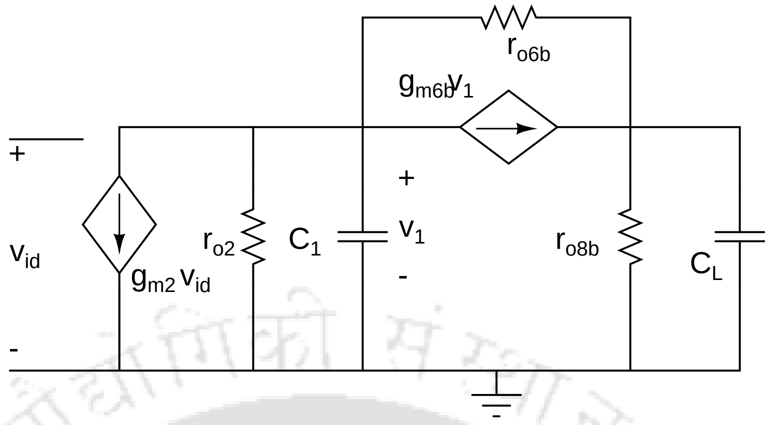


Figure 3.3: Small signal model of symmetric OTA for AC analysis.

The parasitic capacitances due to the increased number of nodes caused by level shifter arrangement do not deter OTA's dominant pole location and is still decided by the load capacitor C_L. As $g_{o2} \ll (g_{m6b} + g_{o6b})$, the dominant pole frequency can be approximated to $f_d = g_{m2}/2\pi A_{v,dc} C_L$.

3.2.3 Noise Analysis of OTA

The input transistor pair, $MN_{1,2}$, self-cascode current mirror devices $MN_{7,8}$, current drivers $MN_{10,13}$ and the composite transistor pairs MP_{3-6} which lie along the signal path contribute to thermal noise. The tail current sources MN_{11-12} and its external reference device does not contribute to noise owing to common-mode cancelation. The OTA input-referred thermal noise voltage, power spectral density (PSD) is given by

$$\overline{v^2_{OTA,th}} = \frac{8KT}{G_M^2} \sum_i (n\Gamma)_i g_{mi} \tag{3.10}$$

where i=1,2,3,5,7,10. Here kT is the product of Boltzmann's constant and absolute temperature. $(n\Gamma)_1$, $(n\Gamma)_3$, $(n\Gamma)_5$, $(n\Gamma)_7$, and $(n\Gamma)_{10}$ are the products of the substrate factor, n, and thermal noise factor Γ for respective devices where Γ is expressed as

$$\Gamma = \frac{\frac{1}{2} + \frac{2}{3} IC}{1 + IC} \tag{3.11}$$

Γ is expressed in terms of inversion coefficient factor IC of the EKV model [53], given by

$$IC = \frac{I_D}{2n\mu C_{ox} {U_T}^2 \left(W/L\right)} = \frac{I_D}{I_S \left(W/L\right)} \tag{3.12}$$

with symbols having their usual meaning. To obtain the noise level the input referred thermal noise density can be integrated over the bandwidth $\Delta f = f_2 - f_1$. Equation (6) integrated in the desired band of frequency, namely f_1 to f_2 gives

$$\overline{V^2_{OTA,th}} = \int_{f_1}^{f_2} \overline{v^2_{OTA,th}} df = \frac{8KT\Delta f}{G_M^2} \sum_i (n\Gamma)_i g_{mi} \tag{3.13}$$

The input referred flicker noise voltage PSD for the OTA is given by

$$\overline{v^2_{OTA,f}} = \frac{2}{C_{ox} f G_M^2} \sum_i \frac{K_{fi} g_{mi}^2}{(WL)_i} \tag{3.14}$$

In the given equation K_{fi} is the flicker noise factor and is a process-dependent parameter, f is the operating frequency.

Integrating flicker noise density over the desired bandwidth gives

$$\overline{V^2_{OTA,f}} = \int_{f_1}^{f_2} \overline{v^2_{OTA,f}} df = \frac{2}{C_{ox} G_M^2} \ln\left(\frac{f_2}{f_1}\right) \sum_i \frac{K_{fi} g_{mi}^2}{(WL)_i} \tag{3.15}$$

The total input referred noise of the OTA consisting of both thermal and flicker noise power spectral density integrated over the bandwidth defined by $\Delta f = f_2 - f_1$ can be computed as

$$\overline{V^2_{n,OTA}} = \sum_i \frac{2g_{mi}^2}{G_M^2} \left[\frac{4KT\Delta f(n\Gamma)_i}{g_{mi}} + \frac{\ln(f_2/f_1) K_{fi}}{C_{ox}(W/L)_i}\right] \tag{3.16}$$

For low frequency applications the flicker noise component dominates the noise spectrum hence the total noise can be reduced by employing larger transconductance gain for input devices $MN_{1,2}$ and by increasing gate area $(W \times L)$. Further reduction in noise level can also be achieved by decreasing the g_m for non-input devices by using lower input bias current I_{BIAS} of OTA or optimally reducing gate area dimensions.

3.2.4 Noise Analysis of Neural Amplifier

The noise of the resulting neural amplifier configured as single-ended OTA in Figure 3.2 can be evaluated using the equivalent noise model as shown in Figure 3.4. The noise power density of the pseudo-resistor having equivalent resistance of r_a is represented by $\overline{v^2_{n,R}} = 4kTr_a$ as the thermal noise. The input referred noise PSD of the proposed OTA $\overline{v^2_{n,OTA}}$ which is given by the sum of equation (6) and equation (10), is placed as a noise source at the input of the noiseless OTA. The noise spectrum

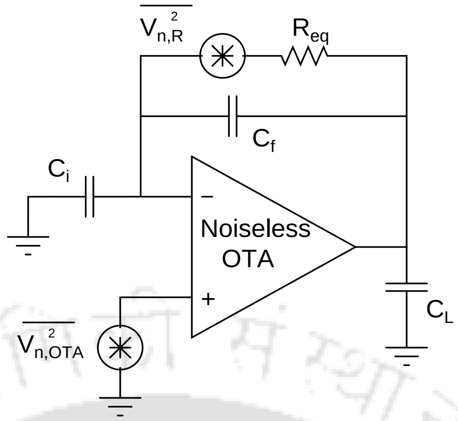


Figure 3.4: Noise model of the neural preamplifier.

of the resistor $\overline{v_{in,R}^2(f)}$ is shaped by low-pass characteristics:

$$\overline{v_{in,R}^2(f)} = \frac{\overline{v_{n,R}^2}}{{A_v}^2\left[1+(2\pi f r_a C_f)^2\right]} \approx \overline{v_{n,R}^2}\left(\frac{f_L}{A_v f}\right)^2 \tag{3.17}$$

The input noise spectral density of the OTA is calculated as

$$\overline{v_{in,OTA}^2(f)} = \frac{\overline{v_{n,OTA}^2}}{{A_v}^2}\left[\left(1+\frac{C_i}{C_f}\right)^2 + \left(\frac{1}{2\pi f r_a C_f}\right)^2\right] \tag{3.18}$$

Therefore, the total input noise spectral density of the neural amplifier is the sum of equation (3.17) and (3.18). The pseudo-resistor has thermal noise PSD $\overline{v_{in,R}^2(f)}$ with power characteristics of brown noise $(1/f^2)$ while the OTA noise is increased due to capacitive transformation. The second term in equation (15) can be ignored since spectrum of $\overline{v_{in,OTA}^2(f)}$ is only significant at much lower frequencies $(f \ll (f_L/A_v))$ than the band of interest where the flicker noise components of OTA are decades lower than that of $\overline{v_{in,OTA}^2(f)}$.

The corner frequency f_c can be obtained by equating thermal and flicker noise components as given in (3.19).

$$f_c = \frac{C_f f_L}{C_i + C_f}\sqrt{\frac{\overline{v_{n,R}^2(f_c)}}{\overline{v_{n,OTA}^2(f_c)}} - 1} \tag{3.19}$$

where $\overline{v_{n,OTA}^2(f_c)} \approx 16kT/3g_{m1,2}$ (if the amplifier is designed so that $C_i \gg C_f$ and $g_{m1,2} \gg g_{mi}, i = 3, 5, 7, 10$) has a flat spectrum at f_c. $\overline{v_{n,OTA}^2(f_c)}$ has thermal noise characteristics at f_c since the corner frequency of $\overline{v_{n,OTA}^2(f)}$ is much lower than $\overline{v_{in,OTA}^2(f)}$. It is worth mentioning that the high-pass cutoff frequency of the gain stage should be kept as low as possible.

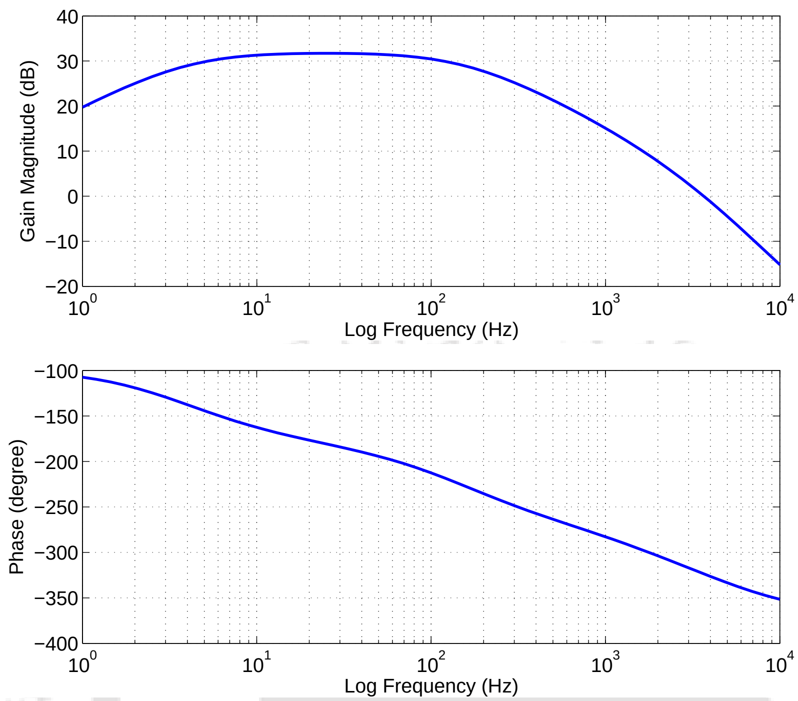


Figure 3.5: Transfer function of the amplifier configured for recording LFP.

3.3 Measurements of Local Field Potentials

The proposed symmetrical OTA was simulated in Cadence Spectre of standard UMC 0.18 μm CMOS process technology operated at $27°C$. The design assumes multi-tanh input transistor pairs $MN_{1a,2a}$ and $MN_{1b,2b}$ to be perfectly matched and working in the deep weak inversion region. Simulation results of the frequency response of the preamplifier transfer function is shown in Figure 3.5. The capacitors were sized for the mid band gain of 31.7 dB in the desired frequency band of 10 Hz to 100 Hz with a phase margin of 58.35°. The lower 3-dB frequency can be tuned via V_{tune} voltage by varying the effective channel conductance in the feedback loop. The amplifier was adjusted to have a high- and a low-pass cutoff frequency of 3 Hz and 164 Hz respectively for LFP-suitable configuration. The total current of our amplifier was measured to be 85.4 nA, corresponding to a power consumption of 68.4 nW from a 0.8 V supply. With the positive gate voltage of $V_{tune} = 0.5$ V applied to transistors $M_{a,b,c}$ the parasitic BJT behavior sets in the lateral path along source, well and drain giving high incremental resistance of the order of peta-ohms. The measured current voltage relationship is shown in Figure 3.6. This high-resistance element can be realized in a small area thus eliminating the need for large off-chip components.

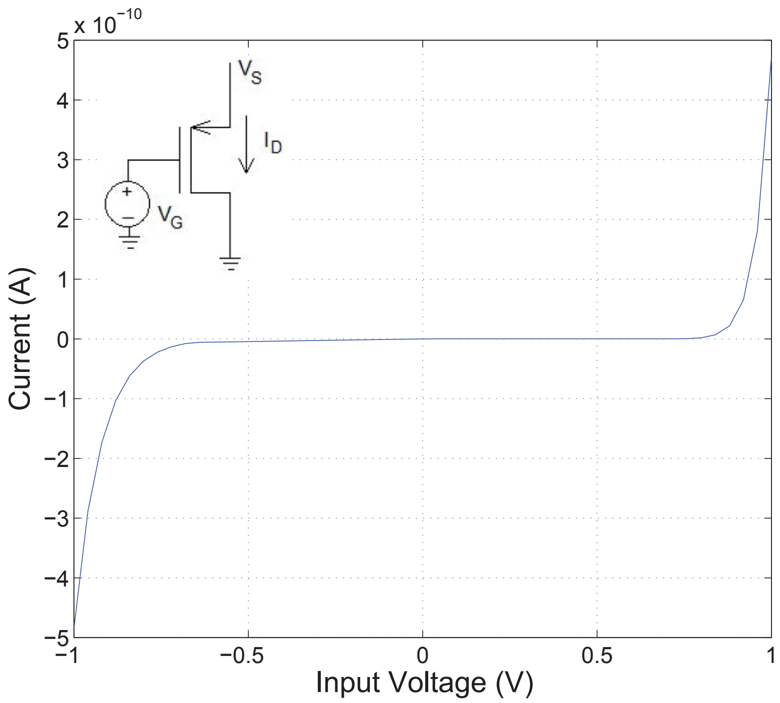


Figure 3.6: Measured $I - V$ characteristic of MOS-bipolar element.

The comparison of measured input-referred noise spectrum of neural amplifier configured for recording LFP and input referred noise of designed OTA from simulation are shown in Figure 3.7. An input-referred noise spectral density of our SCCA configured neural amplifier shows a significant amount of low-frequency noise as compared to that of G_M OTA. The total input referred noise of neural amplifier was found to be 5.62 μV_{rms}, when integrated from 1 Hz to 0.1 kHz of bandwidth. The measured noise efficiency factor (NEF) [25] for LFP recording is evaluated as 6.33. From the noise analysis of neural amplifier shown previously it was evident that the low frequency noise rolls-off as $1/f^2$ which is due to the filtering of the thermal noise in M_{a-c}. Thus in order to reduce this low-frequency noise in the pass-band, the saturation current of M_{a-c} should be made very small, so as to have as small gate-source voltage of M_{a-c} as possible. By setting the gate-source voltage of M_{a-c} such that the high-pass cutoff frequency of the amplifier happens at a very low frequency, the thermal noise of these pseudo-resistor MOS can be filtered out well before the frequency band of interest.

The simulated total harmonic distortion (THD) at 1 KHz stays below 1% for differential input less than 20.2 mV showing there by an improvement in linearity due to current cancelation and DC-shifting techiniques employed in the OTA design. This level of input signal is larger than a typical

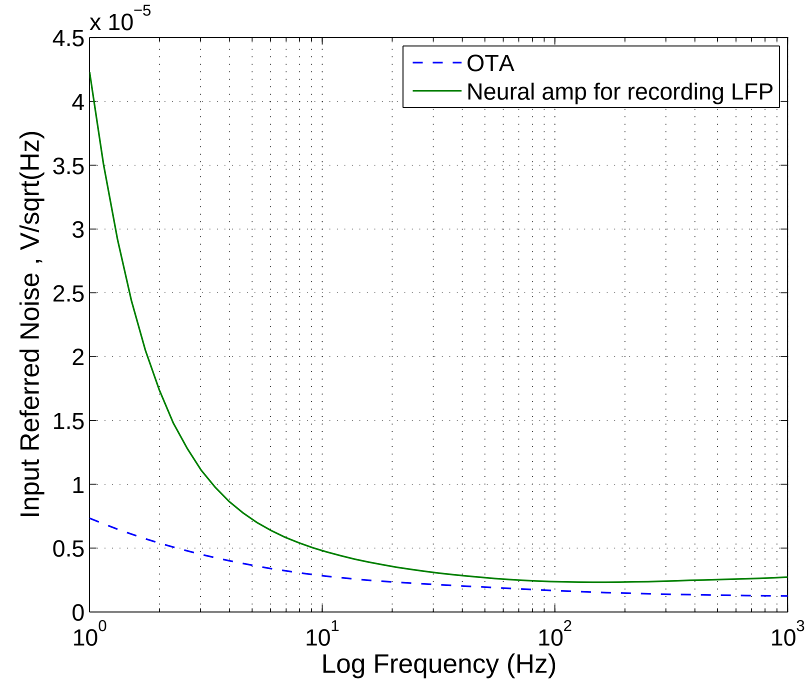


Figure 3.7: Comparison of simulated input-referred noise spectra for the amplifier configured for recording LFP (smooth curve) and G_m OTA (dashed).

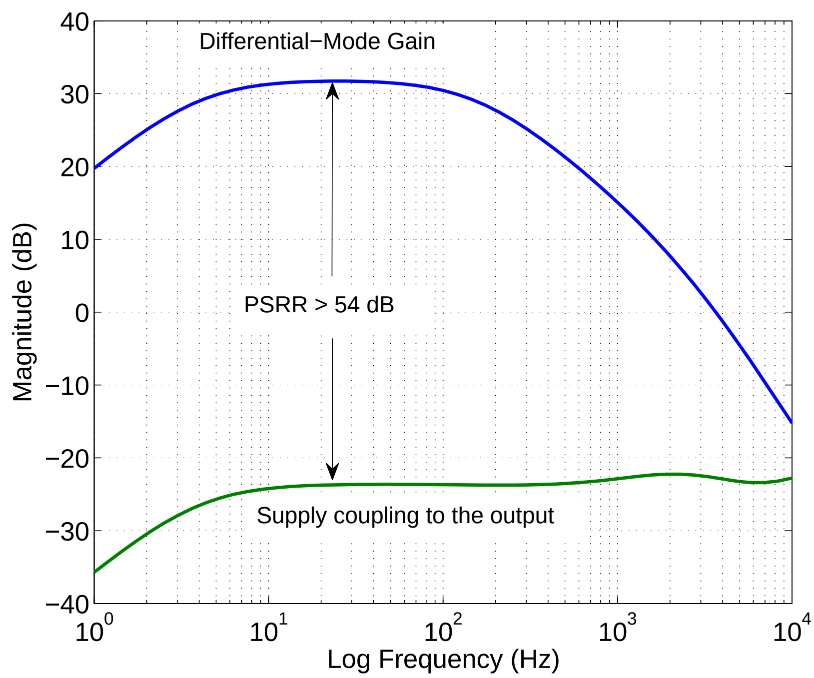


Figure 3.8: PSRR measurement of the neural amplifier configured for recording local field potentials.

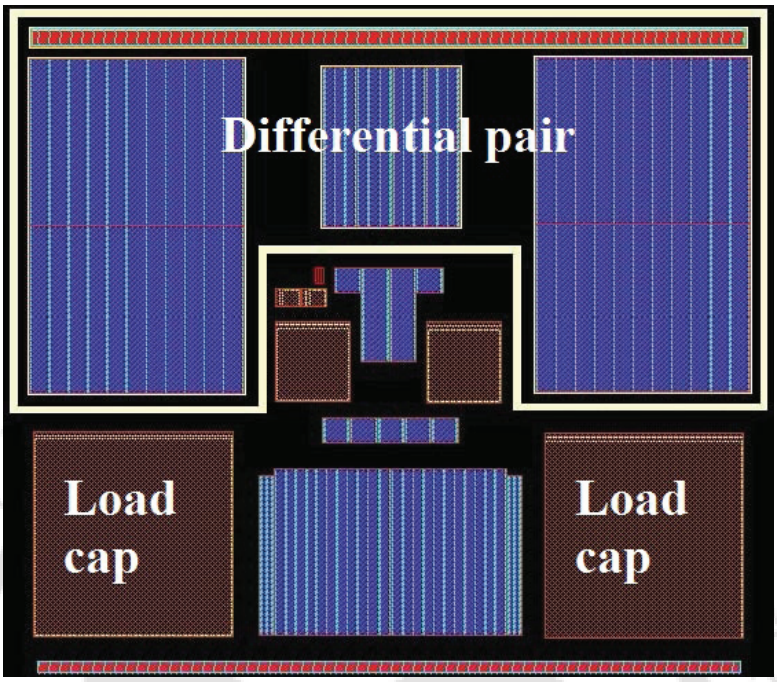


Figure 3.9: Layout of the designed neural preamplifier.

field potential amplitude level (< 10 mV) [74]. The dynamic range of our amplifier assuming 1% distortion is approximately 68 dB. The simulated result of power supply rejection ratio is shown in Figure 3.8 and is calculated as the ratio of the differential-mode gain to the gain from power supply to the output. The measured PSRR exceed 54 dB over the range of 3 Hz to 0.15 kHz which is due to low DC gain at low frequency through increase of impedance of the feedback network so that the supply ripple does not get transmitted in the frequency band of interest.

The three capacitances in integrated circuit constitute a total capacitance of 17.1 pF. The layout of the amplifier occupies an estimated area of 332 μm $\times$ 296 μm ($\sim$ 0.10 mm^2) and is depicted in Figure 3.9 (pads and routing details not shown). The input, load and feedback capacitors consume approximately 30% of the layout area. A 100-channel system made from our amplifier will consume a silicon die area of less than 10 mm^2 in a 0.18-μm process and power of only 6.8 μW allowing for a large scale implantable neural recording system.

3.4 Conclusion

An ultra-low power CMOS biosignal amplifier used in LFP recording application was successfully designed and simulated. The single-ended capacitively-coupled neural preamplifier driven with a

Table 3.1: Performance Comparison with Neural Preamplifiers Configured for LFP

Parameter	[12]	[76]	This work
Supply voltage	2.8 V	1.5 V	0.8 V
Total current	743 nA	*800 nA	85.4 nA
Gain	40.9 dB	35 dB	31.7 dB
Bandwidth	0.39-295 Hz	1-825 Hz	3-164 Hz
Input-referred noise	1.66 μV_{rms}	1.8 μV_{rms}	5.62 μV_{rms}
Noise efficiency factor	3.21	-	6.33
Max. Signal (THD)	1%@7.2 mV_{pp}	1%@8 mV_{pp}	1%@20.2 mV_{pp}
Dynamic Range	63.7 dB	66 dB	68 dB
PSRR	75 dB	60 dB	54 dB
Area	0.16 mm^2	** 2.1 mm^2	0.098 mm^2

*Includes 8 recording channels **Includes reference and biasing circuits

single 0.8 V supply offers a flat-band gain of 31.7 dB over a frequency range of $1-164$ Hz. The neural amplifier gave an input-referred noise of 5.62 μV while consuming 68.35 nW of power and occupied less than 0.10 mm^2 of silicon area. While working with the lowest supply voltage this work is a significant improvement with respect to power consumption, dynamic range, and chip area when compared to [12, 61]. A comparison between this work and a related work of [12] is shown in Table 3.1.

The layout of LFP preamplifier is designed to implement in a CMOS integrated circuit where the LFP output can be further digitized by integrating a unity gain buffer and ADC to complete the quantization process. The integrated amplifier circuit will have a minor effect with respect to the overall input referred noise, power or quantization accuracy in a practical neural front-end making these low power area-efficient CMOS circuits crucial in building high density multi-electrode array for subdural implant. Such amplifier circuits may be useful in deep-brain stimulator BMIs for studying and treating various neuro-degenerative diseases especially Parkinson's disease and power-efficient experimental neurological systems.

4

A Band-Tunable Neural Recording Amplifier with Spike/LFP Separation

Contents

4.1 Neural Recording Modalities

The spectrum of recorded bio-potential data typically include biosignals of different frequencies: Towards lower frequency range from less than 1 Hz to 300 Hz, local field potential (LFP) is dominant which is due to the summation of synaptic currents within a 1-5 mm of extracellular space. At a higher frequency from a 300 hertz to a 6 kHz, the signal consists of extracellular spikes from neurons in close vicinity of the electrode [12, 75]. Both these biopotential signals viz., local field potentials and extracellular spikes are important for information decoding in BMI systems. Moreover, these signals are often superimposed with the DC offset due to electrode-tissue interface which should also be eliminated in order to avoid saturation of front-end amplifier. Some of neural recording parameters for different signal readings at electrode are given in Table 4.1:

Table 4.1: Summary of neural recording parameters

Signal Source	Amplitude	Frequency range	Spatial Resolution
Local Field Potentials	0.5−5 mV$_{pp}$	1 Hz − 300 Hz	0.2 mm
Extra-cellular Action Potentials	0.01−0.5 mV$_{pp}$	300 Hz − 6 KHz	1 mm
EEG (surface)	0.5−100 μV$_{pp}$	0.4 Hz −30 Hz	3 cm

While action potential (AP) or spike recording through micro-electrode arrays offers higher spatial resolution, it has to compromise with the tissue fibrosis which results in signal degradation over the certain period of time [77]. LFPs have less spatial resolution than single-cell microelectrode recordings but they have the advantage of lesser susceptibility to chronic measurement issues. Recent studies have demonstrated that LFP signals can encode the information for BMI necessary for building an effective neuroprosthetic device [69, 78]. Some researches have reported using LFP recordings for decoding voluntary movement activities and in predicting sub-thalamic nucleus neural spikes during deep brain stimulation for understanding neuro-degenerative pathologies which also enhances the development of neural-to-electronic interface [79, 80]. The spectrogram of LFP recordings in the planning and execution period indicates a significant increase in the power spectrum above 15 Hz during reach task movement [68]. Choosing a particular measurement approach requires consideration of various constraints including spacial resolution, power requirement, and desired information content.

In the following sections, we shall be focussing on the major design aspects for neural front-end interface which includes AC-coupled, closed loop neural pre-amplifier followed by a band-limiting analog filter, a high-gain post-amplifier and a post-processing stage for digitization. It also discusses band programmability feature of a neural recording amplifier for sifting spike signals from low-frequency local field potentials for delivery of high-frequency spike signal to the following block which is a spike detector. Most of brain activities related to neurological diseases occur in $\alpha, \beta, \gamma, \delta$ frequency range [75,81]. Therefore, focus of this work will be to shift the 3-dB high-pass corner frequency of neural amplifier into sub-Hz range. This would filter out the DC offsets at the inut due to electrode potentials. The design constraints of the previously designed LFP amplifier are addressed and later on low power analog design techniques are discussed for the improved version of LFP amplifier. A prototype design of LFP recording amplifier implemented in 180 nm CMOS process and experimentation results are also presented which characterizes circuit implementation and validates the design method.

4.2 Motivation for spike/LFP Amplifier

Since biomedical signals require quite different amplification levels and bandwidths, programmability becomes mandatory with respect to both gain and bandwidth. Size and power becomes two major concerns for implantable neural recording systems. A recording system is desired which can acquire the entire biosignal spectrum of interest while rejecting DC offset. In the field of biomedical electronic systems, the acquisition devices for neural signal recording, typically consists of an capacitively-coupled instrumentation amplifier (CCIA). CCIA is a key element, which senses and amplifies the neural signals such as Electrocorticography (ECoG), Electroencephalogram (EEG), action potential, local field potential (LFP) etc., through electrode-tissue interface. The neural signals acquired by microelectrodes are very small in amplitude (1 μV to 0.5 mV range, depending on the distance from the electrode and on the size of cell) and are accompanied with local field potential, neuronal background noise and other electrical disturbances from electrode-electrolyte interfaces. Therefore, the design of CCIA should necessarily include filtering circuitry along low noise signal amplification whose input-referred noise should be below 10 μVrms. It is important to be mentioned that the acquisition process should preserve the information contained in the original signal. Thus the front-end system should not introduce any form of noise that can distort the information. The power consumption of the neural interface circuit should also be kept low (<1-2 μW/channel) since the CCIA has to be integrated with microelectrodes which has short thermal path to the neural tissues. Simultaneously, small area per recording channel is also important so it can be scalable to a high channel count system for its suitability in high performance BMI systems.

The previous design of low-noise amplifier for recording local field potential can be employed for amplifying neural signals including neural spikes in 100 Hz to 6 KHz range. For this, the functionality of the previously designed LFP amplifier needs to be extended to record these signals. However, in order to develop this architecture into a fully implantable system for detecting neural potentials, the design of this amplifier has to meet certain challenges:

(i) Operation with high supply voltage is preferred since a low supply reduces overdrive voltage in some above-threshold transistor which increases noise contribution.

(ii) For real-time recording of bio-potentials, there can be a shift in the amplitude scale at the input, therefore, calibration of equipment is important for its accurate determination owing to which the front-end amplifier should have an adjustable gain feature.

(iii) It is required to increase the upper cut-off frequency since the LFP amplifier is designed to operate up to 170 Hz which can only record LFP signals. Whereas the bio-potential signals at other frequency bands (neural spikes) will be attenuated.

(iv) In the previous design, the lower cut-off frequency of LFP amplifier is high ($>$ 3 Hz) because of which a significant part of low frequency neural activities which lie near 1 Hz frequency range is lost. The lower cut-off frequency should be extended to below 1 Hz since most of neural diseases occur at frequencies in $<$1-30 Hz band.

(v) Finally, in many applications it is desirable to separate LFP from neural spikes. In our application, the amplifier should also have a tunable bandwidth while interfacing with spike detection circuit in order to eliminate local field potentials which can interfere with the subsequent spike detection circuit.

4.3 Front-end Amplifier Design

The front-end of neural interface forms a crucial part of spike detection. The microvolt level of biopotential signal ($\sim$ 10 μV) at the electrodes and large frequency spectrum from below 1 Hz up to several kilo hertz present in the neural signal mandates the design of an analog front-end interface. A gain of the order of 1000 and a reconfigurable bandwidth is required in order to make neural signal

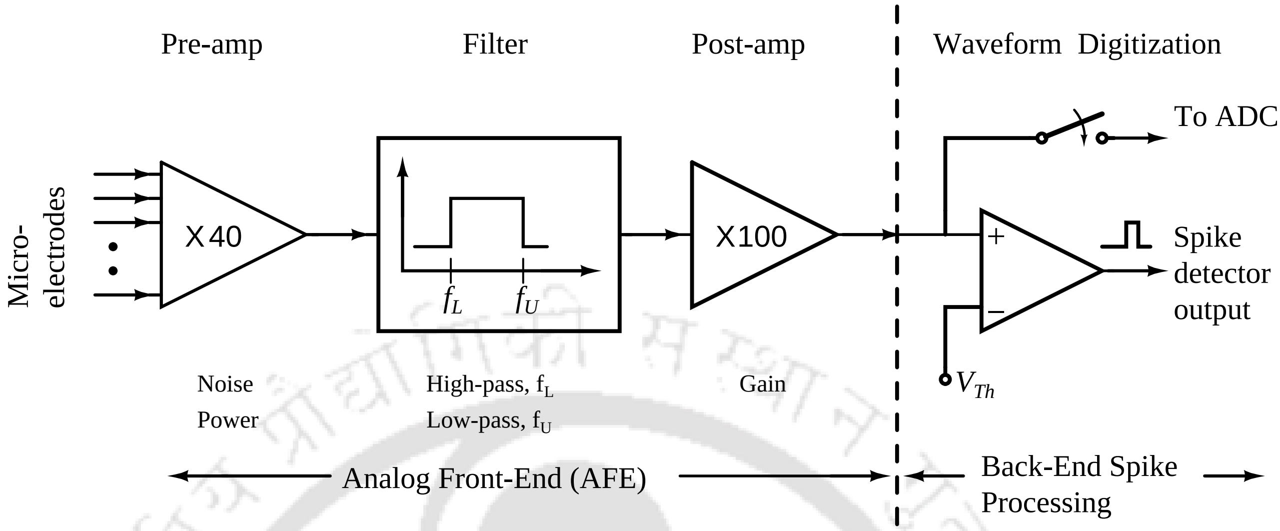


Figure 4.1: The channel architecture of proposed neural signal processing chain: a cascade of pre-amplifier, filter, post-amplifier, and back-end processing blocks for data reduction and waveform digitization.

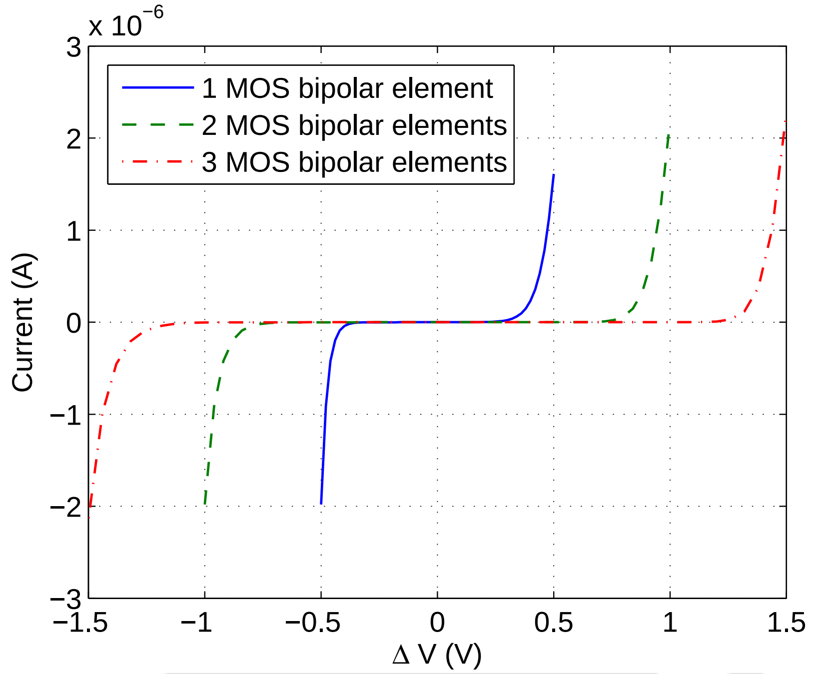
suitable for the spike detection. Due to poor noise performance of CMOS circuits, the pre-amplification stage should also have a low input-referred noise.

In most of neural recording front-end systems, a high-gain instrumentation amplifier (IA) is used after the micro-electrode. This stage is followed by either an ADC and a bio-marker extraction circuitry or a simply a spike detector for neural signal digitization. The channel architecture neural processing signal chain is shown in Figure 4.1. The raw signal picked-up from micro-electrodes is processed at the first stage of analog front-end (AFE). AFE mainly consists of three stages: (1) a low-noise gain stage, (2) bandwidth limiting stage, and (3) programmable gain stage (Figure 4.4). Later on the amplified and filtered signal is processed at the back-end for spike detection and digitization. The tuning parameters are also outlined along with the corresponding block. The pre-amplifier stage mainly determines the noise-power trade-off. This stage can be designed so that an input-referred noise below 10 μV_{rms} can be traded with a sub-micro Watt power while providing a moderate gain of 30 to 35 dB. Gain should not be very high for this stage as the frontend will saturate for large signal input. The subsequent stages of amplifier are designed for high gain and high swing. On the other hand, if the gain is very low it may raise the noise requirement for the following stages. A low-power $G_m - C$ filter stage can be tuned for upper and lower cut-off frequencies with a negligible addition to the input-referred noise. The third stage is designed for providing an additional gain of over 40 dB from DC to 6 KHz so that the neural signal can be further processed by a threshold comparator for spike detection. The voltage threshold V_{Th} is usually set by a spike detection algorithm for reducing detection errors. Important information contained in the amplified neural signal can also be preserved for A/D conversion and waveform digitization.

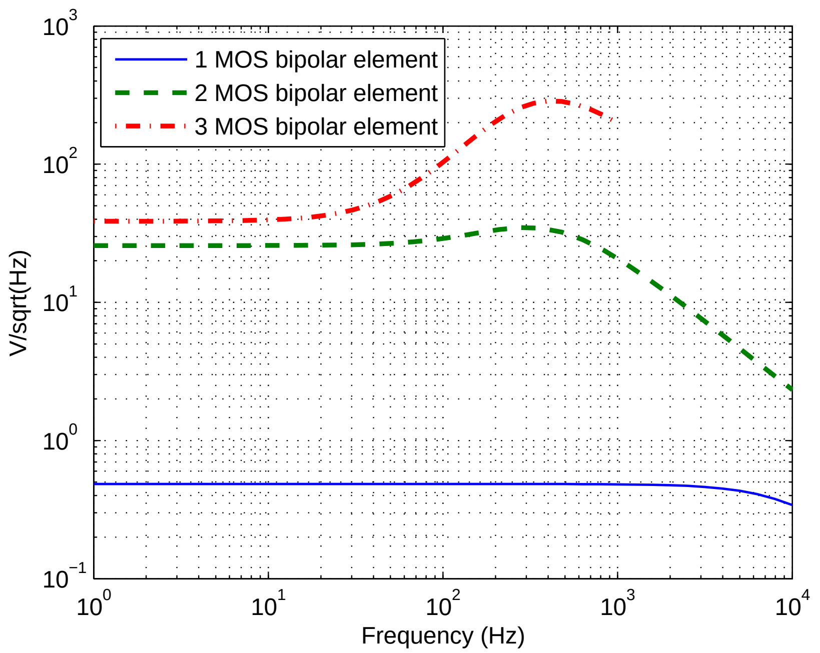
4.3.1 Design of the Pre-amplifier Stage

For pre-amplifier stage, a generic topology of single-ended OTA-based neural signal amplifier with capacitive feedback [14, 82, 83] is adopted. The scheme used for our pre-amplifier is analogous to the one used in earlier chapter. For achieving lower cut-off frequency a very high value of feedback resistance is required (of the order of Giga-Ohm). Previous design of LFP amplifier consists of three series transistors in feedback circuit. Equivalent resistors of such high values can be implemented by means of MOS devices however the corresponding range of linearity is limited. A series combination of identical pseudo resistors can be replaced for a better linearity [84]. However, as the number of MOS elements increases the overall noise contribution also increase. The increase in total noise power is due to noise contribution of the individual pseudo resistors which consists of MOSFET thermal noise and Flicker noise. This trade-off between dynamic range and noise suggests there is an optimum choice for the number of transistors used in neural amplification. The figure showing the effect of the choice of the number of MOS elements on linear range of current-voltage relationship of MOS bipolar element and total input referred noise in 4.2(a) and 4.2(b).

From these figures it is evident that increasing number of series transistors improve the linear performance over a larger range of voltage swing applied across them, but it also contributes to the thermal noise of the amplifier. Therefore, only two transistors M_a and M_b are considered (instead of three transistors) in the feedback path. A modified schematic of Figure 3.1 is reproduced in Figure 4.3(a) for AC analysis. These pMOS transistors work as MOS-bipolar devices collectively termed as pseudo-resistor. With application of a very small V_{GS}, the transistors operate as a parasitic source-well-drain p-n-p bipolar junction transistor (BJT) equivalent to a diode-connected BJT. r_a denotes the incremental resistance of $M_{a,b}$ when there gate-source voltage is close to zero. The loading effect at the output node of neural amplifier is modeled as C_{Lp}, a parasitic capacitance connected between v_{na} and incremental ground. To analyze the operation of neural amplifier, we adhere to a feedback amplifier approach. Figure 4.3(b) shows a preliminary block diagram which describes the operation of feedback amplifier. The feedforward gain G is the open-loop gain $-A(s)$ of amplifier, where $-A(s) = G_{m1} \times (r_o \parallel (1/sC_{Lp}))$. The transfer function K and H denote the input voltage

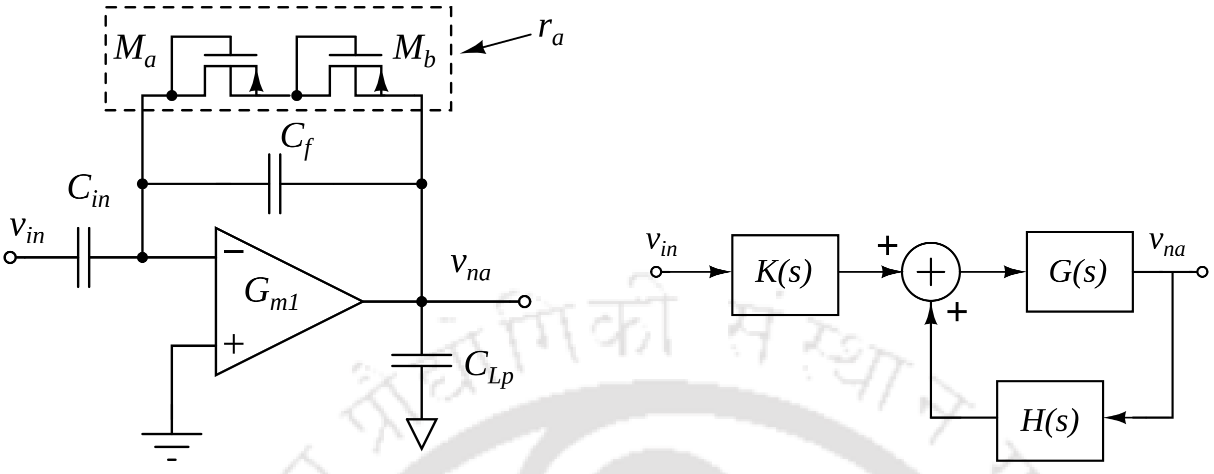


(a)



(b)

Figure 4.2: Comparison of linearity and noise of single, double, and triple pseudo resistor element (a) I-V characteristics (b) input-referred noise spectrum.



(a) AC analysis of pre-amplifier stage.

(b) A feedback block diagram representation.

Figure 4.3: Circuit schematic for small signal analysis of neural amplifier

divider and feedback voltage divider gain given by $Z_f/(Z_f + Z_{in})$ and $Z_{in}/(Z_f + Z_{in})$ respectively. K and H can be derived by substituting the input impedance $Z_{in} = 1/sC_{in}$ and feedback impedance $Z_f = r_a/(1 + sr_aC_f)$ as

$$\begin{aligned} K(s) &= \frac{sr_aC_{in}}{1 + sr_a\left(C_{in} + C_f\right)} \\ H(s) &= \frac{1 + sr_aC_f}{1 + sr_a\left(C_{in} + C_f\right)} \end{aligned} \tag{4.1}$$

Towards the operation of amplifier where the frequency $\omega \gg 1/(r_a \times (C_{in} + C_f))$, H can be approximated as $C_f/(C_{in} + C_f)$. Neglecting channel length modulation (output resistance $r_o \gg 1/sC_{Lp}$), the open loop gain $-A(s) \cong G_{m1}/sC_{Lp}$. The voltage gain of the neural amplifier can be obtained as

$$\begin{aligned} H_{na}(s) = \frac{v_{na}}{v_{in}} &= K.\frac{-G}{1 + GH} = -\frac{K}{H}.\frac{1}{1 + (1/GH)} \\ &= \frac{sr_aC_{in}}{1 + sr_aC_f}.\frac{1}{1 + (sC_{Lp}/G_{m1})[(C_{in} + C_f)/C_f]} \end{aligned} \tag{4.2}$$

The expression in (4.2) contains a high-pass and low-pass transfer function. The expression $(C_{in} + C_f)/C_f$ can be defined as a midband gain A_{m1} of the neural amplifier. This equation suggests that the lower cut-off frequency is governed by the combination of feedback capacitor and pseudo resistor and is situated at $f_{na,L} = 1/2\pi r_aC_f$. The higher cut-off frequency which is due to the parasitic load capacitor is given by $f_{na,U} = G_{m1}.C_f/2\pi(C_{in} + C_f).C_{Lp}$. At a mid-band frequency where

$1/r_aC_f < \omega < G_{m1}.C_f/(C_{in}+C_f).C_Lp$, the gain of the neural amplifier can be approximated as

$$H_{na}(s) = \frac{C_{in}}{C_f} \tag{4.3}$$

A high-level schematic of a three-stage neural front-end system along with corresponding OTA topologies is described in Figure 4.4. The first stage of analog processing chain consists of a pre-amplifier which mainly determines the noise performance of a neural recording channel. This stage is designed with a low input-referred noise and is implemented through a differential-input capacitively-coupled instrumentation amplifier (CCIA) first used in [62].

The amplifier is AC coupled to the micro-electrodes through C_{in} to decouple front-end DC offset arising from electrode-tissue interface. The use of on-chip capacitor creates the high-pass corner frequency and eliminates the need for external components. The pseudo-resistive elements $M_{a,b}$ along with the capacitance C_f in feedback path creates a high-pass corner frequency given by $\omega_L = 1/(r_a \times C_f)$. CCIA also acts as a reconfigurable high-pass filter. The high-pass corner frequency can be tuned with the bias voltage V_{bias} at the gate terminal of pseudo-resistor. The low-pass corner frequency is determined by the 3 dB roll-off of the OTA of CCIA. The ratio of feedback capacitance, $C_f = 140$ fF and the input capacitance $C_{in} = 10$ pF sets the mid-band gain of the preamplifier to approximately 34 dB. The value of C_{in} is chosen sufficiently high in order to avoid tissue fibrosis leading to signal distortion and to reduce the silicon area. The lower input referred noise may be obtained at the cost of higher drain current but severe power constraint on implantable devices imposes the direct limit on increasing the current bias.

The expanded view of Figure 4.4(a) depicts the basic schematic of G_{m1} operational transconductor amplifier (OTA) of first stage preamplifier design. At low frequency, the CMOS transistors exhibit a $1/f$ power spectrum as noted in Figure 3.7, therefore the key challenge is to reduce the $1/f$ noise low. Cascode topologies are impractical because of the low supply voltage, hence two NMOS differential pairs with PMOS loads in parallel are employed. Hence the same OTA configuration used for LFP amplifier can be utilized for gain and band programmability and can be applied generally to neural sensor interfaces. The transconductor employs a multi-tan$h$ scheme implemented by differential pairs $M_{1a,b}$ and $M_{2a,b}$ having equal tail currents. The unequal sizing of input transistors is chosen in order to maximize the linear operation of the transconductor. The series transistor structure formed by common gate configuration of M_{3-6} ensure constant and equal drain voltages of differential pairs. A

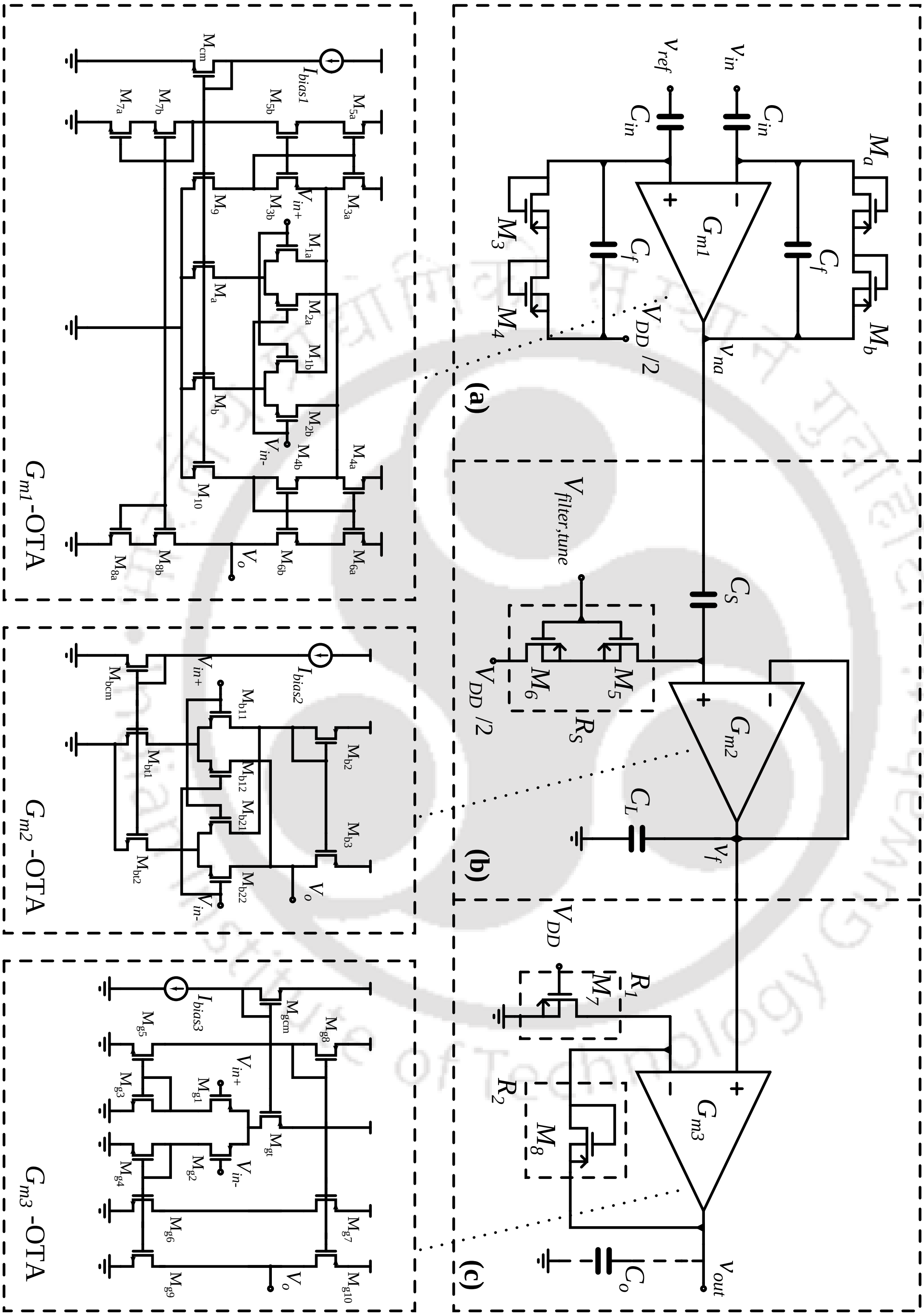


Figure 4.4: System block diagram of the three-stage analog front-end interface and OTA topologies for (a) 1st stage: capacitor-coupled instrumentation amplifier (A_{v1} =37 dB), (b) 2nd stage: tunable bandwidth limiting stage, ($G_{m2}\times 1/C_L$ =1) and (c) 3rd stage: high-gain amplifier (A_{v3} =43 dB).

Table 4.2: Operating Points For Transistors In The G_{m1}−OTA with $I_{bias1} = 0.2\ \mu$A

Devices	W/L (μm)	$I_D(nA)$	g_m/I_D (V^{-1})	$V_{GS} - V_T$ (mV)
$M_{1a,2a}$	100/8	36.51	32.57	-165.8
$M_{1b,2b}$	400/8	152.6	32.5	-164.4
$M_{3a,4a,5a,6a}$	100/4	378.6	22.88	66.1
$M_{3b,4b,5b,6b}$	400/8	191.8	25.61	125.8
$M_{7a,8a}$	140/2	329.3	31.01	-146
$M_{7b,8b}$	400/4	329.3	32.45	-162.9
$M_{9,10,a,b,cm}$	10/8	190	23.71	-24.5

single-ended output current of OTA is drawn from self-cascode structure $M_{8a,b}$ which increases the DC gain without any substantial loss of voltage swing. The transistors are sized so as to increase g_m/I_D ratio (by operating in weak-inversion region) of input differential pairs while reducing g_m/I_D ratio of non-input devices by driving them into a higher moderate inversion regime. The input referred noise power spectral density including flicker and thermal noise of the transconductor in Figure 4.4(a) is expressed as

$$\begin{aligned} \overline{v_{n,th}^2} &= 2\frac{4kT}{G_{m1}^2}\sum_i \gamma_i g_{mi} \\ \overline{v_{n,f}^2} &= 2\frac{1}{G_{m1}^2 C_{ox} f}\sum_i \frac{K_{fi} g_{mi}^2}{(W_i L_i)^2} \end{aligned} \tag{4.4}$$

where i =1,2,3,5,7,10. The value of γ is 2/3 for devices operating in strong inversion and $1/2\kappa$ in weak-inversion. W, L, C_{ox}, f_i are gate width length, oxide capacitance density, and process-dependent flicker noise factor respectively. The output transconductance G_{m1} of OTA is given by

$$G_{m1} = \frac{g_{m2}}{1 + \frac{g_{o2}}{g_{m6b}+g_{o6b}}} \cong g_{m2} \tag{4.5}$$

The total noise of first stage OTA is contribution of both flicker and thermal noise.

$$\overline{v_{n,OTA1}^2} = \overline{v_{n,th}^2} + \overline{v_{n,f}^2} \tag{4.6}$$

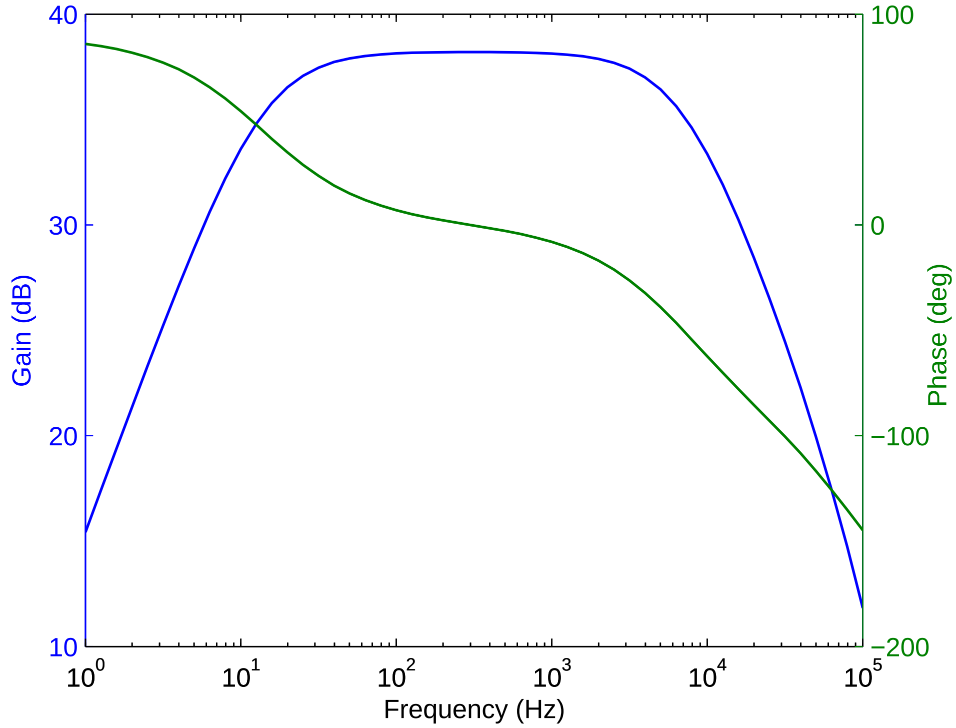


Figure 4.5: The transfer function of pre-amplifier. Midband gain is 38.2 dB, and the high frequency roll-off occurs at 6.9 kHz. Low-frequency roll-off occurs at 13.9 Hz.

If r_a is replaced with a real resistor then its noise spectral density is given as

$$\overline{v_{n,r}^2} = 4kTr_a \tag{4.7}$$

The total input referred noise is because of the contribution of OTA and the resistor

$$\overline{v_{in}^2} = \overline{v_{in,r}^2} + \overline{v_{in,OTA1}^2} \tag{4.8}$$

where $v_{in,r}^2$ and $v_{in,OTA1}^2$ are similar to equations (3.17) and (3.18) respectively. Table 4.2 shows the dimensions of devices. The input differential pair transistors $M_{1,2}$ are sized with a large W/L ratio so as to push them into deep weak inversion region, maximizing their g_m/I_D ratio. The transistors in biasing circuit $M_{3,4,5,6,7,8}$ are sized with small W/L ratio to operate them in moderate inversion region instead of strong inversion to save power. This also contributes to lowest thermal noise and an improvement over the LFP amplifier noise. This improvement in noise is evident from Figure 4.9(b). The bias current is set to 0.2 μA, giving devices $M_{1a,2a,1b,2b}$ small drain currents of 36 and 152 μA. Total power consumed by the pre-amplifier stage including current source is 1.9 μW.

4.3.2 Design of the Filter and Gain Stage

To circumvent noise-power trade-off and to make the instrumentation amplifier more robust to noisy environment, another stage of $g_m - C$ based bandpass filter is added as shown in Figure 4.4 (b). This stage is added to independently shape the pass-band of the preamplifier without effecting input referred noise. With the addition of the filter stage, the neural amplifier can be configured to either amplify local field potential signal (4 Hz − 300 Hz) or record the neural spikes (300 Hz − 6 KHz) to facilitate spike detection. The transfer function of the filter stage is given as

$$H_f(s) = \frac{v_f}{v_{na}} = \frac{sC_sR_s}{1+sC_sR_s}.\frac{1}{1+sC_L/G_{m2}} \tag{4.9}$$

For higher operating frequency in which $1/r_aC_f < \omega < 1/C_sR_s$ the transfer function of the neural amplifier and filter stage can be expressed as

$$H_{na,f}(s) = \frac{v_f}{v_{in}} = -\frac{C_{in}}{C_f}.\frac{sC_sR_s}{1+sC_sR_s}.\frac{1}{1+sC_L/G_{m2}} \tag{4.10}$$

Since action potential signal contains most of energy in 300 Hz−6 KHz band therefore preamplifier should be designed to reject signal outside this band. This can be carried out by changing the value of time constant C_sR_s, where R_s is the equivalent resistance of a floating high-valued resistance composed of two back-to-back pMOS transistors connected in series. The value of R_s can be controlled through the gate voltage $V_{filter,tune}$ with respect to the drain current which provides a wide tuning range. The high-pass corner frequency is set by $f_L = 1/2\pi C_sR_s$. The upper cut-off frequency is given by $f_U = G_{m2}/2\pi C_L$.

Figure 4.4 (c) shows the third stage of AFE which is a high-gain amplifier loaded by a large input capacitance C_o of the following stage (viz. an ADC). The gain of this stage is equal to $(1 + R_2/R_1)$. R_2 is a large fixed pseudo-resistor defined by a diode connected M_8 that sets the input common-mode voltage of the OTA. For the third stage, a current-mirror OTA is used which is suitable for driving capacitive loads. In order to maximize the the closed-loop gain of the amplifier, a high-transconductance OTA should be used. For this reason the input differential transistors are biased in strong-inversion side of moderate inversion which also helps to achieve low noise at lower bias current. It also serves setting the high-cutoff frequency larger than the input neural signal bandwidth.

A band-limiting stage allows the preamplifier to adjust the 3-dB bandwidth without compromising

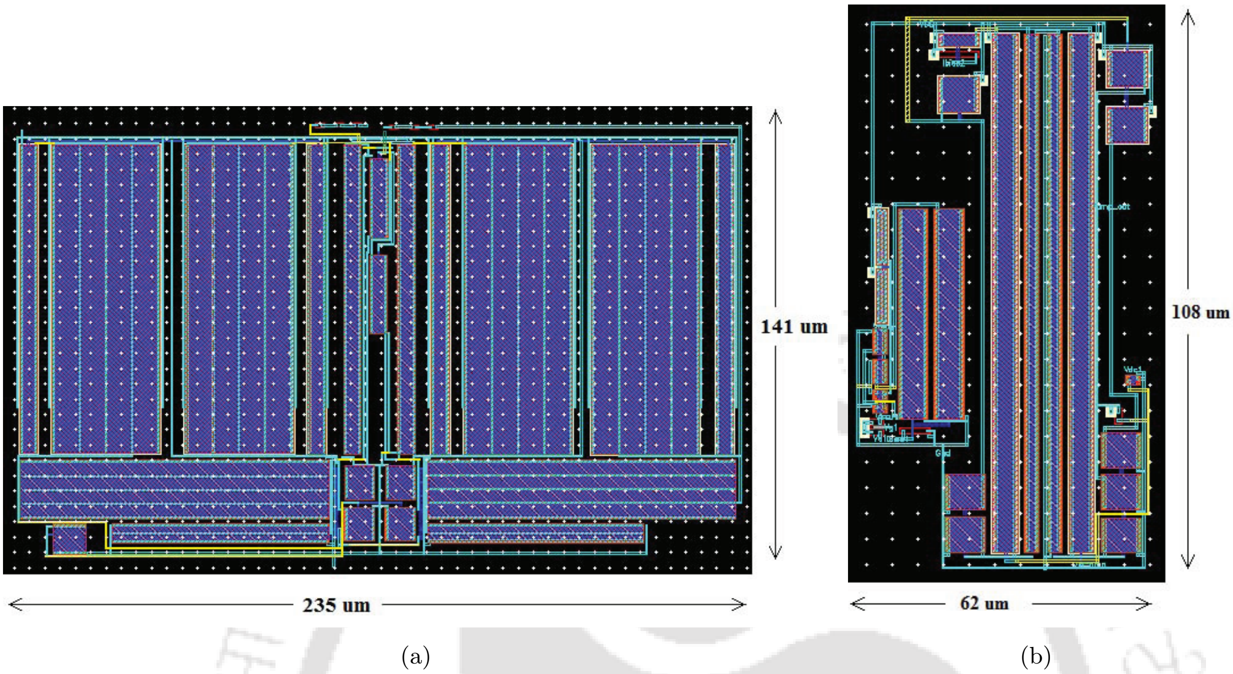


Figure 4.6: Layout area of 3-stage programmable front-end amplifier consisting of (a) the instrumentation amplifier and (b) band limiting filter and high gain stage.

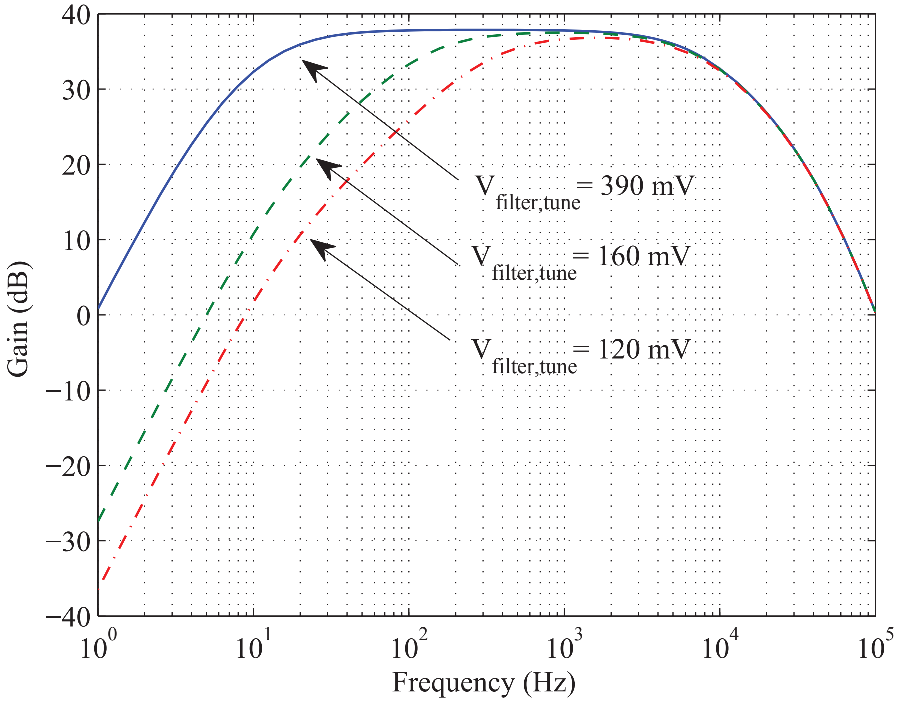


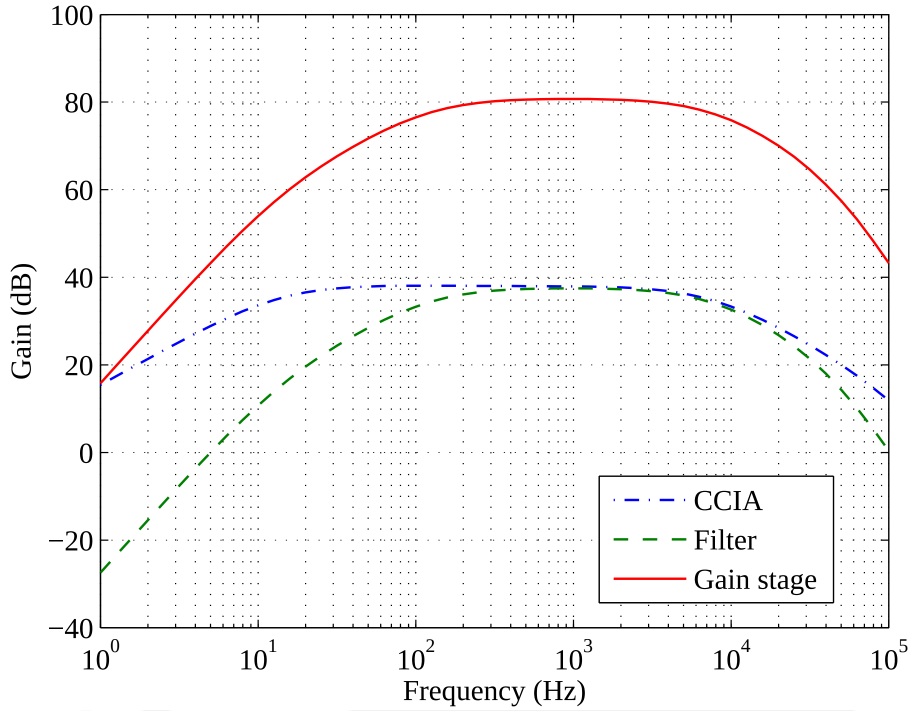
Figure 4.7: The transfer function of neural amplifier after filtering stage. Programmable high-pass filter pole shown at three different cut-off frequencies, f_{HP} = 15 Hz, 124 Hz, and 318 Hz at tuning voltage of 390 mV, 160 mV, and 120 mV respectively.

Table 4.3: Operating Points For Transistors In The G_{m2} and G_{m2} OTA with I_{bias2} = 5 pA and I_{bias3} = 0.1 μA

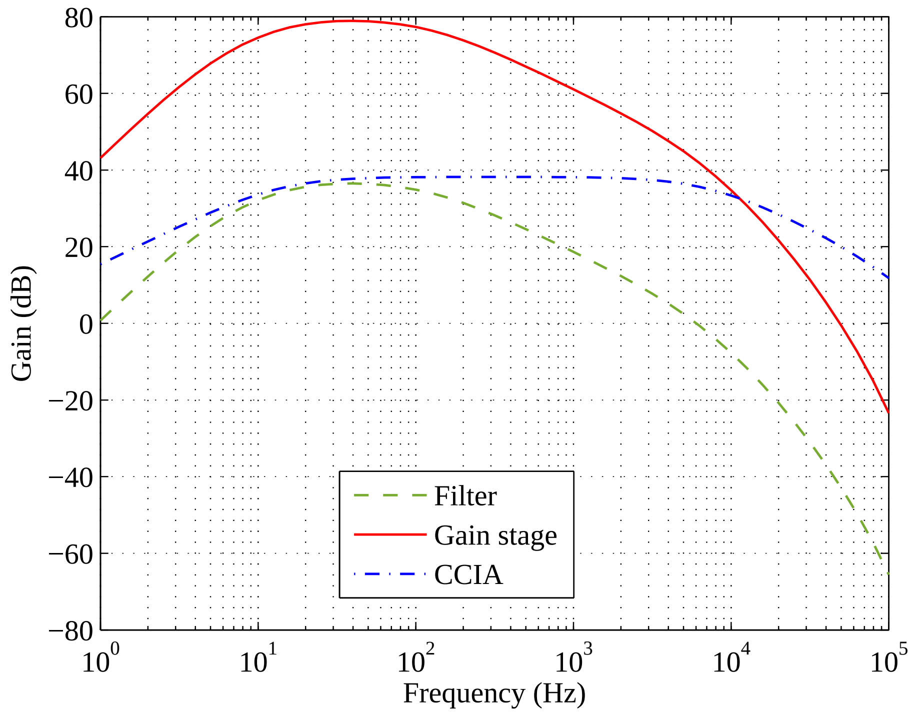
Devices	W/L (μm)	I_D	g_m/I_D (V^{-1})	$V_{GS} - V_T$ (mV)
$M_{b11,b21}$	1/1	15.57 pA	36.35	-324.23
$M_{b12,b22}$	4/1	361 pA	35.85	-273.7
$M_{b2,b3}$	80/0.2	245.2 pA	25.79	493.58
$M_{bt1,bt2}$	50/4	376.9 pA	36.23	-297.85
M_{bcm}	0.24/4	4.97 pA	35.72	-262.15
$M_{g1,g2}$	100/3	257.6 nA	24.19	95.8
$M_{g3,g4,g5,g6,g9}$	6/6	257.6 nA	22.17	-2.0
$M_{g7,g8,g10}$	1/6	247.8 nA	7.8	-222.9
M_{gcm}	0.24/6	100 nA	6.57	-275.7

on the low input-referred noise of first stage while the noise contribution from bandpass filter remain insignificant. The transconductor employed in the design of OTA is a parallel asymmetric differential pair as shown in the inset of Figure 4.4(b). The high- and low-pass corner frequency can be adjusted by varying $V_{filter,tune}$ and transconductance of OTA. The third stage of amplification includes linear gain stage. The OTA is implemented with a current-mirror transconductor topology whose gain can be programmed through bias current I_{bias2} from 27 dB to 33 dB. Post layout simulations were carried out separately for the neural amplifier at the first stage and band-limiting and gain amplifier at the second stage. The layout area occupied by first stage is 235 $\times$ 141 μm while the layout area of second stage is 108 $\times$ 62 μm when designed in Cadence UMC PDK as shown in Figure 4.6(a) and 4.6(b) respectively.

A mid-band gain of 80 dB, an input-referred noise of 5 μV_{rms} (10 Hz $-$ 6 kHz), a programmable high-frequency cutoff from 100 Hz to above 6 kHz, and a tunable low-frequency cutoff from 300 Hz down to 10 Hz are some of the important functional specifications measured for this circuit. The high-pass corner frequency f_L can be controlled with the time constant $\tau_L = R_s \times C_s$. Figure 4.7 shows programmability of high-pass corner frequency at three different voltage settings of $V_{filter,tune}$. At control voltage of $V_{filter,tune}$ = 120 mV, 160 mV, and 390 mV, the cut-off frequencies are 318 Hz,

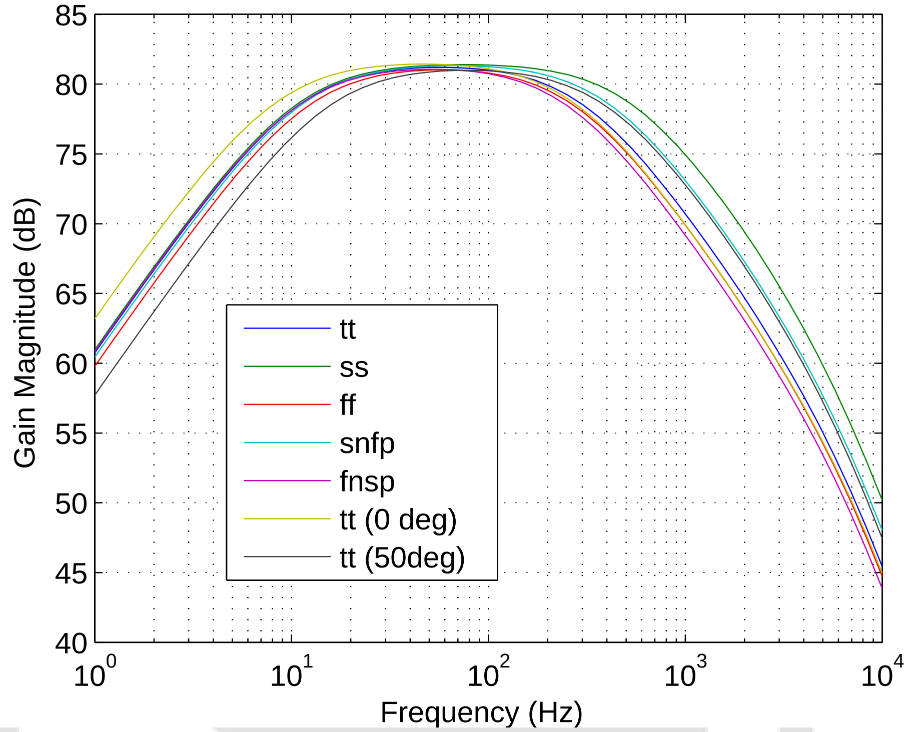


(a)

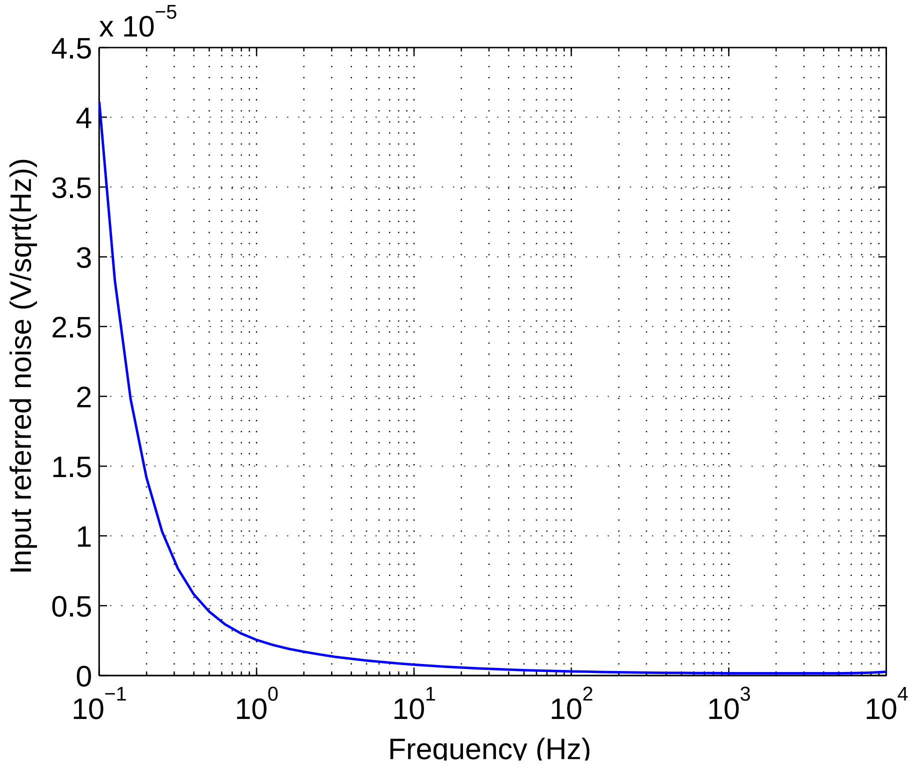


(b)

Figure 4.8: Magnitude response of different stages of FEA namely CCIA, band-limiting filter stage, and high gain stage. FEA can be configured to record either (a) spike in the frequency band 100 Hz to 6 KHz or (b) local field potential in the range 14 Hz to 100 Hz.

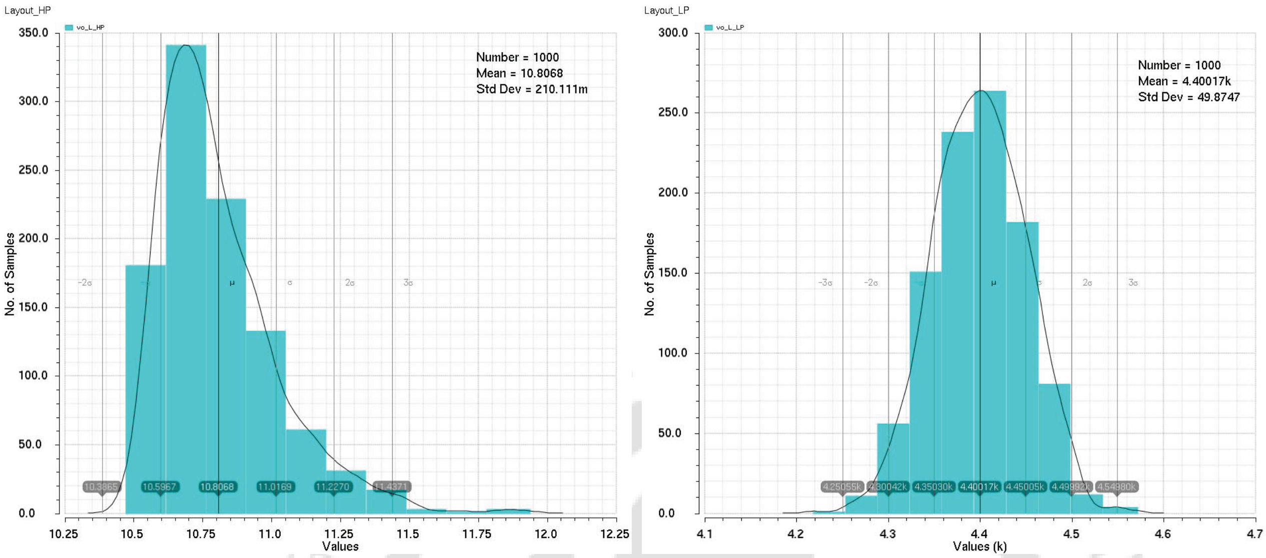


(a)



(b)

Figure 4.9: Corners and noise analysis of AFE (a) Frequency response at different process corners (b) Input referred noise.



(a) (b)

Figure 4.10: Monte Carlo analysis of pre-amplifier with a 1 pF load capacitor when the bandwidth was tuned for (a) high pass corner. (b) low pass corner.

124 Hz, and 15 Hz respectively. Mid-band gain and high cut-off is constant at 36.8 dB and 7.8 KHz respectively. Low-pass corner frequency can be controlled with the time constant $\tau_U = C_L/G_{m2}$. The load capacitor C_L can be varied through which the low-pass corner frequency f_U can be programmed. The transconductance G_{m2} can be also varied but since thermal noise is inversely proportional to the transconductance, therefore value of transconductance of second stage cannot be lowered beyond certain value. For this reason, bias current at the first stage should be kept fixed and should not be adjusted for decreasing input referred noise. There is also a reason why first stage can not be chosen for setting of f_U. The combination of pseudo-resistor r_a and capacitor C_f determine the lower cut-off frequency in the first stage. The thermal noise is contributed by a large value of pseudo-resistor r_a. Noise of the first stage should be the lowest since it will be amplified in second and third stage of AFE. Consequently to reduce the noise contribution due to this pseudo-resistor, the product of r_a and C_f should be less then the product of r_s and C_s so that the pole due to second stage becomes dominant for setting the low-cutoff frequency of AFE amplifier.

The composite transfer curve is illustrated in Figure 4.8. The transfer characteristics at different stages of FEA are depicted in Figure 4.8(a) and 4.8(b) for two different configurations: spike and LFP recording. It is shown that by tuning $V_{filter,tune}$ voltage, the FEA can be switched between two modes for recording either LFP or neural spikes. The first stage gain is set to 36.8 dB with a band of 6 KHz.

At the second stage, the low-pass filter pole can be tuned for recording high frequency spike by setting the load capacitor to 1 pF. For the load capacitor of 10 pF, the bandwidth can be reduced to 100 Hz which is suitable for recording of LFP. The third stage adds the gain of 43 dB without modifying the bandwidth resulting in the overall gain of over 80 dB.

Table 4.3 shows the dimensions and operating points of devices in G_{m2} and G_{m3} OTA. The input transistors $M_{b11,b12,b21,b22}$ in the filter stage are biased in deep weak inversion region. The transistors in biasing circuit $M_{3,4,5,6}$ are operated in moderate inversion region. Total power consumed is 1.9 μW. The pre-amplifier presented in this work was designed and simulated in 0.18 μm CMOS process. AC analysis of AFE output in typical process condition and in all the four process corners as well as at different temperature is shown in Figure 4.9(a). Figure 4.9(b) shows the input-referred voltage noise spectrum. The noise analysis shows the flicker noise of 38 $\mu V_{rms}/Hz^{1/2}$ at 100 mHz while noise is below 500 nV at 1 Hz. Since the CCIA stage is designed for a fixed bandwidth therefore Monte-Carlo simulations with post-layout extracted parasitics were carried out to verify the sensitivity of the first stage amplifier circuit to the device mismatches. Figure 4.10 shows a one thousand runs of Monte Carlo simulation of the high pass and low pass cut-off frequency at room temperature of 27 °C. It can be seen that the histogram in Figure 4.10(a) exhibits a narrower spread of frequency with a standard deviation (σ) of 0.21 Hz around a mean value of 10.8 Hz. The histogram in Figure 4.10(b) also has a standard deviation of 49.9 Hz for a mean of 4.4 KHz. Both the analysis for the low and high cut-off frequency shows the effectiveness of the proposed configuration for setting the bandwidth of the CCIA stage.

4.3.3 Design of the Spike Detector Stage

Real time spike detection is an essential requirement for developing brain machine interfaces. Neural amplifiers are directly interfaced with the electrodes and are typically succeeded by spike detectors or analog-to-digital converters (ADC) for digitizing the recorded signals. In the proposed channel architecture in Figure 4.1, the output of the post-amplifier is connected to a voltage threshold comparator that performs binary spike detection. Spike detection is the process of identifying that an extracellular spike event has occurred which has been recorded by the neural amplifier system. Spikes carry important neural information. To reduce the amount of data transfer only spike information needs to be conveyed. This reduces bandwidth requirement of back-end processing viz, ADC, multiplexer, and wireless telemetry. After detection of spikes following are three degrees of bandwidth reduction in an increasing order: (i) transmit a portion of the waveform around the spike, (ii) transmit morphological features needed for spike sorting, (iii) only transmit the spike times. The later case offers ultimate bandwidth reduction which is used in our spike detection method.

One of the major challenges for a Spike detection system is to be able to adapt to the varying rate of incoming spike as it can drastically vary from 10 Hz up to 100 Hz. Hence spike detector should be robust with respect to changing firing rate patterns. Simultaneously, it should precisely be able to distinguish spike events with respect to the background noise. Robustness is necessary because of varying noise sources, firing rates, and SNR fluctuations. For implantable devices, these spike detector should also occupy less silicon area. An area-efficient algorithm is proposed whose design and circuit implementation are discussed in detail in the Chapters 5 and 6.

4.4 LFP Amplifier with a Modified Pseudo Resistor

In recent studies, LFP from intracortical region is shown to be useful in decoding kinematic parameters viz. position and velocity for controlling external devices and neuro-prosthetics [85–87]. However the usefulness of LFP signals is often masked by low frequency noise. In addition to low-frequency noise, suppressing the electrode induced DC offset voltage while leaving the LFP signal (from 10 mHz to 300 Hz) untouched represents a considerable circuit design challenge in an integrated device, since this requires the implementation of a high-pass cut-off frequency well below 1 Hz. The design of such an amplifier should start by lowering the input-referred noise. The main source of noise of the front-end amplifier is the OTA and the pseudo-resistor present in the first stage (CCIA) which significantly contribute to the overall input-referred noise of the amplifier which was derived for first stage amplifier noise analysis in equation (4.8). The first stage OTA noise can be minimized by increasing the transconductance G_{m1} through transistor sizing. For minimizing noise contribution from the pseudo-resistor, we note from equation (3.19) for a similar CCIA configuration, that f_c should be as low as possible. Simplifying (3.19), we obtain the corner frequency f_c for first stage of FEA as

$$f_c \approx \frac{1}{2\pi}\sqrt{\frac{3G_{m1}}{2{C_{in}}^2 r_a}} \tag{4.11}$$

Since f_c is inversely proportional to $C_{in}\sqrt{r_a}$, therefore r_a need to be chosen as large as possible (C_{in} can not be increased as it will decrease the input impedance) so that the high-pass corner of CCIA stage can be pushed into sub-Hertz range. Also, local field potentials are small of the order of tens of

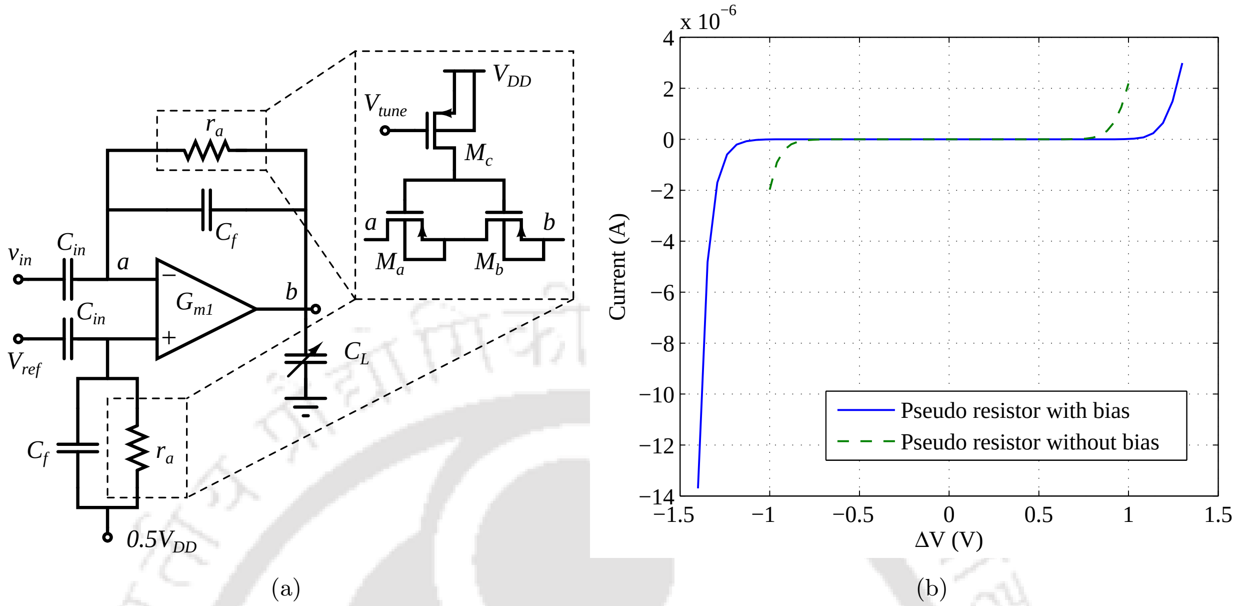


Figure 4.11: Effect of adding bias circuitry to MOS pseudo element (a) Modified on-chip pseudo resistor added to CCIA stage. (b) Linear range enhancement.

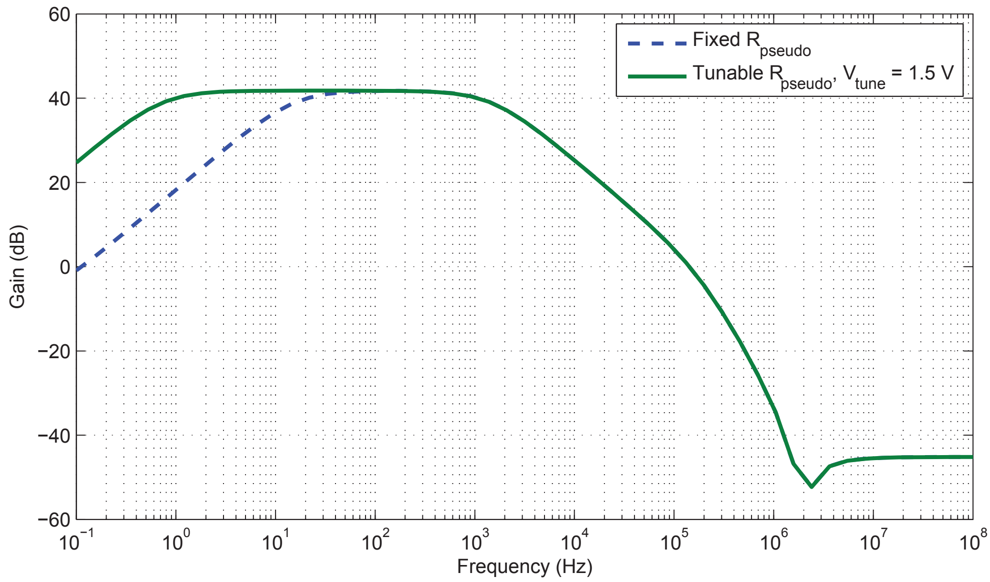
μV and have lower bandwidths of less than 200 Hz which make them susceptible to excess noise. At such low frequency flicker noise is dominant therefore $1/f$ and popcorn noise needs to be eliminated in order to prevent signal loss at lower frequencies and to filter out low frequency noise. For realizing the lower 3-dB corner frequency of the amplifier in the sub-hertz range, a high-pass filter having large C_f and r_a values is embedded with the front-end neural amplifier. The lower high-pass pole frequency is determined by the combination of feedback capacitor C_f and resistor r_a. As described in Chapter 3, the larger value of r_a can be realized with pseudo resistor elements. As a design trade-off between high gain and low corner frequency and because of the area limitation, the value of C_f is typically chosen to be in the order of 100 fF, which requires r_a to assume a large value ($>$ 10 GΩ).

The high resistance r_a is implemented with pseudo resistors which can be realized by biasing the gate voltage of pseudo resistors in a very small area without requirement of bulky passive resistors. An important property of a gate biased pseudo resistor element is that it can be tuned to the required bandwidth which allows for attenuation of the low frequency offsets and also produces a lower high-pass pole frequency to filter out low frequency noise and to correct for process variations. Neural amplifiers reported by Horiuchi et al. [88] and Olsson et al. [66] used gate controlled NMOS pseudo resistor to achieve tunable low-cutoff.

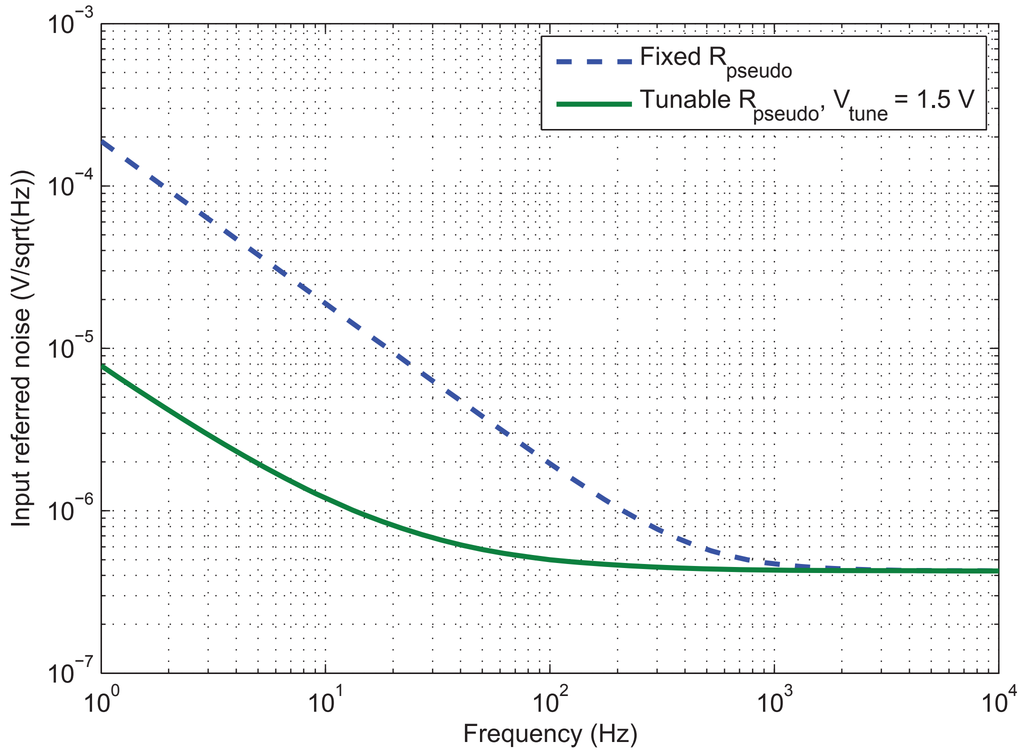
In the LFP amplifier of Figure 3.1, the resistive component r_a is realized with gate-biased PMOS transistors. The device acting as a pseudo resistive bipolar element is biased in weak inversion region. Assuming $V_{SG} = 0$, the effective resistance is exponentially dependent on its V_G in weak inversion region therefore it is essential to maintain a stable dc bias. The drawback of these pseudo resistors is that although the resistance is constant through a voltage range of -0.5 to 0.5 V but V_{eff} of MOS device depends upon the voltage V_{SD} across it as a result there effective resistance is hard to control. This makes V_{eff} sensitive to the changes in input offset voltage and process variations which limits the linear range. The pseudo resistor r_a can also be realized with diode-connected MOSFET M_a and M_b as shown in the first stage of Figure 4.4. It can maintain a stable DC operating bias point for the OTA but it is not tunable. The value of these pseudo resistors are process dependent and has low linearity for large output swings. Moreover, the DC current flow through these transistors as a result their transconductance increases which increases their noise spectral density.

To reduce the sensitivity of fixed-V_{GS} pseudo-resistors, the gate can separately be controlled by another biasing circuit consisting of transistor M_c. The resistance of pseudo resistive element is increased by setting gate of the MOS transistor to a suitable voltage bias where the device is operated in subthreshold. The drain-source leakage current is very low while the drain bulk current dominates the device's behavior. MOS bipolar element has a much higher impedance than a weak inversion MOS transistor. For increasing the resistance value, the V_{SG} for transistor M_c and should be decreased. For $V_{SG} \approx 0$, the resistance of the device is of the order of tera-Ohms which is due to the low leakage current of the transistors in 0.18 μm SCL technology. This further decreases the low-cutoff frequency of the amplifier. Figure 4.11(a) shows the feedback resistor r_a implemented with MOS bipolar pseudo resistor element (MBE) with biasing arrangement. The resulting improvement in linear range of MBE is shown in Figure 4.11(b). This scheme has a suitable linearity over the desired frequency band. The scheme also shows less sensitivity to process variations, common-mode variations and offset voltage of OTA.

CCIA1 was designed with fixed biasing of pseudo resistor. While in CCIA2, the double tunable pseudo resistor are utilized as feedback element. The frequency response of the two configurations were simulated in Spectre simulator with SCL PDK. It is observed that there is an improvement in bandwidth of FEA as shown in Figure 4.12(a). The neural preamplifier is seen capable of amplifying low frequency signal in the range of 0.3 Hz to 6 KHz while rejecting large DC offset at the electrode-tissue interface. The noise performance in Figure 4.15 shows that the randomly generated electrode dc



(a)



(b)

Figure 4.12: Improvement in Frequency response and noise of FEA simulated in SCL PDK. (a) Lower corner frequency lowered to 100mHz. (b) Lowering of flicker noise at low frequency.

offset is reduced and flicker noise of circuits are attenuated in the front-end stage preventing saturation of neural preamplifier. However, it is observed that the noise level is higher than the expected value in SCL PDK. The results indicate that thermal-noise PSD increases by as much as a factor of 100 above the expected value for a 0.18 μm process. The reason can be given from the comparison result given in Figure 4.15 where the thermal noise is inherently higher for a single NMOS transistor in case of SCL foundary. The difference in excess thermal-noise measurements can be attributed to the high-flicker-noise corner frequency present for short-channel devices operating at high levels of inversion. This explains the input referred noise is thus much higher in Figure 4.28 when compared to UMC PDK though the flicker noise is much improved with the suggested scheme.

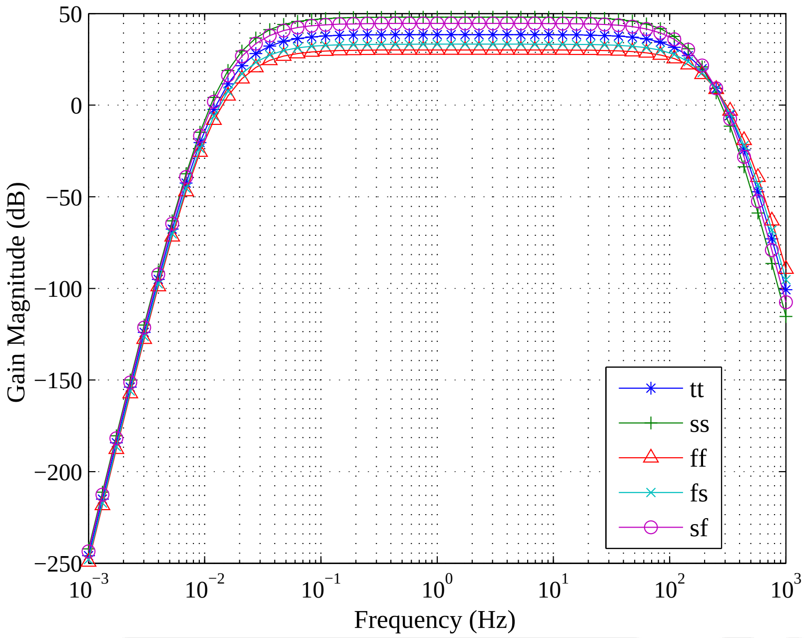
AC analysis of AFE output in typical process condition and in all the four process corners is shown in Figure 4.13(a). By maintaining a constant temperature, the worst case corners yield 42 and 48 dB of gain magnitude. Since the CCIA stage is designed for a programmable bandwidth with a cut-off frequency in sub-Hz range therefore Monte-Carlo simulations were carried out to verify the sensitivity of the LFP amplifier circuit to the device mismatches. Figure 4.13(b) shows a one thousand runs of Monte Carlo simulation of the high pass cut-off frequency at room temperature of 27°C. It can be seen that the histogram exhibits a standard deviation of 16.2 mHz around a mean value of 18.5 mHz. Both the analysis for the process corners and low cut-off frequency shows the effectiveness of the proposed configuration for bandwidth programmability of the LFP stage.

4.5 Prototype Results

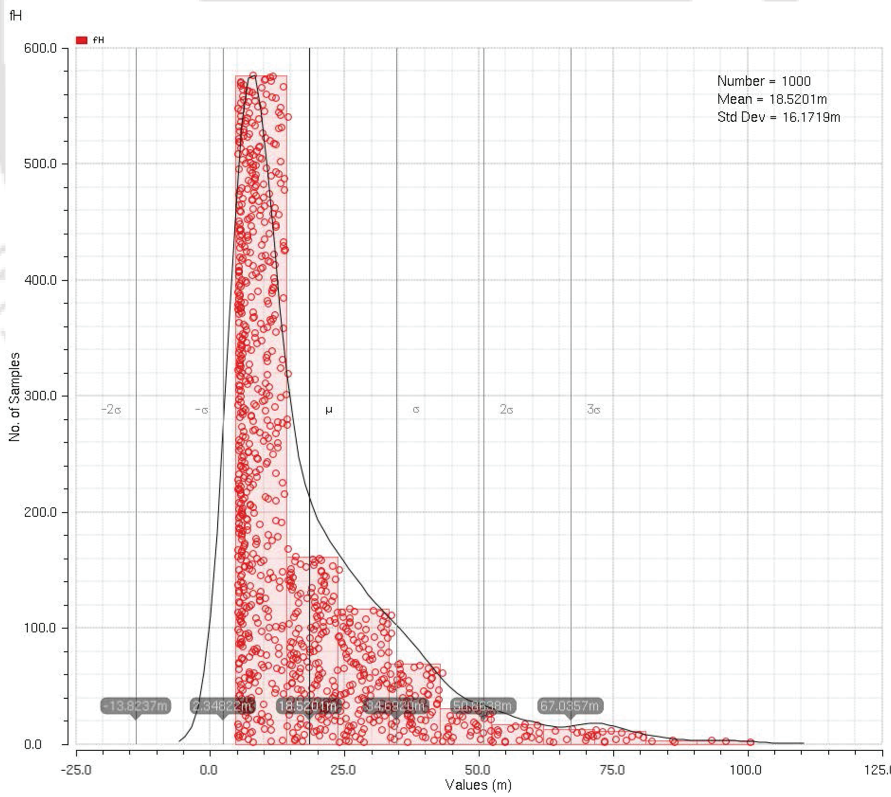
In this section, we describe experimental measurements of the two versions of our neural amplifier. For carrying out frequency-response measurements, swept-sine technique is used with an input amplitude of 10 mV.

4.5.1 CMOS Circuit Realization and Packaging

We fabricated the prototype neural amplifiers (CCIA1 and CCIA2) in the 0.18 μm two metal, one poly CMOS process of SCL (Semi-Conductor Laboratory) foundry. The amplifiers were designed for an approximate gain of 40 dB, setting C_{in} to 20 pF and to C_f to 200 fF. Both C_{in} and C_f were built as polypoly capacitors for maximum linearity. The bandwidth-limiting load capacitor were connected externally with a value of 100 pF. The amplifier circuits use 0.86 mm^2 of active silicon area, and 61.5% of this area is taken up by capacitors. All capacitors are implemented using metalinsulatormetal (MIM)

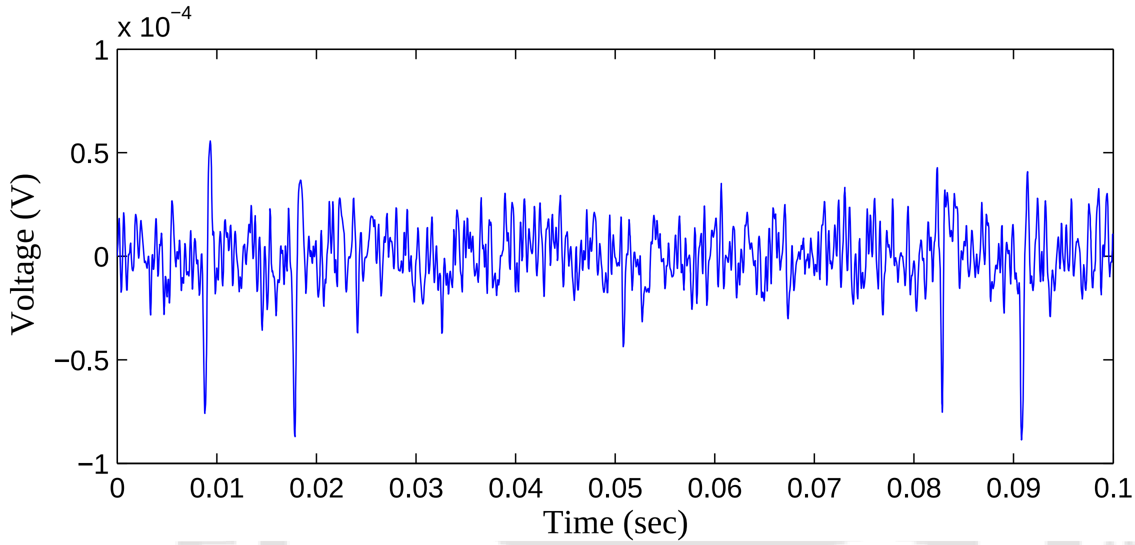


(a)

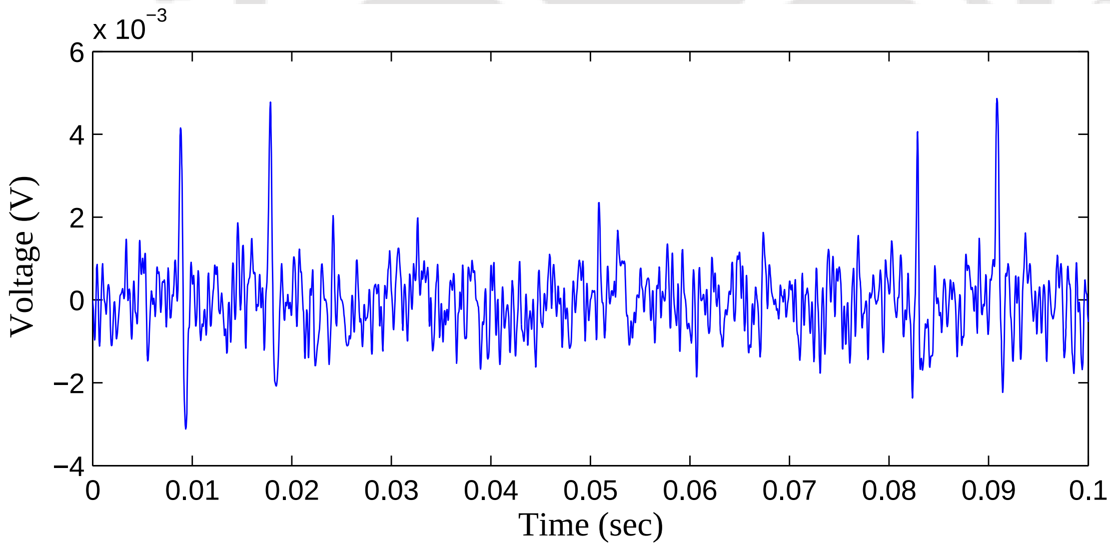


(b)

Figure 4.13: Improvement in Frequency response and noise of FEA simulated in SCL PDK. (a) Lower corner frequency lowered to 100mHz. (b) Lowering of flicker noise at low frequency.



(a)



(b)

Figure 4.14: Transient response of CCIA2 for amplitude of 50 μV (a) input neural waveform (b) amplified output.

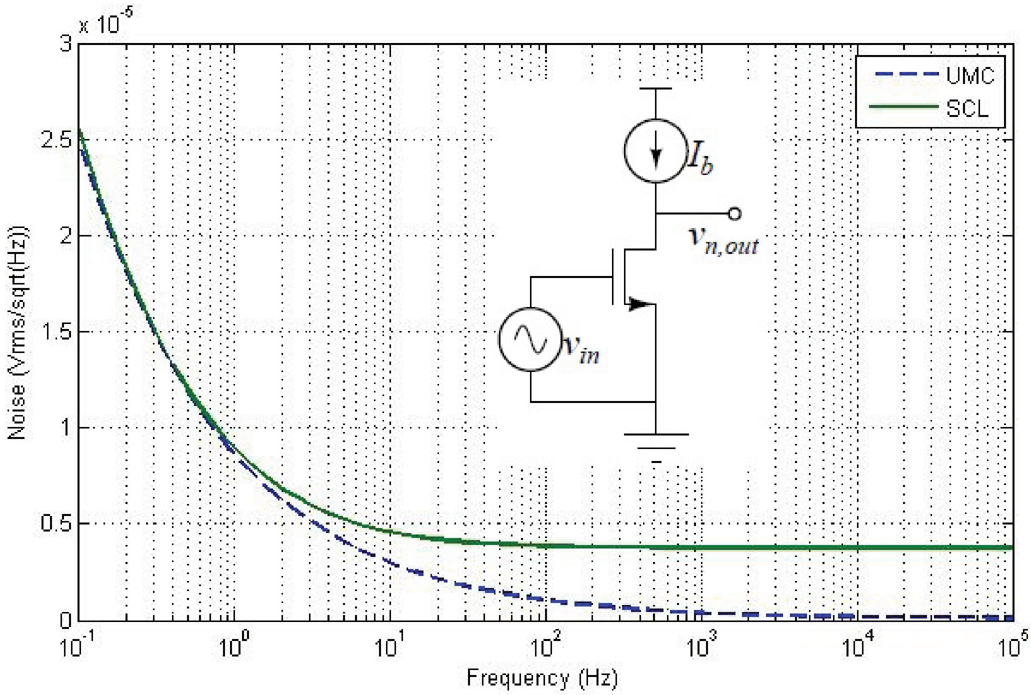


Figure 4.15: Noise comparison of two foundries.

capacitors. Figure 4.16 illustrates the finished layout of CCIA1 based neural amplifier along with bond pads. The CCIA2 based neural amplifier also has a similar structural layout. The core part of the layout which includes amplifier and capacitors is surrounded by the guard rings (shown with thick, heavy lines). The corresponding die microphotograph of a 1.112 mm $\times$ 0.932 mm chip containing a single amplifier is shown in Figure 4.17(b) out of which the active area is 0.056 mm^2. A small outline package (SOP) was used to encapsulate the integrated circuit chip with bond wires and pads. The pin diagram of amplifier chips packaged in a 14-pin SOP along with there dimensions are provided in Figure 4.18. Detailed pin description for CCIA1 and CCIA2 are given in Figure 4.19(a) and Figure 4.19(b) respectively.

4.5.2 Test Setup

Both amplifiers are tested separately on custom-made PCBs (printed-circuit board) . A schematic diagram of one of the PCB is shown in Figure 4.20(a). The schematic diagram is same of other PCB except for minor changes. All power supply lines and the common mode signal have a DC decoupling capacitor to ground to provide a low-resistance path to the disturbance. These decoupling capacitors are built using 0.1μF connected in shunt. The 0.1μF capacitor provides a low-resistance

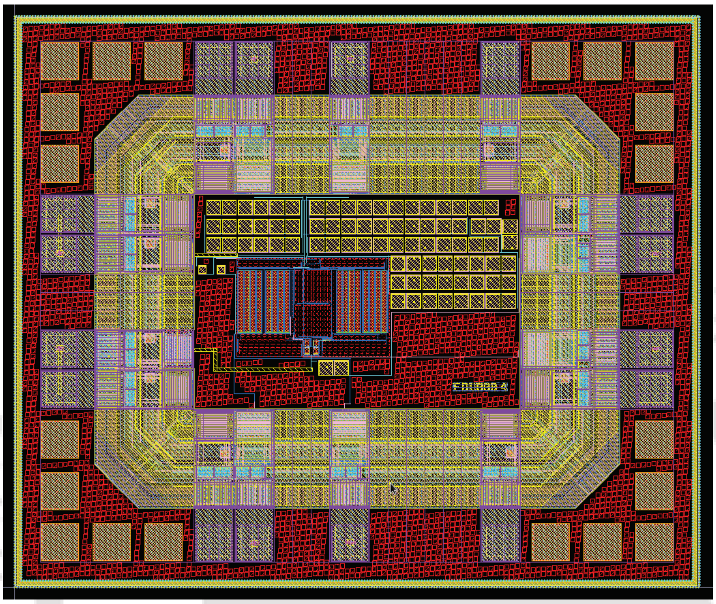
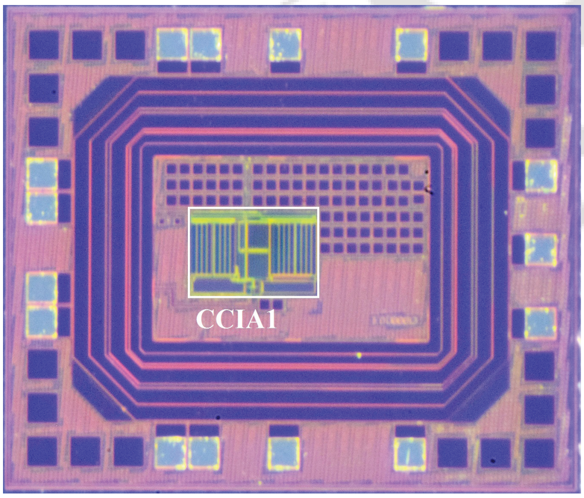
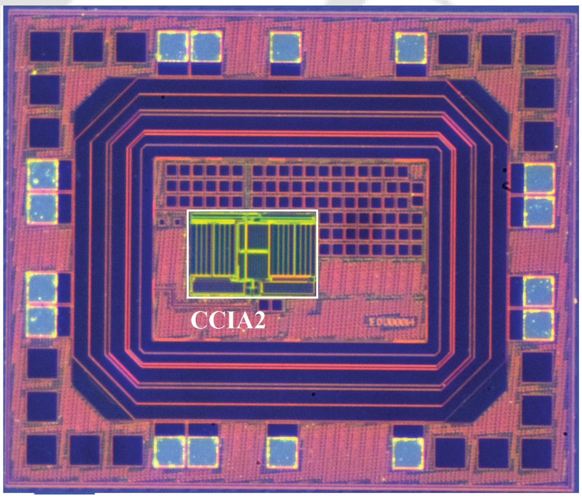


Figure 4.16: Final layout recommendation of the pre-amplifier (CCIA) chip along with bond pads placed for power supply and input-output connections.



(a)



(b)

Figure 4.17: Microphotograph of 1.112 × 0.932 mm chip containing a neural pre-amplifier (a) CCIA1. (b) CCIA2.

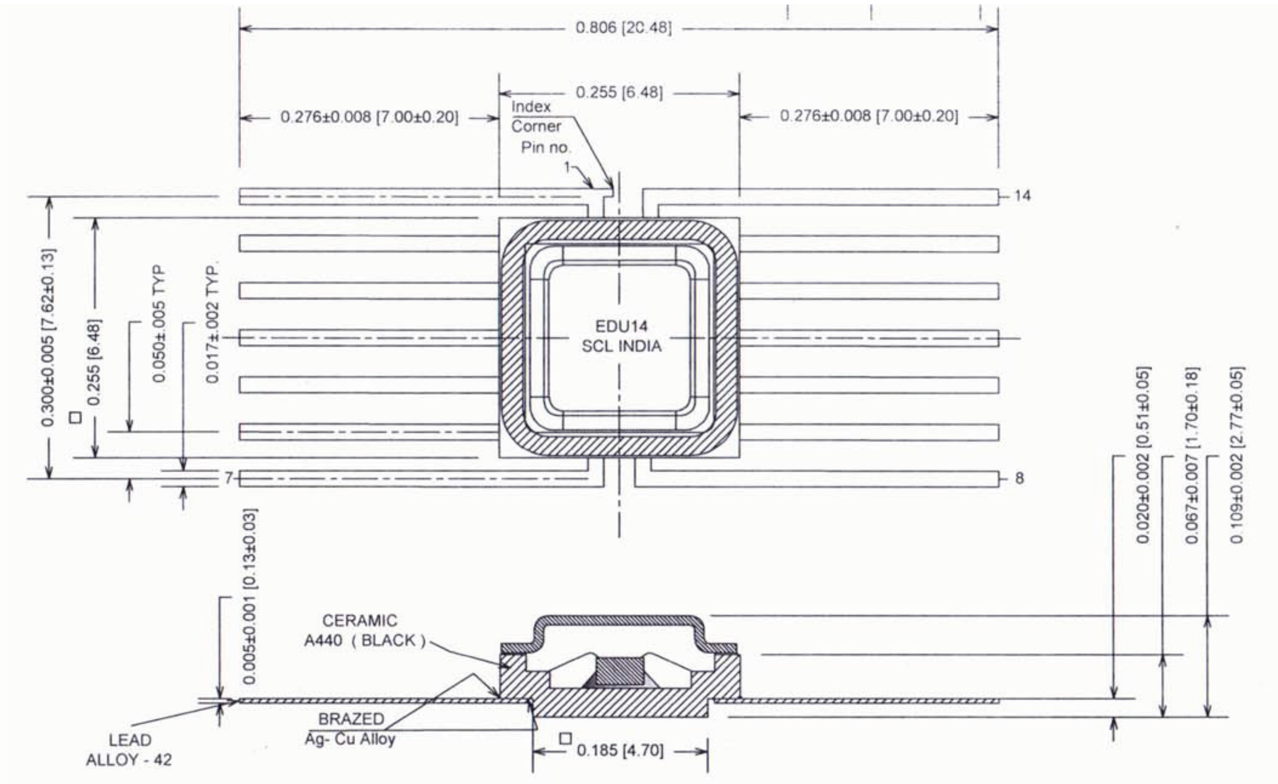


Figure 4.18: Pin diagram of 14-pin SOP-packaged neural amplifier chip.

Pin Description

Pin No.	Bond Finger	Pad No.	Pad Name	Pin No.	Bond Finger	Pad No.	Pad Name	Pin No.	Bond Finger	Pad No.	Pad Name
1	1	10	VP1	6	6	----	----	11	11	6	VOUT
2	2	----	----	7	7	2	VREF	12	12	----	----
3	3	13	VDD	8	8	3	IBIAS	13	13	----	----
4	4	14	VSS	9	9	----	----	14	14	9	VN1
5	5	----	----	10	10	----	----				

Remarks:- 1. Package pin / Bond finger no. 2, 5, 6, 9, 10, 12 & 13 are NC.
2. Die Pad no. 1, 4, 5, 7, 8,11, 12 & 15 are NC.

(a)

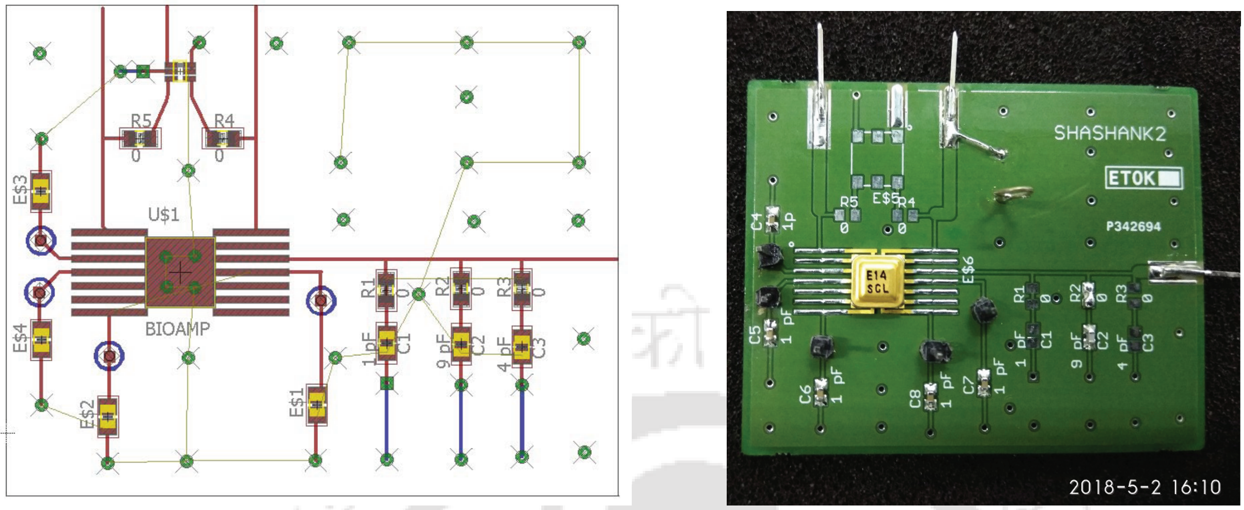
Pin Description

Pin No.	Bond Finger	Pad No.	Pad Name	Pin No.	Bond Finger	Pad No.	Pad Name	Pin No.	Bond Finger	Pad No.	Pad Name
1	1	11	VP1	6	6	----	----	11	11	6	VCON
2	2	----	----	7	7	2	VREF	12	12	7	VOUT
3	3	14	VDD	8	8	3	IBIAS	13	13	----	----
4	4	15	VSS	9	9	----	----	14	14	10	VN1
5	5	----	----	10	10	----	----				

Remarks:- 1. Package pin / Bond finger no. 2, 5, 6, 9, 10 & 13 are NC.
2. Die Pad no. 1, 4, 5, 8, 9, 12, 13 & 16 are NC.

(b)

Figure 4.19: The pin description showing configuration and function of the packaged neural amplifier chip representing (a) CCIA1 and (b) CCIA2.

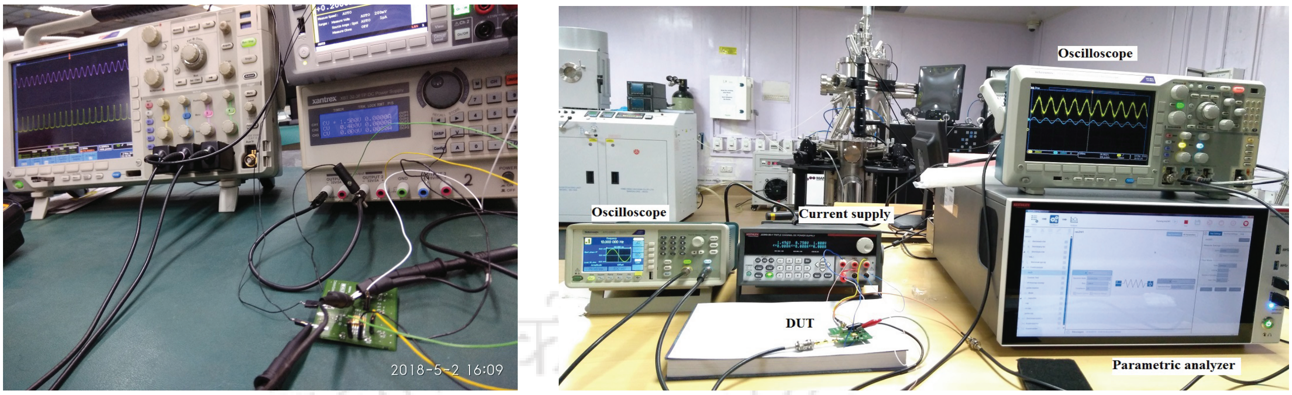


(a) (b)

Figure 4.20: Die photograph and PCB (a) Schematic diagram of the custom PCB (b) Populated test board with IC mounted.

path for low frequency signal. Frequencies above these are decoupled using 20 pF on-chip decoupling capacitor. All external passive components are 0805 surface mount devices. The packaged IC was then mounted on a custom-designed 59 × 47 mm printed-circuit board. The layout of the PCB is shown in Figure 4.20(b). A two-layer PCB is used for cost reasons. The bottom layer of the PCB is used as a ground plane for noise immunity and strong ground connection. The traces between SMA connectors and pins of the test-chip are micro-strip lines with 50 Ohms of characteristics impedance. They are terminated with a 50-Ohm resistor right next to the pin to minimize any refection. The differential input traces are symmetrical and are of equal length. In actual circuit testing one of the input terminals (v_{in}^{-}) was connected to PCB ground. The signal supply was provided from the other terminal (v_{in}^{+}). We have tested CCIA with closed-loop amplifier topology to ensure gain uniformity without the need for calibration. The functionality of amplifier was tested by populating the PCB using discrete components. Figure 4.21(a) and 4.21(b) respectively shows the set-up of the experiment for various electrical measurements carried out at SCL Chandigarh and at centre for Nano-Technology (CNT) in IIT Guwahati.

The measurement setup for characterizing the neural preamplifier ICs is shown in Figure 4.21. The test setup for frequency response and distortion analysis is shown in Figure 4.21(a) which was carried out at SCL Chandigarh. The remaining tests including noise measurements were done at IIT Guwahati whose setup is shown in Figure 4.21(b).



(a) (b)

Figure 4.21: Test setup for electrical measurement of PCB containing neural amplifier chip at (a) SCL and (b) CNT.

4.5.3 Measurement Results of CCIA1

The circuit design and integrated circuit layout of the neural amplifier stage was also done using SCL PDK. The neural amplifier performance was evaluated in Spectre simulator (Cadence tools). CCIA1 based neural amplifier layout as shown in Figure 4.17(a), was also simulated for post-layout results in HSpice using the standard 0.18-μm CMOS process with BSIM3v3 model. The analog power supply is set to 1.5 V. An external voltage reference of 0.75 V is used for biasing which is generated off-chip. Figure 4.22 presents the input (A3 channel) and output (A4 channel) waveforms of CCIA1 connected as a 30 dB gain for a power supply of 1.5 V. The input is a 1-kHz sinusoidal signal of about 12-mV amplitude. Voltage gain of instrumentation amplifier can be programmed with the change in bias current. Figure 4.23 shows the measured magnitude response of CCIA1 in the spike-recording setting for four different gains for a bias current setting of 50 nA, 200 nA, 250 nA, and 300 nA. The midband gain can be adjusted from about 28 dB to 37 dB with the increase in bias current from 50 nA to 300 nA. The upper and lower corner frequency is constant throughout this range. The lower 3-dB corner frequency is 0.8 Hz. While upper corner is at 1099 Hz without any capacitive loading. Due to the presence of large parasitic capacitance of band pads, the gain magnitude was found to have been reduced by several dBs while the power consumption increased significantly. Bandwidth is lowered because of secondary poles due to parasitic capacitances of output pad.

Since the designed CCIA1 is targeted to record LFP signal, our instrumentation amplifier can be configured to record field potentials in the range $1 - 300$ Hz by introducing a load capacitor C_L. Figure 4.24 plots the measured frequency response of CCIA1 for a load capacitor C_L of 100 pF when the bias

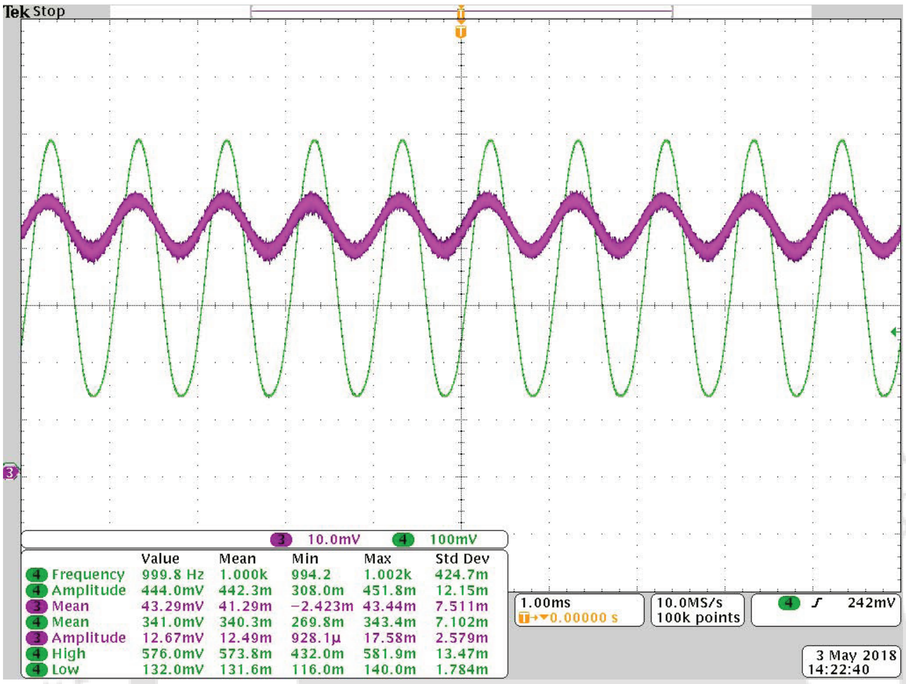


Figure 4.22: Amplifier input (A3 channel) and output (A4 channel) waveforms at a 1500-mV power supply. Measured output gain of CCIA1 based neural amplifier chip at I_{bias1} = 80 nA.

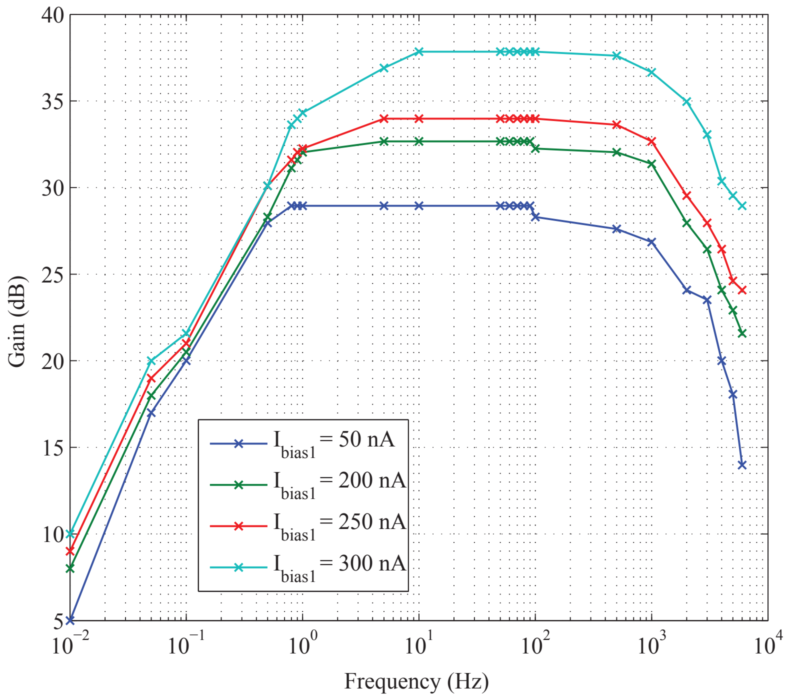


Figure 4.23: Measured transfer characteristic spectrum of CCIA1 with adjustable gain for a bandwidth of 0.8−1099 Hz.

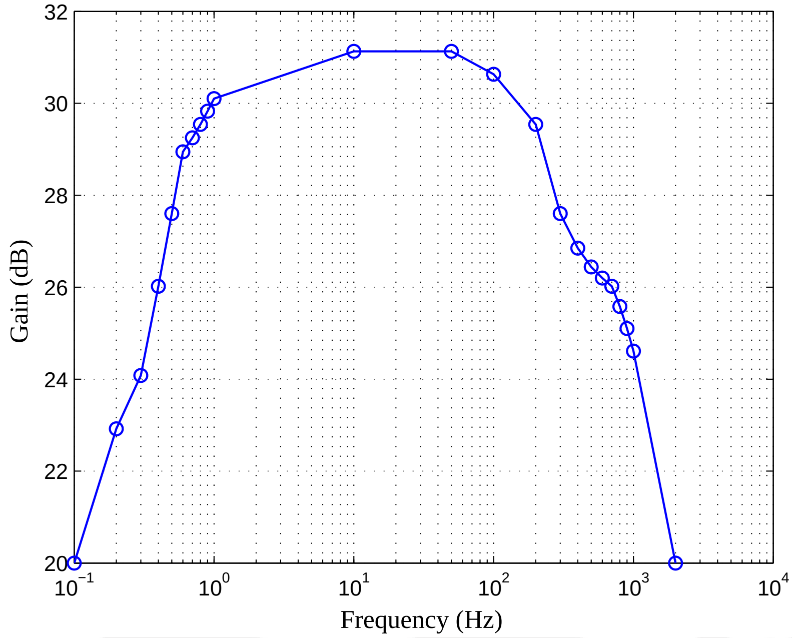


Figure 4.24: CCIA1 configured as LFP amplifier at $I_{bias1} = 200$ nA with load capacitor $C_L = 100$ pF.

current is set to 200 nA. The load capacitance should be larger than both feedback capacitance and the parasitic capacitance of the OTA, so that the dominant pole of the OTA is determined at the output node given by $f_U = G_{m1}/2\pi A_{m1}(C_L + C_{pL})$. The mid-band gain A_{m1} is around 31 dB. The lower and upper corner frequency are noted as $f_L = 0.8$ Hz and $f_U = 300$ Hz respectively. Power consumption is about 4942 nW. It may be noted that as C_f is very small, it is required to be well above neighboring feedback parasitic capacitances. Despite the fixed C_{in} and C_f, the mid-band gain was lower than design value. It is inadvertently caused by large parasitic capacitance C_{pf} due to bond pads (> 5 pF) and also because of fringing field effects. Dynamic range was much reduced due to output noise from sources including interference from power supply and electrically or magnetically induced interference. Linearity performance is illustrated in Figure 4.25. The figure shows the frequency response plot of CCIA1 for a 200 Hz input tone with 190 mV$_{pp}$ amplitude. It can be seen that the third harmonic is about 48 dB below the fundamental.

4.5.4 Measurement Results of CCIA2

Another set of experiments were performed with the CCIA2 chip. The microphotograph of CCIA2, implementing a bias control circuit for the pseudo resistor is shown in Figure 4.17(b). For this prototype chip, the tuning voltage V_{tune} was set to 1.0 V. For LFP operation, the bias current of

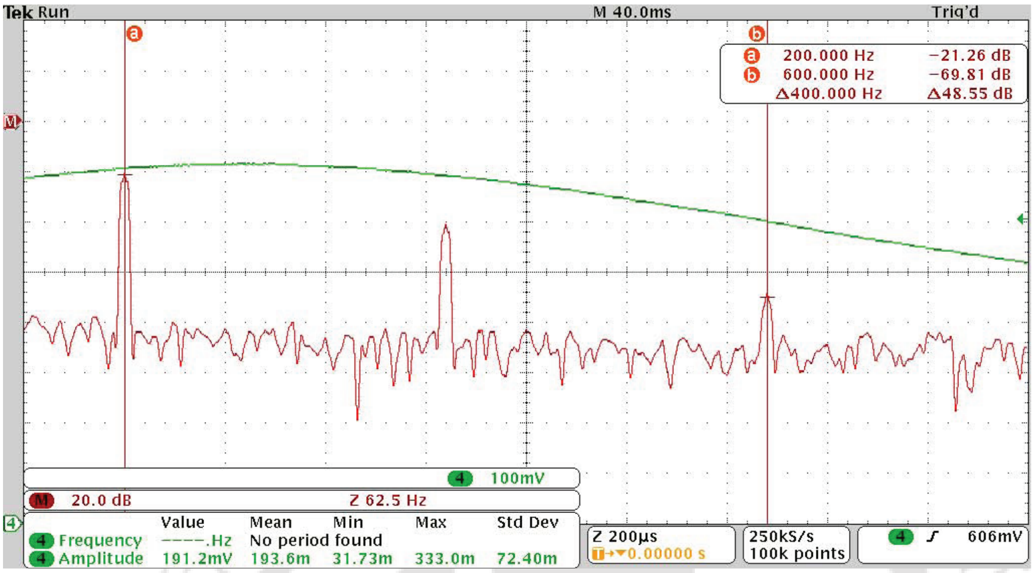
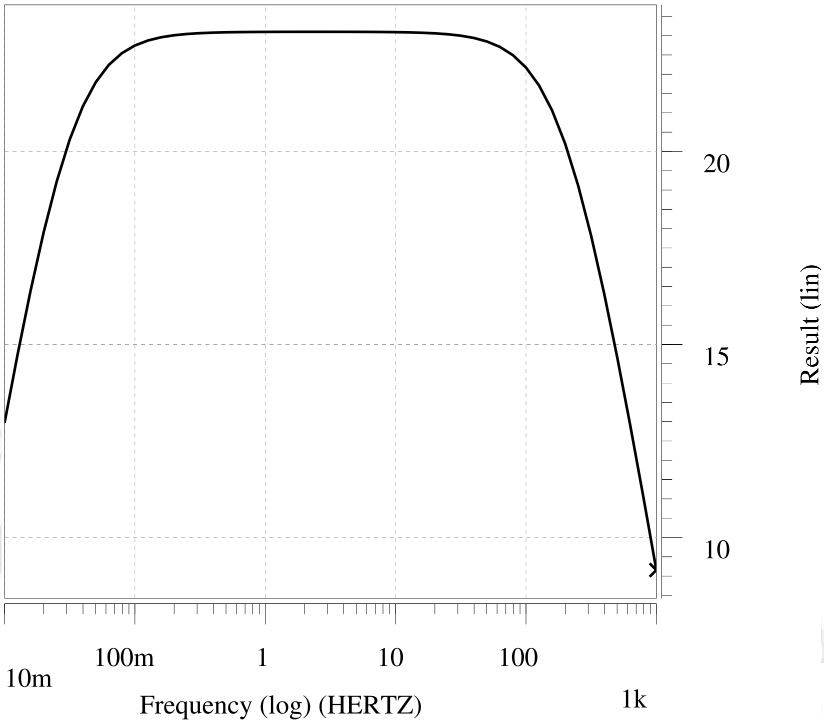


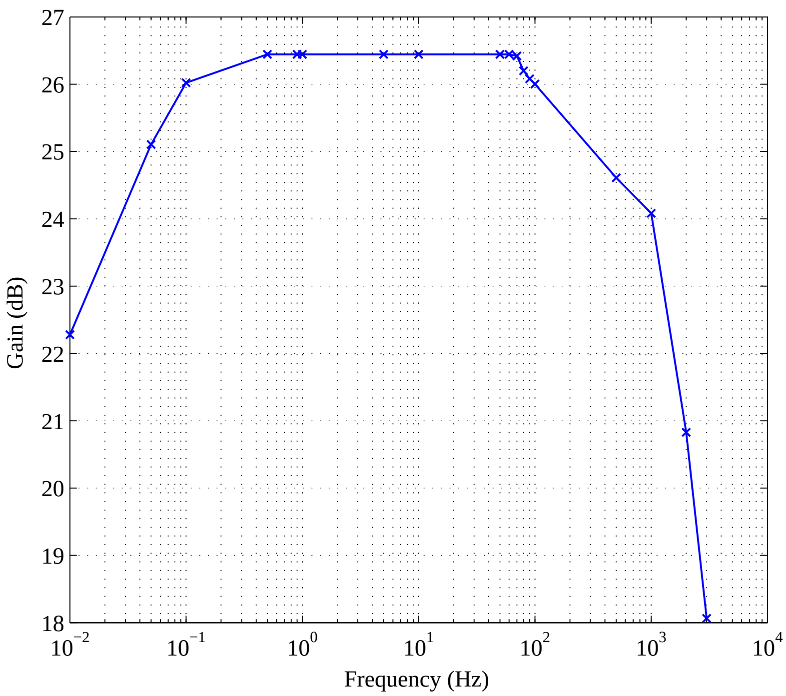
Figure 4.25: Measured output distortion of CCIA1 based neural amplifier chip. HD_3 =48 dB

CCIA2 was increased to 140 nA. The post layout simulation of frequency response of CCIA2 is plotted in Figure 4.26(a) for a bias current settings of 140 nA. The fabrication result of CCIA2 for same bias current is presented in Figure 4.26(b). The simulated and experimental results are in good agreement. The mid-band gain was found to be 26.65 dB. The 3-dB bandwidth is noted as 1 KHz. The gain magnitude and unity-gain bandwidth is low due to presence of large parasitics as well as the interconnect non-ideal components together with the ESR protection circuits of pads, and bond wires forming various parasitic RLC circuits. As a result, the current bias has to be increased to achieve the same voltage gain as the simulated result due to which there is an increase in power dissipation which is found to be 9.03 μW.

To configure the amplifier for recording LFP signal the load capacitor value was increased to $C_L =$ 100 pF. The measured result of CCIA2 for bias current of 140 nA is plotted in Figure 4.27. The lower corner frequency was shifted down to 3 mHz. This filters out any DC offset added by the surface electrodes. The lower corner frequency was also found to be lower than expected which is because of the feedback capacitive effect in the equation $f_L = 1/2\pi(C_f + C_{pf})R_f$, where C_{pf} is the parasitic capacitance in feedback. The upper corner frequency was found to be 340 Hz. Such high value of load capacitor C_L was chosen due to large value of parasitic capacitance ($C_{pL} \sim 40$ pF) at the output of OTA. A high-cutoff frequency should not to be ideally realized in the first stage since it needs an external capacitor. Since only first stage is realized for LFP configuration, therefore our amplifier



(a)



(b)

Figure 4.26: (a) Simulated (post-layout) frequency response of CCIA2. (b) Measured output gain of CCIA2 based neural amplifier chip at I_{bias1} = 140 nA without load capacitor.

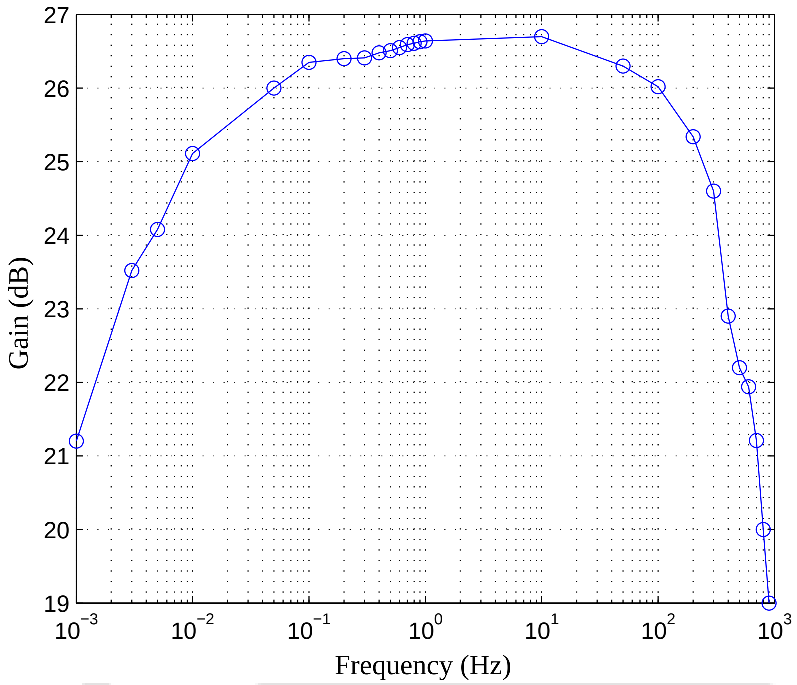


Figure 4.27: Measured output gain of CCIA2 based neural amplifier chip at I_{bias1} = 140 nA with load capacitor of 100 pF.

does need an external passive component to define the frequency response of the CCIA2 amplifier. Nevertheless, it is possible to replace the external passive components of the CCIA stage with the MOS based capacitors integrated in the next prototype AFE IC.

A comparison plot of the measured input-referred noise spectrum for zero input signal of CCIA1 and CCIA2 in the LFP-recording setting is shown in Figure 4.2(b). Integrating the input-referred noise spectral density curve from 1 Hz to 300 Hz for both the amplifiers yield a total input-referred noise of 9.64 and 5.11 μV_{rms} with a amplifiers bias current of 200 nA and 140 nA respectively. It is interesting to note that with the improvement of bias configuration CCIA2 has a better noise characteristics than CCIA1 even at lower bias currents. However, the overall noise curves for both the amplifiers are somewhat higher than the simulated results. It is due to increase in thermal noise due to inadvertently high parasitic capacitances (of the order of pF) which have exceeded the designed feedback capacitance. Moreover, the noise models available from SCL are not very accurate and the intrinsic device noise parameters are inherently high in the reported documents. This discrepancy can also be confirmed from a comparison of foundries given in Figure 4.15. The scaled-up noise is also due to noisy environment inside the laboratory where these tests were conducted. The experiments

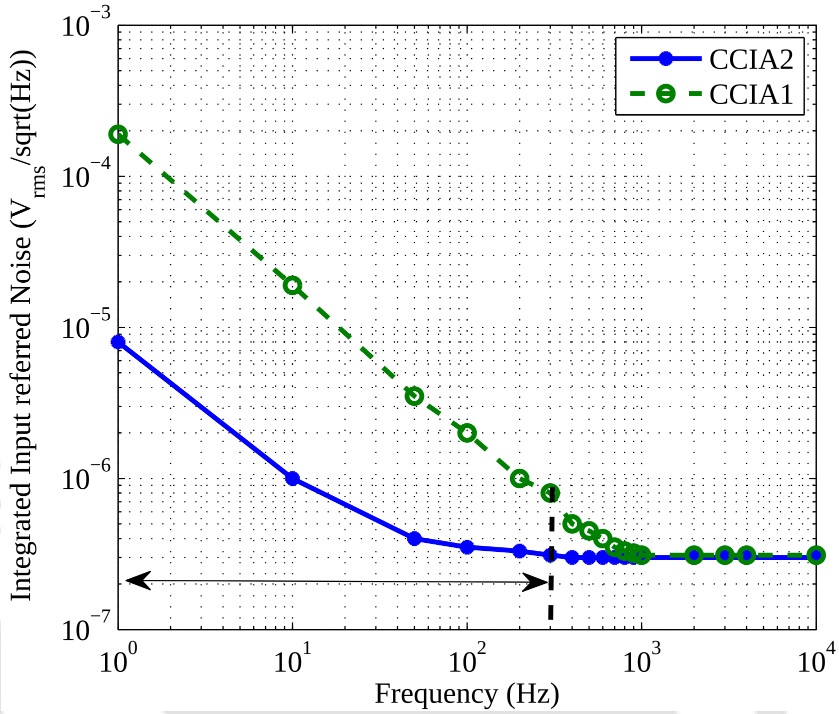


Figure 4.28: Comparison of input noise spectrum of CCIA1 and CCIA2 showing lowering of flicker noise of CCIA2 at low frequency.

were carried out in an unshielded environment which is depicted in a photograph of the test setup in Figure 4.21. This made the system more susceptible to interferers such as 60 Hz noise, which impact the measured noise performance. Another concern is the coupling noise from the connecting wires, cables and power supply instruments that have also made a significant contribution to the increase in the overall noise. Despite these differences, the order of improvement in input referred noise of the proposed CCIA2 circuit when compared to CCIA1 is found to be the same as the simulated results.

Table 4.4 summarizes the comparison of the measured parameters of this AFE chip with those of reported AFE IC works. It can be seen that the proposed AFE IC offers technical merits of lowest cut-off frequency, reduced power consumption, reasonable low layout area yet offers comparable measured results of the critical performance parameters such as noise and dynamic range. Although the proposed front-end interface implementation is not fully integrated with CCIA stage but it is justifiable in terms of performance in the context of the IC realization of the previous state-of-the-art devices.

Table 4.4: Performance Summary and Comparison with Prior Work

Parameters	[89]	[90]	[91]	[81]	[92]	[93]	This work (CCIA1/2)
Topology	Folded cascode	Cross-coupled + ADC	Chop amp + ADC	Chop amp + MADC)	VCO + NLC	self-biased	Double diff-amp
Signal	LFP	spike-LFP	LFP	spike-LFP	LFP	spike-LFP	LFP/LFP
Technology (nm@V)	180 @1.8	65 @0.5	65 @0.5	130 @1.2	40 @1.2	180 @0.5/1	180 @1.5
Area/Ch. (mm^2)	1.7	0.013	0.025	0.018	0.135	0.098	0.086
Power/Ch. (μW)	2	5.04	2.3	9.1	7	15	5/9
Bandwidth (Hz) Low cut-off High cut-off	0.05 100	10 300	1 250	1 5K	1 200	0.4 9.2K	0.8/0.003 310/340
Noise BW (Hz)	0.05-100	10-300	1-500	1-1K	1-200	0.4-9K	1-300
In ref Noise (μV)	1	4.3	1.3	4.2	5.2	5.18	9.6/5.1
Peak input	5 mV	3.5 mV	$\pm$0.5 mV	1 mV$_{pp}$	$\pm$50 mV	-	20 mV
SFDR (dB)	60	40	52	50	79	75	48
Z_{in}(MΩ)	8	∞	28	∞	∞	∞	∞

4.6 Conclusion

This work presents a micropower front-end amplifier for neural signal recording. The amplifier can be configured to record either neuronal spikes or local field potentials. A band-pass filter stage is added to the gain stage with a constant current bias. By varying the bias current of the gain stage, noise can be lowered without effecting the total bandwidth as the pole due to the load capacitor is located at a higher frequency than the pole provided by the filter stage. This allows bandwidth tunability through bias voltage of the filter stage while maintaining same supply current at the first stage. Changing the bandwidth in this manner makes the amplifier suitable for spike recording which eliminates LFP signals and aids in the subsequent low-noise spike detection. In the second part of the chapter, the first stage amplifier is configured for recording LFP signals in the band $1-300$ Hz. Since the frequency band of LFP signal extends below 1 Hz where most of epilepsy related neural activity occur [81, 94], the second version of the amplifier is designed using a biasing technique to push the high-pass corner frequency into sub-hertz range.

Both the amplifier versions were realized in the form of a fully integrated circuit. Experiments were conducted to verify the chip functionality and performance. Measurement results have shown that with a gate bias voltage configuration of the pseudo resistor, the amplifier chip can achieve a high-pass corner frequency as low as 0.003 Hz without using off-chip capacitors which is the lowest cut-off frequency reported till date. The measured noise was, however much higher then expected value. Much of the increase in noise level is attributed to the IC manufacturing process which in itself is not controllable. It needs better modeling for analyzing, understanding and potentially optimizing the effect of technology parameters on the performance. This would greatly reduce the design uncertainty and allow to design the chip more precisely for the expected power and noise levels. For other parameters, the fabricated micro-chips used for measurements and the experimental tests validated the consistency of test results. Our amplifier may be useful in BMI applications for neural prosthesis and epileptic detection systems for chronic monitoring.

5

A Novel Adaptive Algorithm for Real-Time Spike Detection

Contents

5.1 Introduction to Spike Detection

Brain-machine interface (BMI) enables communication between the human brain and a machine (e.g., computers) with an ultimate goal of restoring full motor function for people suffering from conditions such as spinal cord injuries or loss of limbs. One of the most important part of brain-machine interface is the development of fully implantable micro-electrode arrays (MEAs) for monitoring extracellular neural activities. Nordhausen [2] used high-density microelectrode array featuring a sensing area of 4×4 mm^2 that can record up to 100 extracellular neural signals simultaneously. Action potentials (AP), or spikes, generated electrochemically by individual neurons are the fundamental neuronal measure for neurobiological experiments. The time instants of spike occurrence encode important neural information [7, 13] which is critical in interpreting motor commands for BMI applications. Spike detection allows BMI to transmit only the useful information contained in the spike waveforms and their respective time stamps while ignoring data between the spikes. To achieve this data reduction, it is necessary for a high-performance spike detection to clearly discriminate neural spikes from the background noise.

Spike detection must be very accurate because missed detection errors propagate through the system as missed neural spike. False detections also propagate through the system as incorrect information. A systems that transmit windows of data around the spike often use lower thresholds to increase detection performance, but the lower the threshold the higher the power dissipation and required bandwidth which can cause a bottleneck in wireless data transmission so care must be taken when selecting the thresholds. One solution is to monitor the bandwidth and adjust the thresholds to maximize its utilization while remaining within the power limits. In addition, the spike wave shape needs to be preserved as it enables on-chip feature extraction required for treating neurological disorders such as seizure detection system for epileptic patients [95]. Section 5.4 presents the spike detection strategy and the algorithm.

5.2 Basic Problems in Spike Detection

Extracellularly recorded spike trains are invariably corrupted by variability of noise. Noise is present from various sources in the recorded signal. These sources are Johnson noise from individual electrodes, field potentials from distant active populations of neurons, and EMG signals from entire body. Another problem may be the variation in the level of background noise. If the background noise remains constant, the classifications will be consistent throughout the trial. If the background noise level fluctuates, there will be many more misclassifications especially during high levels of noise.

Often encountered problem is the variability of spike wave shape [96, 97]. During recording, the electrode drifts slowly to a new position as the neural tissue settles in response to pressure from the advancement of the electrode. This results in a gradual change in the shapes of the action potentials. These variations are dependent on position of recording electrodes, individual cell geometry, and distribution of ionic currents. Hence, the detection of spikes in noisy extracellular environment is a challenging problem. There are many reported algorithms that carry out the task of spike detection and are classified as supervised and unsupervised, manual or automated. In this work, our focus is on spike detection methods that are automated and unsupervised.

Also the spike height can vary greatly if there are other neurons in the local region that generate action potentials of significant size [96]. If the peak of the desired unit and the dip of a background unit line up, a spike will be missed. The spikes may have a sharp negative going transition, but the time phase and peak amplitude vary significantly. Hence, using digital filters, or frequency domain methods becomes difficult.

5.3 Literature Survey

It is well known that for obtaining best spike detection performance, it is important to capture multiple waveshape features so that extracted biopotential information can be utilized in subsequent processing steps [96, 98]. The essentials waveform features may include the time of occurrence, the maximum and the minimum amplitude value that can be associated with a specific neuron. Neural data processing is usually performed after the signal is amplified and filtered in a neural amplifier so that it is less sensitive to noise and occupies large dynamic range.

5.3.1 Spike Detection Methods

Various spike detection methods are adopted which compete for a high detection performance while restraining power consumption, area, and computational complexity within a specified limit. A bilateral amplitude threshold detection is discussed in [99] which was previously validated in [15] and [100]. This scheme allows preservation of biopotential waveshapes. However, capturing complete waveshapes require large SRAM memory blocks which require addition silicon area and increases complexity. Wavelet transform methods such as matched filtering [101], principal components [102] are based on template matching which require prior knowledge of spike shape and cannot be classified as unsupervised detection algorithms. Moreover, these methods require significant memory management. Other digital algorithms for spike detection and classification [103–109] are unsupervised but they are sophisticated and require high-speed computing.

5.3.2 Spike Detection with AT Method

For designing of implantable neural recording microsystems with more number of recording channels where size and power are two major concerns it becomes necessary to minimize the complexity and the amount of signal processing on the implant. Therefore, such detection methods cannot be employed in this type of system and spike detection is typically performed by simple thresholding or Amplitude Threshold (AT) triggering method. It is a conventional scheme for measuring the neuronal activity in the presence of background noise where the spikes are detected whenever the amplitude of a neural signal crosses a threshold level [110, 111].

Chandra [112] and Olsson [113] have described AT method by choosing a predetermined and fixed threshold on the basis of statistical characteristics of the neural signal. This method is simple in terms of hardware implementation; however, their performance is very sensitive to the background noise. Bharti [114, 115] has also adopted the amplitude threshold method. The analog detection technique uses few processing blocks although the power dissipation is high. Moreover, the detection performance is not discussed. Harrison suggested a spike detection scheme, in which the threshold level is adaptively set according to the standard deviation of background noise [14]. Although, the scheme adapts to the background noise reasonably well, it is not robust with respect to varying occurrence-rate of spikes. Besides, when there are large-amplitude spikes, the background noise estimation further departs from the real value. As reported by Quiroga, the noise estimation may deviate as much as twice of its real value with a moderate firing rate of 40 Hz [116]. It may be mentioned that, in practice, the occurrence rate of the spikes may vary from 10 to 100 Hz depending on the activity of neurons [64, 117].

Here it is shown that taking the standard deviation of the signal (including the spikes) could lead to very high threshold values, especially in cases with high firing rates and large spike amplitudes. In contrast, by using the estimation based on the median, the interference of the spikes is diminished (under the reasonable assumption that spikes amount to a small fraction of all samples) when the voltage threshold (v_{Th}) is set to

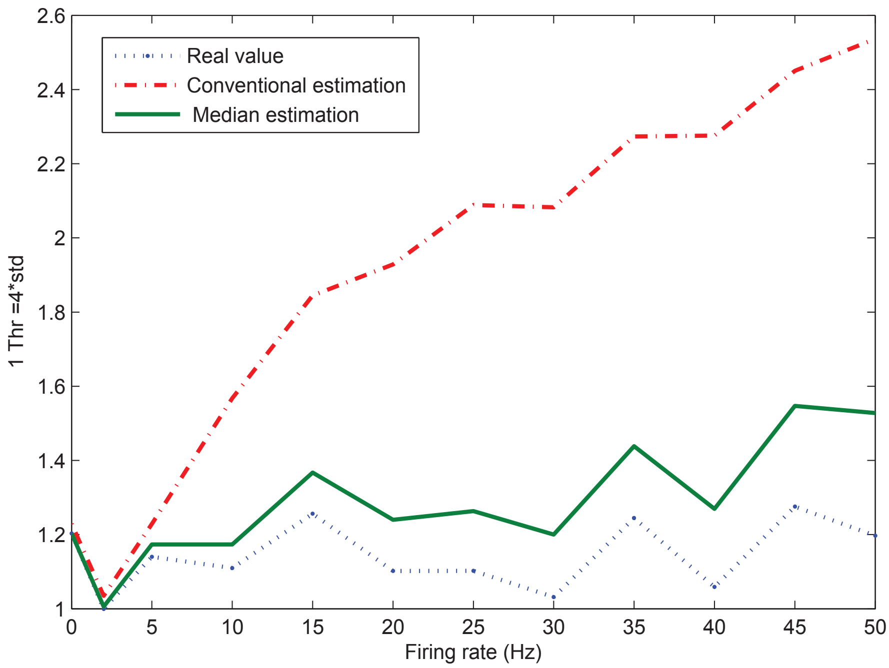


Figure 5.1: Estimation of noise level used for determining the amplitude threshold.

$$v_{Th} = 4\sigma_n; \sigma_n = median\left\{\frac{|Si|}{0.6745}\right\} \tag{5.1}$$

where Si is the bandpass-filtered signal and σ_n is an estimate of the standard deviation of the background noise [118]. To demonstrate this, we considered a segment of 1 sec of background noise taken *in-vivo* (Available: *http://web.mit.edu*) with a standard deviation of 0.147 and having 30 spikes of same class with different firing rates. Next, we eliminated a 2 *msec* of window around the spikes to obtain a pure noise data. Figure 5.1 shows that for noise alone (i.e., zero firing rate), both estimates are equal, but as the firing rate increases, the standard deviation of the signal (conventional estimate) gives an increasingly erroneous estimate of the noise level, whereas the median estimate from (5.1) remains close to the real value. But median based methods [116] require extensive computation that are difficult for hardware realization.

In order to further improve spike detection performance, NEO (non-linear energy operator) pre-processing along with adaptive thresholding have been suggested by Holleman [26] and Koutsos [27]. However, detection performances are not reported in these papers. NEO algorithm exploits the instantaneous frequency and amplitude information of the input signal. This approach enhances signal-to-noise ratio (SNR) and thereby improves the spike detection performance even under low SNR conditions [119]. In a recent work by Yao [1], a spike detector is proposed which estimates background noise by filtering NEO coefficients. The background noise estimation improves considerably, still, it is sensitive to the spike firing rate. For the spike firing rate of 50 Hz, the background estimate deviates by 20 times its real value. A NEO spike detection algorithm proposed by Yang is based on digital processing for calculation of threshold level in real-time [120]. The method performs accurately over a significant range of spike firing rate. However, the accuracy performance shows a slightly decreasing trend for firing rates greater than 40 Hz. In general, it is quite a bit of challenge to improve spike detection performance over a wide range of firing rates. A more robust methodology needs to be developed to extenuate these effects on the adaptive estimation of threshold value for larger variations in spike firing rates.

5.4 A Real-Time Spike Detection Algorithm

We present a robust method for the adaptive setting of threshold in real-time which is nearly independent of the input spike firing rate as well as the variations in spike amplitude. An algorithm is proposed which can eliminate spike waveforms from the raw data by means of energy-of-derivative method. The instantaneous background noise measured from the spike-eliminated data allows setting the adaptive threshold for spike detection. The improved performance of the proposed spike detector may be attributed to the following factors:

- The absence of spike attained by spike elimination technique makes it possible to maintain an unbiased estimate of the background noise and improves the accuracy of those measurements.
- The background noise estimation is independent of incoming spike amplitudes and their variations.
- The adaptive threshold is less susceptible to different filter cut-off frequencies and thereby overcoming the limitation of conventional NEO-based algorithms [1, 26, 27].
- It provides closed-loop stability to the adaptive threshold circuit without the need for calibration.

Figure 5.2 illustrates the fundamental blocks of the proposed CMOS detector in which a detailed schematic of the adaptive spike detection technique is described. The continuous-time neural data-set, $v_{in}(t)$ is passed through a preprocessing block. The preprocessor implements the energy-of-derivative

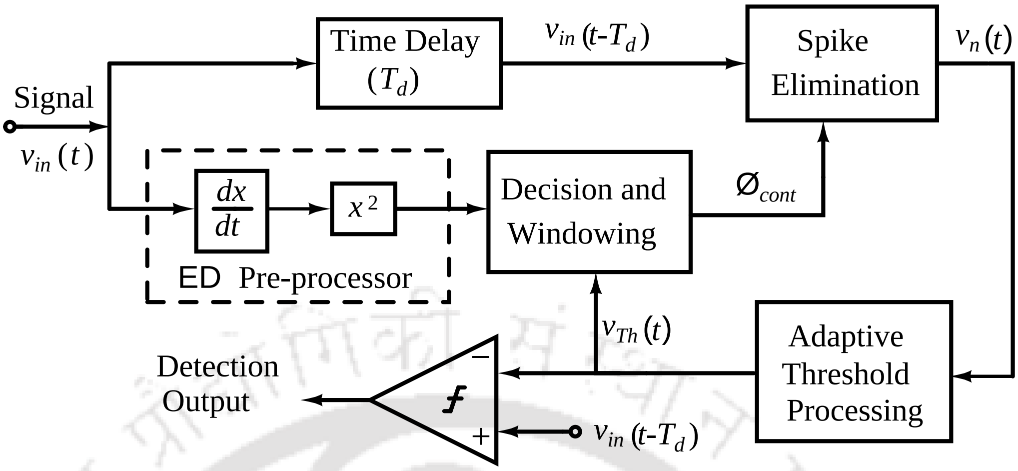


Figure 5.2: System block architecture of single, positive threshold based adaptive spike detection circuit.

function (ED), $\psi(.)$. The ED method is an approximation of the conventional non-linear energy operator (NEO). It is used for accentuating spike shapes with respect to the background noise in the presence of large baseline variations [121]. Energy-of-derivative method is particularly appealing because it can be implemented using simple hardware, keeping the complexity relatively low. This makes it more suitable for low-power analog implementation. ED algorithm processes the continuous-time signal $x(t)$ according to the equation

$$\psi\left[x\left(t\right)\right] = \left(\frac{dx\left(t\right)}{dt}\right)^2 \tag{5.2}$$

As shown in Figure 5.2, adaptive threshold v_{Th} is applied to the pre-processed waveform $\psi\left[v_{in}\left(t\right)\right]$ in the decision- and window-generation block. The location of spike event is determined from the threshold crossing mark as described in [122]. The threshold level v_{Th} is set above the background noise activity so as to reduce false detections. Upon crossing the threshold level, a spike event is detected which generates a control signal denoted by ϕ_{cont}. The spike indicator signal ϕ_{cont} controls the spike cancellation process.

5.4.1 Spike Elimination Strategy

The procedure adopted for spike elimination (referred here as spike elimination technique or SET) is illustrated with the timing diagram of a neural waveform in Figure 5.3. The neural signal $v_{in}(t)$ in Figure 5.3(a) shows a single spike of duration $t_3 - t_1$. It is delayed by T_d $(= t_2 - t_1)$ in Figure 5.3(c). The time delay T_d is introduced to hold the original waveform, for the duration which includes the

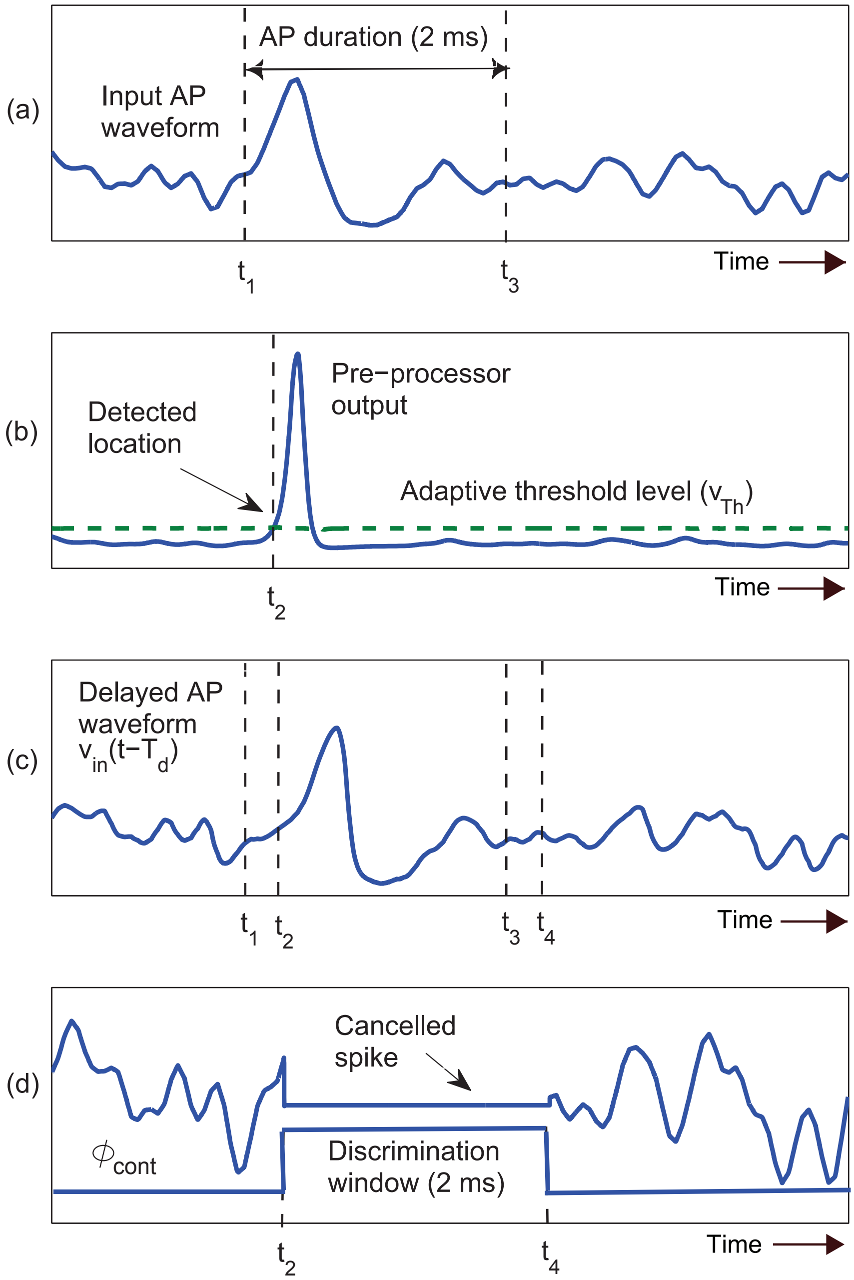


Figure 5.3: Spike detection strategy. (a) Input AP waveform, v_{in} with duration $(t_3 - t_1)$. (b) Preprocessed AP, $\psi(v_{in})$ showing the detection location at threshold crossing, t_2. (c) v_{in} after a time delay of $T_d = t_2 - t_1$. (d) waveform v_n with the canceled spike. Discrimination window $(t_4 - t_2)$ is generated upon threshold crossing.

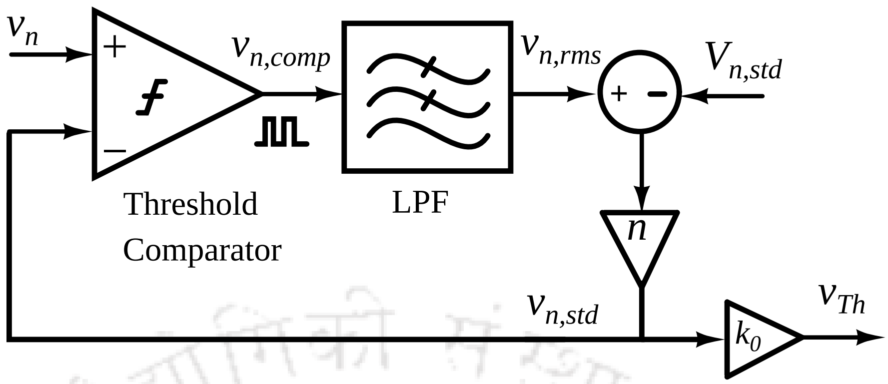


Figure 5.4: Circuit schematic of the adaptive threshold biasing loop.

time elapsed from the onset of AP activity to the threshold crossing point and the post-processing time of the window generator block. This operation preserves waveform integrity. T_d is determined from the average duration of the part of AP waveform happening before the detection point. The duration of AP onset to the threshold crossing was empirically estimated to be 390 μsec while the post-processing time of comparator and window generator was around 20 μsec making up a total delay of $T_d = 410$ μsec.

The delayed waveform $v_{in}(t - T_d)$ is fed to a digital switch for spike cancellation. With each threshold crossing of preprocessed waveform (marked by the time instant t_2 in Figure 5.3(b)), a control signal ϕ_{cont} of $T_s = t_4 - t_2$ ($\approx$ 2 msec) duration is asserted. This duration is set in confirmation with a typical spike length in order to adequately delimit the action potential of the incoming neural waveform. The spike cancellation block annuls a 2 msec window around the action potential in the delayed version of the original signal leaving only the noisy background activity, v_n as depicted in Figure 5.3(d). The start time and stop time of the annulling process is controlled by the ϕ_{cont} pulse which bears the timestamp of each spike occurrence.

5.4.2 Adaptive Threshold Generation

Figure 5.4 shows the schematic for the generation of adaptive threshold through a proportional feedback control mechanism. The threshold voltage is derived from the "spike-free" signal v_n (SNR = 0 dB) which is obtained at the output of spike elimination block. A continuous threshold comparator compares the amplitude information of v_n with its extracted root mean square (RMS) value $v_{n,std}$ to generate a binary stream $v_{n,comp}$. Because the spike peaks are absent, the average value of $v_{n,comp}$ is the RMS value of v_n which is equivalent to σ_n, the standard deviation (std) of the background noise.

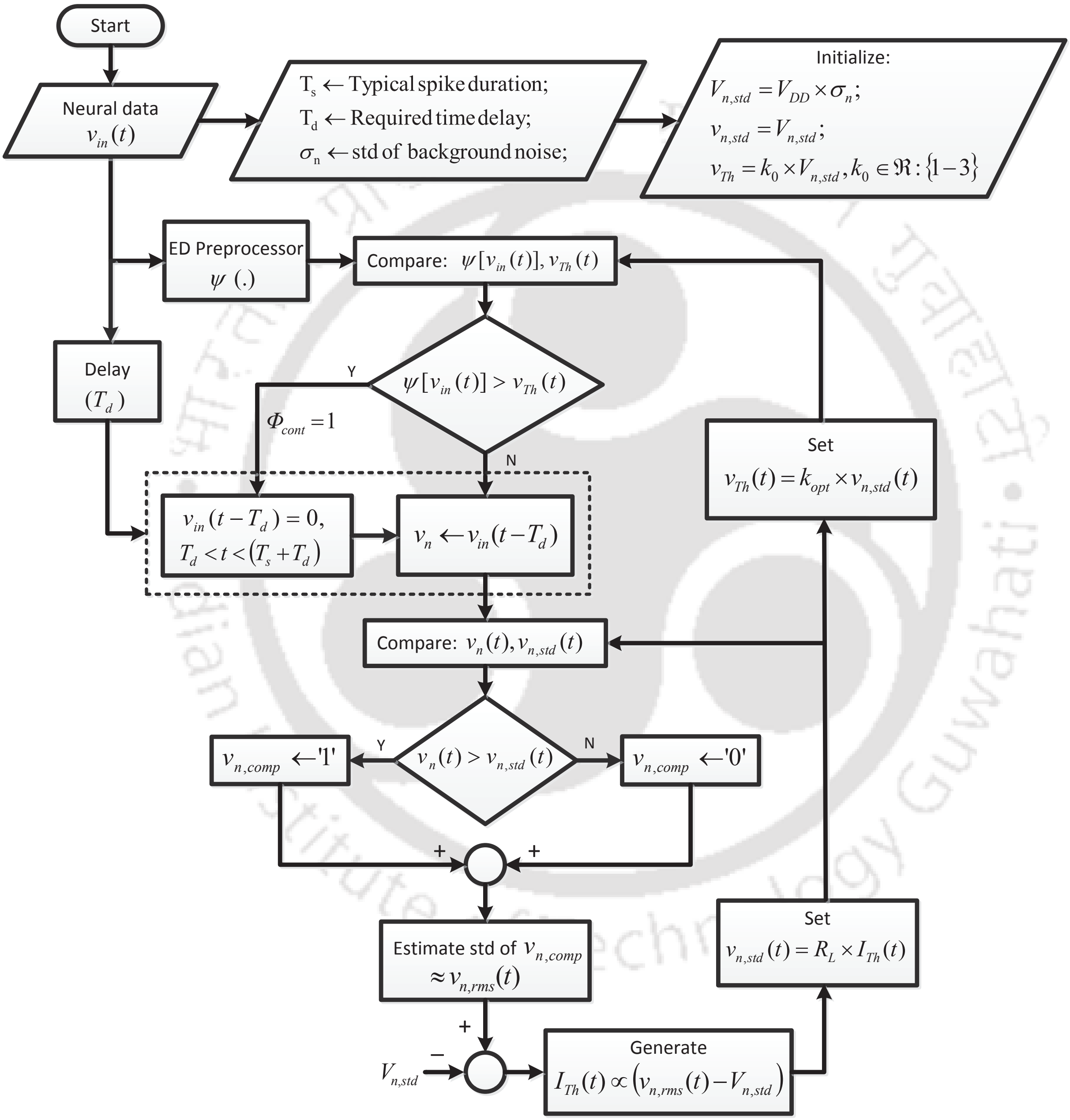


Figure 5.5: Flowchart of algorithm for adaptively setting threshold, v_{Th} based on $v_{n,std}$ and optimized scaling parameter k_{opt}.

A low pass filter (LPF) in the subsequent stage computes a weighted moving average of $v_{n,comp}$. The average value $v_{n,rms}$, at the filter output varies in proportion with the duty cycle of $v_{n,comp}$ signal. As the duty cycle increases, $v_{n,rms}$ approaches the maximum value V_{DD}. The averaged voltage level $v_{n,rms}$ is subtracted with $V_{n,std}$ which is preset to one std of background noise. An output voltage $v_{n,std}$ proportional to the difference between the two voltages is fed back to the threshold comparator. The feedback loop is stabilized as the difference in the magnitude of noise voltages becomes negligible. The estimated RMS noise $v_{n,std}$ is scaled by a constant k_0 so as to float over the background noise. The scaled value directly sets the adaptive threshold v_{Th} for spike detection. This threshold value is used for online data compression where the raw neural signal is transmitted whenever a spike is detected.

The flowchart of proposed algorithm for adaptively setting threshold based on $v_{n,std}$ and optimized scaling parameter k_{opt} is outlined in Fig. 5.5. The variable $V_{n,std}$ is defined as V_{DD} times the standard deviation σ_n of the background noise. $v_{n,std}$ and v_{Th} are initialized to $V_{n,std}$ and $k_0 \times V_{n,std}$, respectively. The dashed box shows the spike cancellation process. As a spike is declared by the decision logic, the delayed neural signal $v_{in}(t - T_d)$ is annulled for a duration of ϕ_{cont} pulse. The ON time of ϕ_{cont} pulse is specific to a one cycle time of action potential which is T_s. Next, $v_{in}(t - T_d)$ is assigned to a variable v_n. The spike-eliminated waveform v_n, when compared with its standard deviation $v_{n,std}$ gives a pulsed output $v_{n,comp}$. The digital signal $v_{n,comp}$ is filtered within system pass-band to provide a running average $v_{n,rms}$. $v_{n,rms}$ is compared with voltage $V_{n,std}$ to produce a proportional voltage $v_{n,std}$ for threshold comparison with v_n. $v_{n,std}$ scaled by k_{opt} times sets v_{Th} as an adaptive threshold level which updates its previous value. k_{opt} is the optimum value of k_0 determined empirically for the best detection performance.

5.5 Conclusion

A real-time adaptive threshold detection scheme using background noise estimation by the method of spike cancelation is presented. A linear phase delay filter need to be designed so that the neural waveform integrity could be preserved. Most of the conventional NEO based spike detection methods need a low-pass filter to improve detection performance. The proposed algorithm is based on spike cancelation scheme. The elimination of spikes makes the neural data free of spikes which should make the detector performance less sensitive to the choice of filter cut-off frequency. This would also suggest that the performance of spike detection should not vary with the increase in the spike firing rate. The next chapter deals with the implementation of algorithm with CMOS circuits. The topologies implementing individual building blocks should be carefully chosen and circuits should be properly designed such that the resulting layout should occupy less area and consume less power.

6

An Area-Efficient CMOS Spike Detector Circuit

Contents

6.1 Circuit Implementation

The input neural signal v_{in} is assumed to be amplified by the gain of more than 1000 (60 dB) prior to the detection stage so that the neural signals to be processed should have full-scale amplitude in 100 mV range. The operational transconductamce amplifiers (OTAs) implementing the building blocks of the detector are designed in weak inversion region. This allows ultra-low power operation over a lower range of bandwidth (1 Hz - 8 KHz) which is suitable for most bio-potential signals. In this chapter low power circuits are described which implements the spike detection scheme proposed in previous chapter. Simulation results validates the performance of the adaptive algorithm. The detection algorithm, when implemented as analog CMOS circuit achieves small size providing a low power realization of the scheme. We show that our topology offers a reliable spike detection performance at firing rates in excess of 100 Hz. The system implementation is described in Section 6.1. Section 6.2 presents the detailed simulations and performance comparison with reported spike detectors. Finally, Section 6.3 draws the conclusion on this work.

6.1.1 Adaptive Threshold processor

Figure 6.1 shows the delayed signal, $v_{in}(t - T_d)$ fed to a digital switch S_1 for spike cancellation. Upon receiving ϕ_{cont} signal, it performs a switching operation by grounding $v_{in}(t - T_d)$ signal for a period equal to T_s from the onset of the positive phase of ϕ_{cont}. This switching action removes the spike within a window of duration T_s with each detection threshold crossing. This ensures that only noise part remains so that the voltage threshold comparator B receives only the background noise v_n. Comparator B compares the noise signal v_n with voltage level equivalent to its standard deviation $v_{n,std}$ and outputs a binary signal $v_{n,comp}$ with a duty cycle T_{duty}. The average value of this binary waveform is calculated by the local averaging filter whose output $v_{n,rms}$ varies in proportion to its duty cycle.

The local averaging filter shown in Figure 6.1 is a single-input lossy integrator consisting of three OTAs having very low transconductances. The transconductor elements are implemented with a cross-coupled differential pair similar to the topology shown in Figure 6.3(c). MOS transistors employed in these transconductors are used to set the low-pass cut-off frequency of filter. The transfer function of the filter can be written as

$$H(s) = \frac{v_{n,rms}}{v_{n,comp}} = \frac{g_{m2}}{g_{m3}} . \frac{g_{m1}}{g_{m1} + sC_1} = \frac{A}{1 + s/\omega_n} \tag{6.1}$$

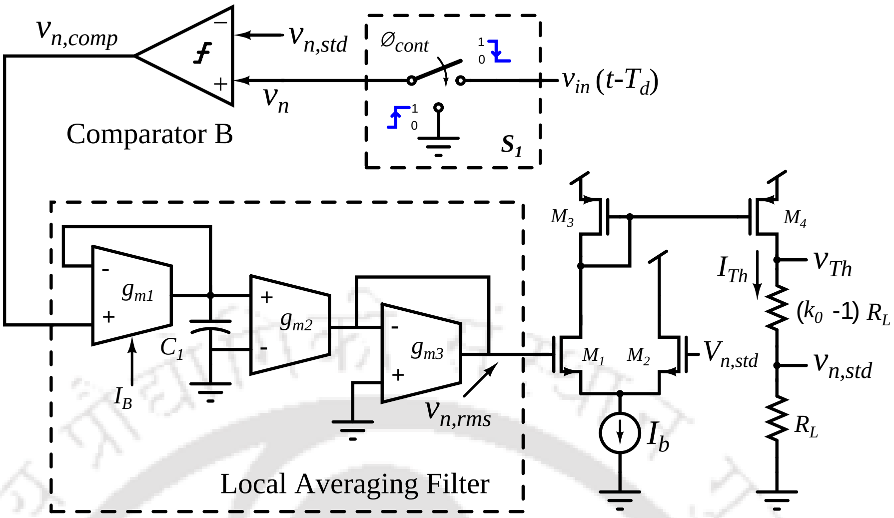


Figure 6.1: A low-pass $g_m - C$ filter for local averaging of $v_{n,comp}$ signal and feedback circuit for adaptive threshold adjustment and scaling.

where A is the mid-band gain. Frequency ω_n of the averaging filter is suitably chosen to adapt to the background noise and to smooth out the spurious transients present in noise voltage v_n. The transconductance, g_{m1} can be tuned with the help of bias current I_B so that f_n can be varied in the range of 1−70 Hz. Figure 6.2 depicts the plot showing linear variation of frequency f_c with the change in bias current. f_n is given by

$$f_n = \frac{\omega_n}{2\pi} = \frac{g_{m1}}{2\pi C_1} \tag{6.2}$$

The differential stage consists of transistors M_{1-4}. The input nMOS devices M_1 and M_2 compares the available time-averaged voltage $v_{n,rms}$ with the reference voltage $V_{n,std}$. The reference voltage is predefined to V_{DD} times the RMS value σ_n of the background noise. A fixed bias I_b sets the output current I_{Th} through the load resistor R_L. The current I_{Th} is converted to v_{Th} and $v_{n,std}$ by driving across a series combination of load resistor $r_1 = (k_0 - 1)R_L$ and $r_2 = R_L$. The voltage $v_{n,std}$ equals one standard deviation (σ_n) of the background noise. v_{Th} is typically set to $k_0 = 1$ to 3 times the $v_{n,std}$ value from the baseline. The resistor R_L is implemented with an nMOS transistor operating in deep weak inversion. It offers a linear resistance of 10 KΩ and also saves chip area.

6.1.2 Spike Enhancement Stage

The block diagram of energy-of-derivative (ED) pre-processor (inset of Figure 5.2) is shown in Figure 6.3(a). The hardware implementing ED pre-processor consists of a differentiator cascaded to a

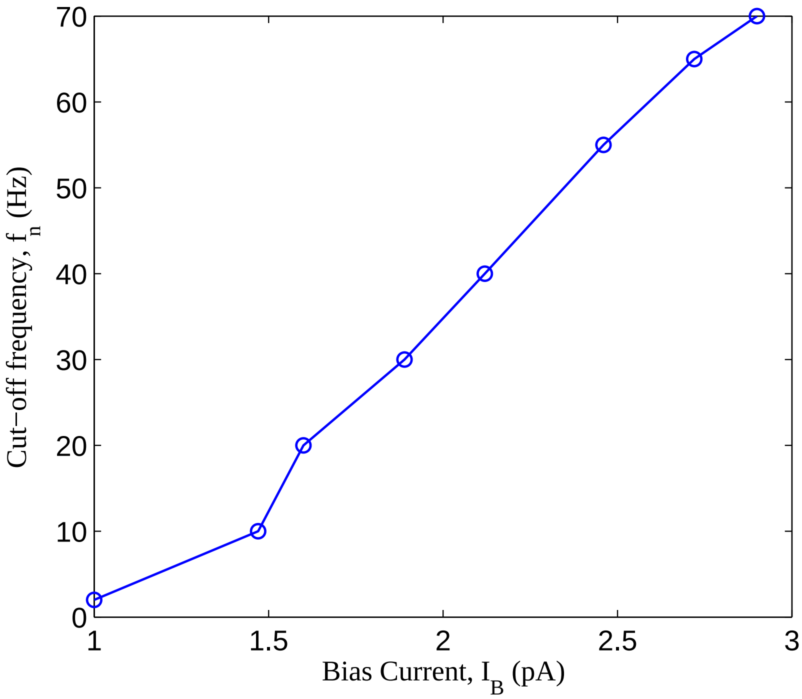


Figure 6.2: Variation of low pass cut-off frequency low-pass $g_m - C$ filter with the change in the bias current.

squaring circuit. A g_m-C based differentiator (Figure 6.3(b)) calculates input signal's first derivative. It assesses the slope of the neural signal by preemphasizing fast changing signal components which exceeds user-defined threshold. The differentiator circuit comprises of three transconductance blocks g_{m1}, g_{m2}, and g_{m3} along with one grounded capacitor C_2. The transfer function of the differentiator circuit is given by

$$H_{diff}(s) = \frac{v_{diff}}{v_{in}} = \frac{g_{m2}}{g_{m3}}.\frac{sC_2}{sC_2 + g_{m1}} \tag{6.3}$$

The transconductances are implemented by an ultra-low power OTA consisting of two cross-coupled N-type differential pairs whose circuit schematic is shown in Fig. 6.3(c). For realization of the differentiation function, the OTAs are configured by grounding bulk terminal of input transistors $M_{1,2,3,4}$ so that $v_{Y1} = v_{Y2} = 0$. The differential current i_0 in terms of differential input voltage $v_X = v_{X1} - v_{X2}$ can be obtained as

$$i_0 = KI_B \tanh\left(\frac{v_X}{2nU_T} + \frac{1}{2}\ln m\right) + KI_B \tanh\left(\frac{v_X}{2nU_T} - \frac{1}{2}\ln m\right) \tag{6.4}$$

where $U_T = kT/q$ is the thermal voltage and n denotes the subthreshold slope of MOS. I_B is the bias current, m is the aspect ratios of transistor pairs $M_{1,2}$ and $M_{3,4}$, and the constant scalar K

is the aspect ratio of the transistor pair $M_{7,9}$. The value of the scaling factor m is chosen as 4 in order to introduce equal and opposite input differential voltage offset for the maximum nonlinearity cancellation thereby enhancing the linearity of the OTA.

The differentiator is followed by a squaring circuit which evaluates energy of the signal proportional to its amplitude and frequency. The voltage squaring is achieved through analog multiplication. The topology in Figure 6.3(c) is configured for implementing a four quadrant multiplier by driving the gate and bulk terminals of the device based on the design suggested in [123]. For normal operation of MOS transistors $M_{1,2,3,4}$ source-bulk and drain-bulk junction should be reverse-biased and $v_X = v_{X1} - v_{X2} \leq V_T/\kappa$ and $v_Y = v_{Y1} - v_{Y2} \leq V_T/(1-\kappa)$. By using above conditions, the output current i_0 can be shown to be the product of two inputs.

$$i_0 = \frac{I_B.\kappa^2}{4.V_T^2}.v_X.v_Y \tag{6.5}$$

where $\kappa(= 1/n)$ is the gate coupling coefficient. A simulation result reported in Figure 6.3(d) shows a segment of raw neural waveform in the lower trace being applied to ED function. The upper trace shows the ED method successfully pre-emphasizing spikes from the neural signal. By providing correct discrimination of spikes from the background activity it makes the selection of threshold level for determining the location of spikes to be a trivial problem. The ED method also enables accurate detection of spikes even at lower SNR values. Processing by derivative method improves spike detection rate while reducing false detection of spurious spikes.

6.1.3 Decision and Window Generation

A low power comparator based on a regenerative track and latch operation is used to implement Comparator A. The same topology is also used to implement comparator B. Comparator A determines the time at which the positive going edge of $\psi(v_{in})$ at the node v_+ equals the positive threshold voltage V_{Th} at the node v_-. Comparator A provides a binary pulse $V_{comp,A}$ at the output, which carries the location of the detected spike.

With the arrival of the positive edge of $V_{comp,A}$, the time-window generator is triggered. Figure 6.4 shows block scheme used to generate the detection window. The SR flip-flop changes to a 'high' state switching-on the transistor M_N which quickly discharges the capacitor C_{slew}. The output pulse V_{slew} consists of a constant slope due to capacitor charging whose slope is given by $\Delta V_{slew}/\Delta t = V_{DD}.g_{ds,P}/C_{slew}$. The Schmitt trigger following the capacitor is implemented with transistors M_{N1-3}

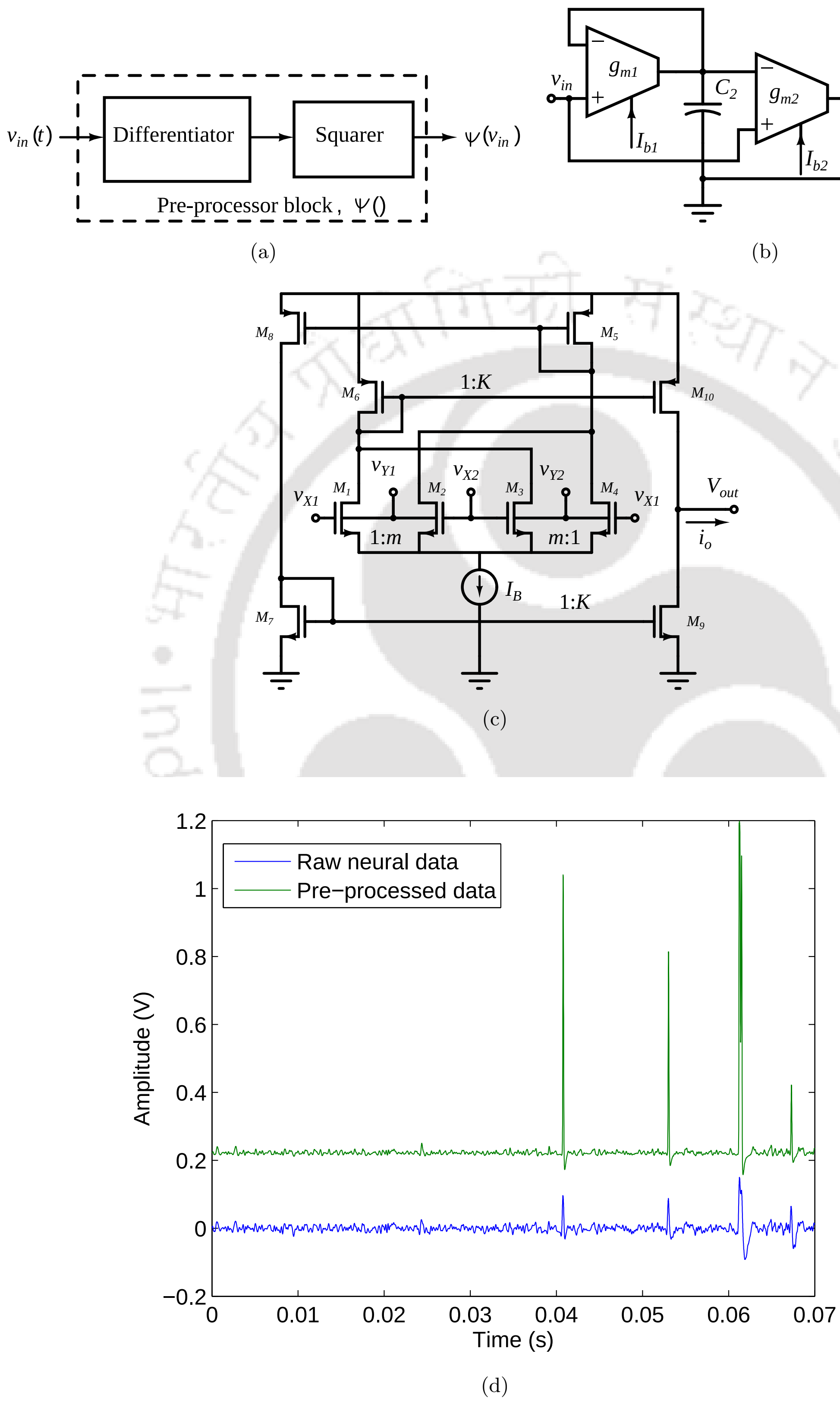


Figure 6.3: Major building blocks for implementation of the energy-of-derivative (ED) algorithm (a) The pre-processing stage (b) A $g_m - C$ based differentiator circuit (c) The topology of cross-coupled OTA used for the realization of differentiator and four-quadrant multiplier (d) ED preprocessor used for initial signal enhancement generates high SNR signal at the squarer output.

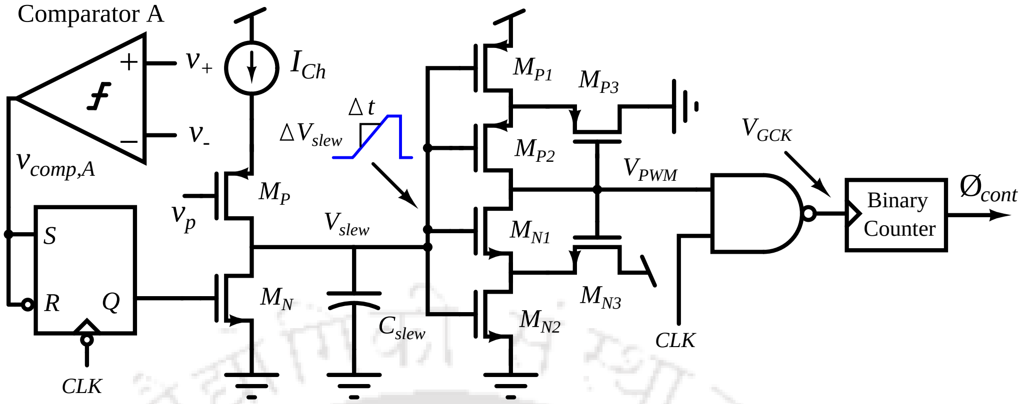


Figure 6.4: Implemented schematic of window generation. The detected pulse V_{comp} triggers the window generator which outputs 2 ms window discriminator, ϕ_{cont}.

and M_{P1-3}. It modulates the pulse width by converting the analog voltage stored in C_{slew} to the pulse timing of output signal V_{PWM}. The width of sampling pulse V_{PWM}, should be long enough for the following serial counter so as to clock through a 2 msec pulse at the serial output. The V_{PWM} signal is converted to a gated clock pulse, V_{GCK}. It is then fed to a 5-bit binary ripple counter clocked at 16 KHz (signal CLK). A positive impulse (ϕ_{cont}) of 2 msec duration is generated by the counter which resets after 32 clock cycles from the time an AP is detected. The window generation process rejects any positive threshold crossing that occur within 2 msec from the detection point. A 62.5 μsec clock is chosen to make sure it is greater than the overall propagation delay of digital sequencing blocks. The length of detection window can also be altered by varying the frequency of 5-bit counter in order to properly delineate the detected AP waveform.

6.1.4 A Linear Phase Delay Filter

A seventh-order all-pass filter was designed for a linear-phase response over the bandwidth of interest. The filter introduces a constant delay T_d to equalize for the time-shift the AP waveform undergoes from the instant of threshold crossing. The time of the onset of AP, which is the closest zero value preceding the detection point has been empirically determined to be ~ 390 μsec. The filter is made up of a cascade of a first-order all-pass filter along with three biquadratic all-pass section [124] yielding a phase delay of 410 μsec. The transfer function of first-order and the biquad section is [124]

$$H_1(s) = \frac{g_a}{g_b}\frac{sC - g_1}{sC + g_1} \tag{6.6}$$

$$H_{n+1}(s) = \frac{g_a}{g_b}\frac{-s^2C^2 + sCg_{2n+1} - g_{2n}g_{2n+1}}{s^2C^2 + sCg_{2n+1} + g_{2n}g_{2n+1}} \tag{6.7}$$

where integer n = 1,2, and 3. Group delay $T_d(\omega)$ is defined as $T_d(\omega) = -\frac{\partial\theta}{\partial\omega}$ where $\theta(\omega)$ is the phase component of the filter transfer function. The transfer function of (6.6) is implemented with

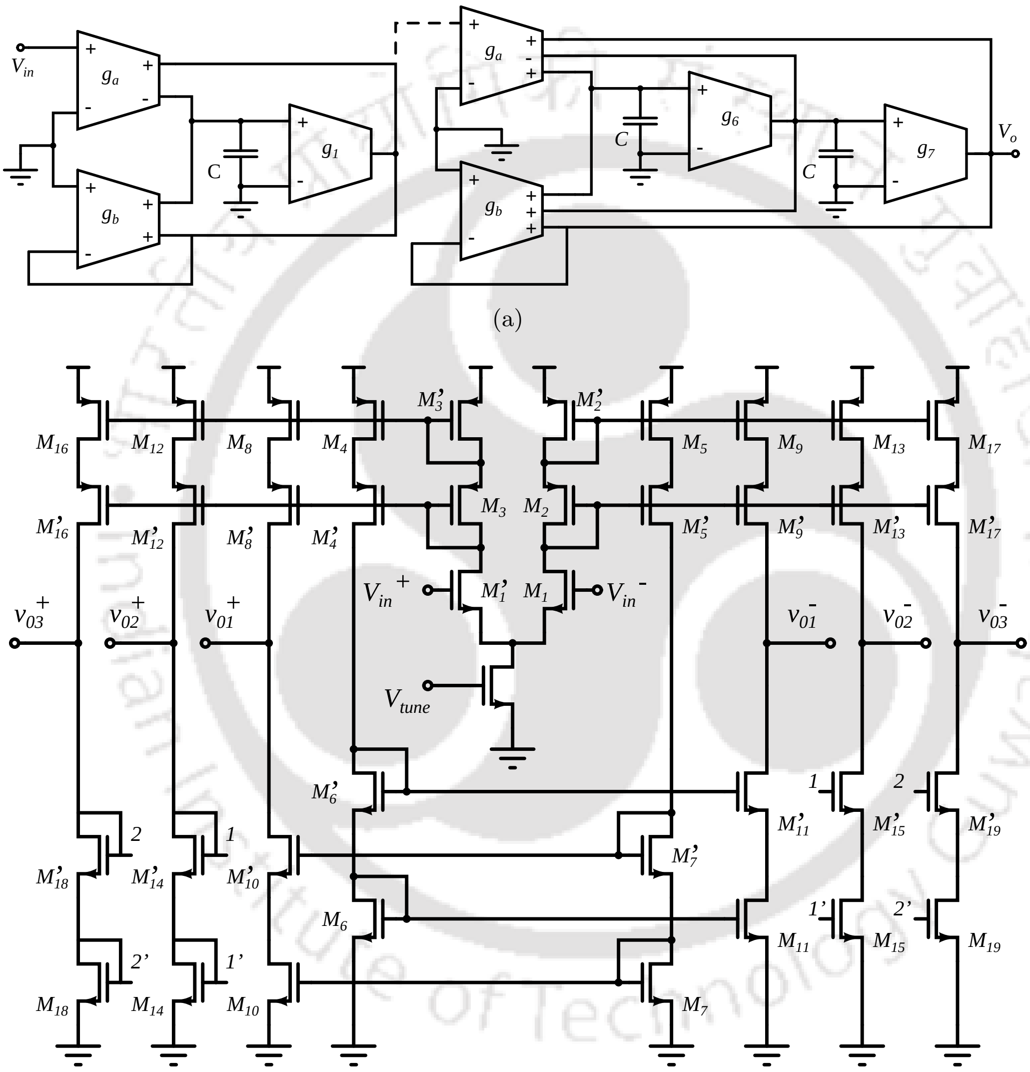


Figure 6.5: (a) Cascaded seventh-order all-pass filter. (b) A triple-output schematic of symmetric OTA with V_{tune} is used to implement the filter blocks.

three OTAs and a single capacitor C. For implementing the polynomial of (6.7), four OTAs and two capacitors of same value C is required. The seventh-order continuous-time OTA-C filter based on cascaded topology [124] is shown in Figure 6.5(a). The filter section employs OTAs with a single-ended input and grounded capacitors minimizing the effect of parasitic poles that cause excess phase shift.

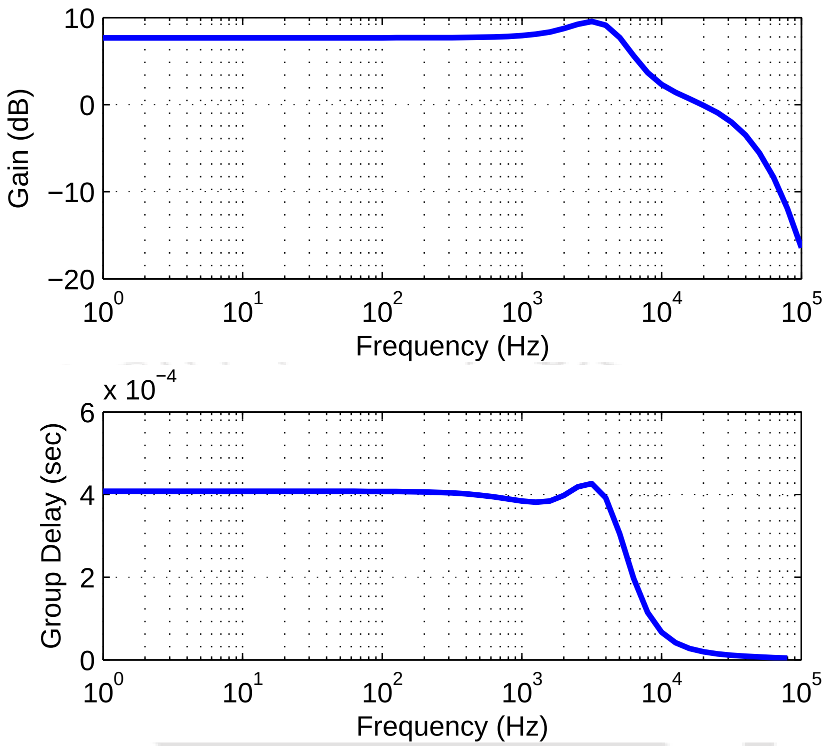


Figure 6.6: Frequency response and corresponding group delay of seventh-order all pass filter.

Grounded capacitors absorb parasitic capacitance and require less area as compared to floating once.

Figure 6.5(b) depicts the schematic of symmetrical OTA used to realize all-pass filter sections. The three outputs at v_{01}, v_{02}, and v_{03} have different transconductances tuned by sizing respective W/L ratios of cascoded transistors. Cascoded output stage is chosen to increase the output impedance of OTA. The transistors with small geometry are used to reduce parasitic capacitance and input referred noise. Figure 6.6 shows the group delay and frequency response of the 7th order filter. Transconductances of OTA were tuned to give a linear phase response yielding a constant delay bandwidth of up to 5.4 KHz with an overall phase delay of 410 $\pm$ 5% μsec while consuming a power of 1.5 μW. This delay bandwidth (300 Hz to 5.6 KHz) contains most of the neural energy hence is suitable for the present application. The filter offers a slight amplification however the waveform is not distorted which is evident from the delayed response shown in Figure 5.3(c). The filter design provides an adjustable phase shift by tuning the bias current of transconductors g_k and g_{k+1} ($k = 2$, 4, and 6) with a delay bandwidth trade-off.

To evaluate the filter performance we note that the design of symmetrical OTA is crucial in minimizing the total input referred noise of the filter. The input-referred thermal-noise voltage, power spectral density of the OTA can be evaluated as

$$v_{n,OTA}^2(f) = \frac{16kT}{3g_{m1}}\left(1 + 5\frac{g_{m2}}{g_{m1}} + 2\frac{g_{m6}}{g_{m1}} + 2\frac{g_{m7}}{g_{m1}}\right) \tag{6.8}$$

For the filter operating at biopotential frequencies, noise contribution of the cascode transistors can be neglected. The input referred noise voltage of this OTA can be minimized by increasing the transconductance of input transistors M_1 and M_1' such that $g_{m1} \gg g_{m2}$, g_{m6}, and g_{m7}. These transistors should be sized so as to have a maximum transconductance efficiency (g_m/I_D) while the non-input devices are designed to operate in strong inversion where g_m/I_D ratio is minimum. The minimum signal-to-noise ratio for an error free spike detection can be derived as

$$SNR_{\min} = 20\log\left(\frac{2 \times EAP_{\max}}{EAP_{min}}\right) \tag{6.9}$$

where EAP_{max} and EAP_{min} are representations of the value of maximum and minimum amplitude of extracellular action potential, respectively. According to the specifications of neural signals recorded extracellularly minimum and maximum signal levels are 30 μV and 200 μV, respectively in amplitude. This gives an minimum SNR of 22.5 dB above which the filter processes the neural signal without distortion. From (6.9), the minimum acceptable input referred noise, $v_{n,rms}$ for a given input signal $v_{in,rms}$ can be derived as

$$v_{n,rms} = \frac{v_{in,rms}}{10^{SNR(dB)/20}} \tag{6.10}$$

For a signal amplitude of 200 mV (p-p) at the filter input and a SNR of 22.5 dB, the maximum allowed noise referred to the filter input should be 5.3 mV$_{rms}$. The RMS noise voltage of filter was found to be in μV range which is below the safety margin value as calculated from (6.10). It denotes the filter noise to be insignificant as compared to in-band system noise, hence design constraint on the input referred noise of delay filter can be relaxed.

6.2 Validation of Method using Synthetic Data

6.2.1 Test Data

For the simulation of the proposed spike detector and comparison with other spike detection methods synthetic neural signal constructed from the real recordings were used. Synthetic neural signals are preferred over recorded extracellular signals since it provides ground truth about timestamps of spikes. Synthetic signals also allow for the noise amplitude and spike firing rate to be varied accord-

Table 6.1: Power Consumption Of Blocks

Block	Power dissipation (nW)	Percentage
Differentiator	29.55	0.58
Multiplier	608.5	11.87
Comparator ($\times 2$)	1938.8	37.81
Window generator	749.4	14.62
$g_m - C$ filter	0.29	0.006
Gain stage	0.031	0.0006
Delay filter	1800.3	35.12

ing to the test requirement. The dataset was obtained by superimposing the real neural spikes on to a noisy background by the method described in [125] and is available from the public database (http://nit.felk.cvut.cz/∼wildj1/ssc). The level of background noise is measured by its standard deviation with reference to the spike amplitude being normalized to one. For a different noise level and uniform firing rate distribution, a total of 5 traces are considered according to the noise standard deviations of 0.1, 0.15, 0.2, 0.3, and 0.4. Later on, each trace is synthesized with firing rates varying from 0 Hz up to 100 Hz in accordance with realistic neural recordings.

6.2.2 Hardware Performance

The designed circuit operating at 0.8 V supply was simulated in 0.18 μm CMOS process using Cadence Spectre simulator. The layout of the adaptive detector circuit integrated in UMC 0.18-CMOS process is shown in Figure 6.7. The overall circuit fits within an effective area of 150 $\times$ 119 μm^2 (excluding pads and routing details). This constitutes only 9.9% of the 400 $\times$ 400 μm^2 area allotment per channel [64] which is typically allocated for interfacing circuits in multichannel devices to fit within the pitch of a microelectrode array (MEA). The chip area is dominated by comparator, delay filter, and window generator circuits which occupies approximately 81% of the total area. Table 6.1 shows the individual power consumption of the building blocks of the proposed spike detector. The delay circuit consumes 35% of total power while occupying 11% of the silicon area. The whole spike detector dissipates 5.1 μW and occupies 0.018 mm^2 per channel.

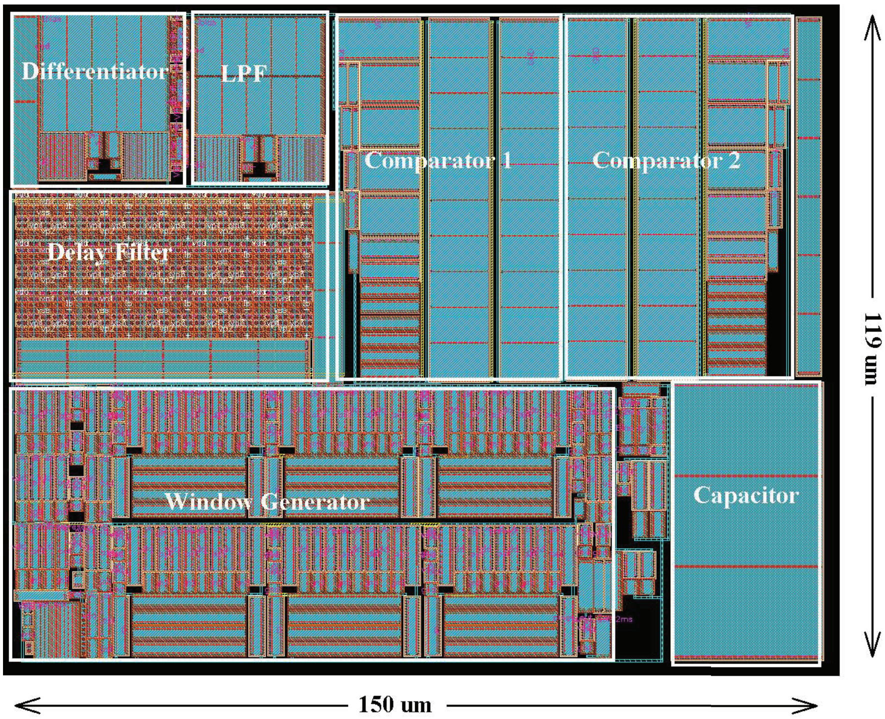


Figure 6.7: The layout of the proposed spike detector circuit.

6.2.3 Detection Performance

The performance of the noise estimation accuracy of our algorithm can be determined by subjecting it to synthesized waveform at different threshold level with varying firing rates. In our experiment with synthesized signals, we define SNR as the ratio of peak-to-peak amplitude of mean spike waveform, $V_{s,pp}$ and twice the standard deviation of noise, $V_{n,std}$ in accordance with [126]. It may be mentioned here that the background noise is the RMS value of the neural and other noise signals. The equation for SNR is given as

$$\mathrm{SNR} = \frac{V_{s,pp}}{2 \times V_{n,std}} \tag{6.11}$$

The synthetic data stream having maximum firing rate of 100 Hz is applied as the test input. In Figure 6.8, the effect of firing rate on the background estimate for three different standard deviations of noise are presented. The simulation results are plotted by normalizing the noise estimate with respect to the true background noise obtained at zero firing rate. Each subplot shows the background noise estimates at two different low pass corner frequencies. It may be mentioned that the tunable corner frequency set in the integrated local averaging filter is denoted by f_n as given in (6.2). It is observed that for a low noise standard deviation of 0.1, the background estimate, at conventional frequency of

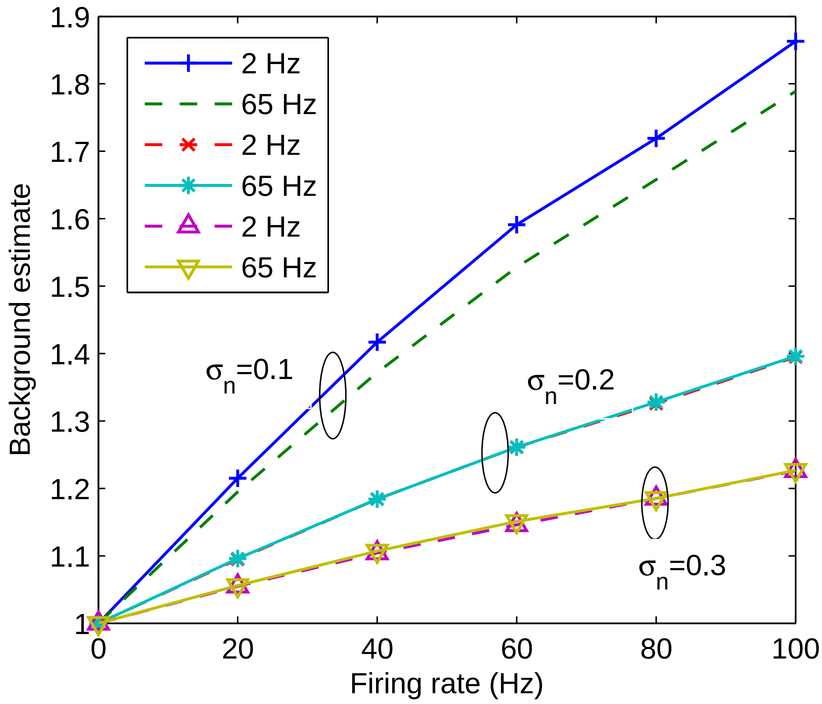


Figure 6.8: Background noise estimate based on different values of f_n with increasing firing rate for noise standard deviation of 0.1, 0.2, and 0.3. The noise estimate is apparently insensitive to the variations in f_n especially under high noise conditions.

f_n being 2 Hz, increases monotonically with the firing rate. When the corner frequency reaches f_n = 65 Hz, comparatively lesser variations are observed above the normalized background noise. This estimation error is further decreased at the moderate noise level of 0.2. At higher noise level of 0.3, the background estimate is apparently independent of the variation of f_n and remains constant within 20% of the true background estimate up to the firing rate of 100 Hz. Setting the cut-off frequency above 65 Hz makes the background estimation noisy. These observations demonstrate weak dependence of the proposed technique to the variations in filter frequency in high noise recording environment and its robustness to adapt for higher firing rates.

The relative performance of spike detection methods are evaluated using false positive ratio (FPR) and false negative ratio (FNR).

$$\text{FPR} = \frac{\text{number of false alarms}}{\text{total number of detections}} \tag{6.12}$$

$$\text{FNR} = \frac{\text{number of missed spikes}}{\text{total number of actual spikes}} \tag{6.13}$$

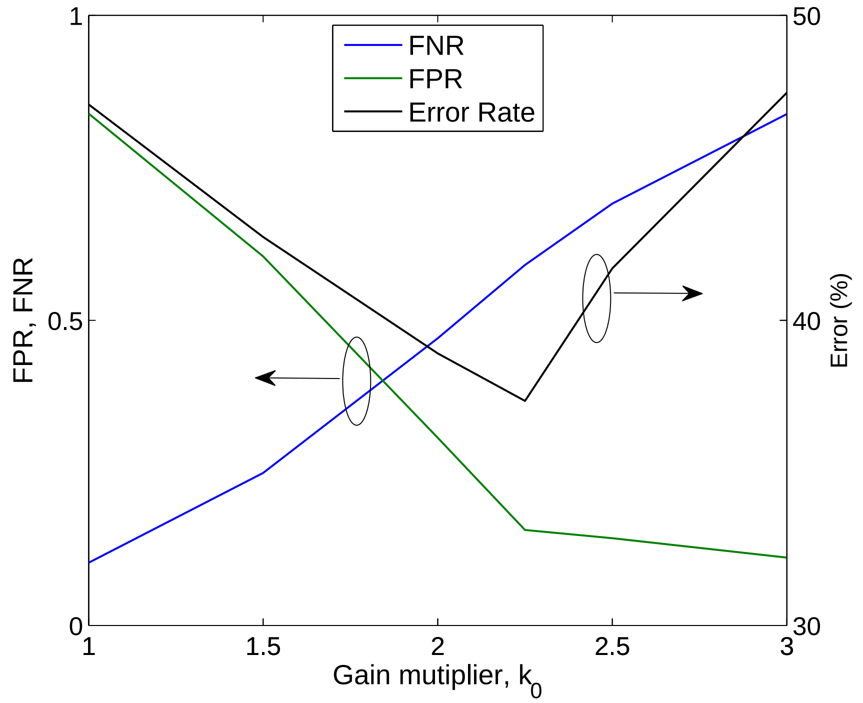
To minimize the erroneous inclusion of noise and maximize the detection of true spikes we find an optimum value of threshold level. The threshold scaling parameter k_0 determines the ratio of threshold voltage to the background noise level. If the value of k_0 is set to be very small then there will be more number of false detections whereas for a larger value of k_0, the number of missed spike may increase. To decrease the number of false detections, we attempt to find an optimum value of k_0, denoted as k_{opt}. The impact of the correctly and falsely detected spikes on detection performance was analyzed. The simulation results for the false positives and false negative detections versus positive threshold amplitude per σ_n were noted. This was done under worst case noise scenario. The detection error was plotted by sweeping the threshold value by varying k_0 ($\approx 1-3$). To quantify the total detection error, we define the error rate as the average of the number of false positive noise detections and the number of false negative misses.

$$\text{Detection Error} = \frac{\text{FPR} + \text{FNR}}{2} \times 100 \tag{6.14}$$

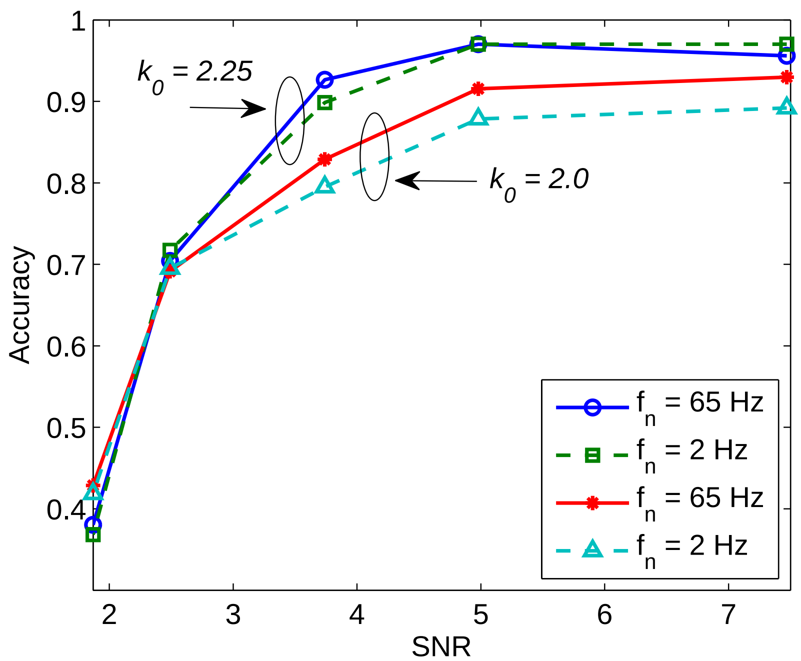
Figure 6.9(a) shows the average error rate evaluated in percentage derived from the false negative rate and false positive error over different values of k_0. The total error rate reaches a minimum value of k_0 which represents an optimal multiplier for evaluating the performance of our spike detector. It is observed that this trend is consistent for all the noise standard deviations. To demonstrate the validity of these results the detection accuracy was evaluated for SNR values varying from 2 to 7. The accuracy of the proposed algorithm is defined following [117] and is given in terms of correctly and falsely detected spikes as

$$\text{Accuracy} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{6.15}$$

where TP is the number of true positives, FN is false negative due to the number of missed spikes, and FP is number of false positives due to detection of noise as spikes. As depicted in Figure 6.9(b), our algorithm achieved the detection accuracy of more than 90 % for SNR level greater than 5.0 when k_0 was set to the value of 2.25. For this value of k_0, the accuracy performance remained insensitive to cut-off even when f_n was increased from 2 Hz to 65 Hz for the entire range of SNR. For $k_0 = 2.0$, the accuracy was sensitive to filter cut-off frequency. It was observed that he performance decreases by 2 % at LPF cut-off frequency of 65 Hz for higher SNR values. While, as the cut-off was decreased to $f_n =$ 2 Hz the accuracy degraded by 8 % for the same range of SNR. However, as a performance trade-off,



(a)



(b)

Figure 6.9: (a) Percentage error deduced from FPR and FNR as a function of gain multiplier, k_0. (b) Illustration of the accuracy of SET method at LPF cut-off frequency of $f_n = 2$ Hz and 65 Hz for gain multiplier of $k_0 = 2.25$ and 2.0.

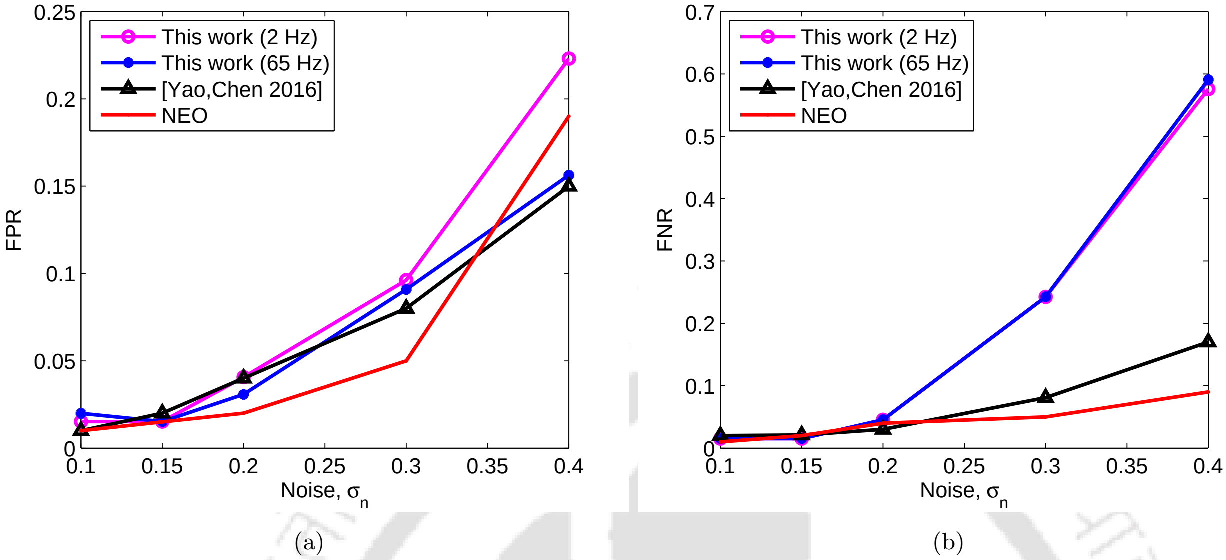


Figure 6.10: Performance comparison with the reported NEO based methods (a) False positive ratio with noise level. (b) Missed spike rate with noise level.

accuracy showed slight improvement towards lower SNR values. Similar results were observed for the values of k_0 other than 2.25. Hence, for performance comparison with the given data set, we choose $k_{opt} = 2.25$ as an optimum value of scaling parameter, so that $v_{Th} = k_{opt} \times v_{n,std}$.

The detection performance achieved with the present technique has been compared at the frequency of $f_n = 2$ Hz and 65 Hz with implementations of a similar reported ED based spike detector (SPD) circuit by Yao *et al.* [1] and a conventional NEO thresholding method reported in [119]. Figure 6.10 shows the performance comparison of our spike detector with different noise standard deviations, σ_n being normalized to the spike amplitude. For false positive error, it can be noted from Figure 6.10(a) that the detection performance at $f_n = 65$ Hz is better than the performance at $f_n = 2$ Hz especially at high noise standard deviation indicating that at lower SNR level, a higher cut-off of $f_n = 65$ Hz is able to better track the background noise variation. The false positive detection rate is found to be even better as compared to Yao's work at lower noise level. For false negative rate the number of spike misses increased significantly with the increase in noise level above 0.2 as shown in Figure 6.10(b). A possible reason is because the detection method considered here is based on single positive threshold determination. This results in more number of spikes being missed for events where the negative phase of action potentials have a higher SNR value than the positive phase. For the cut-off frequency of $f_n = 2$ Hz, the FNR performance of our method is marginally better, since the threshold estimate

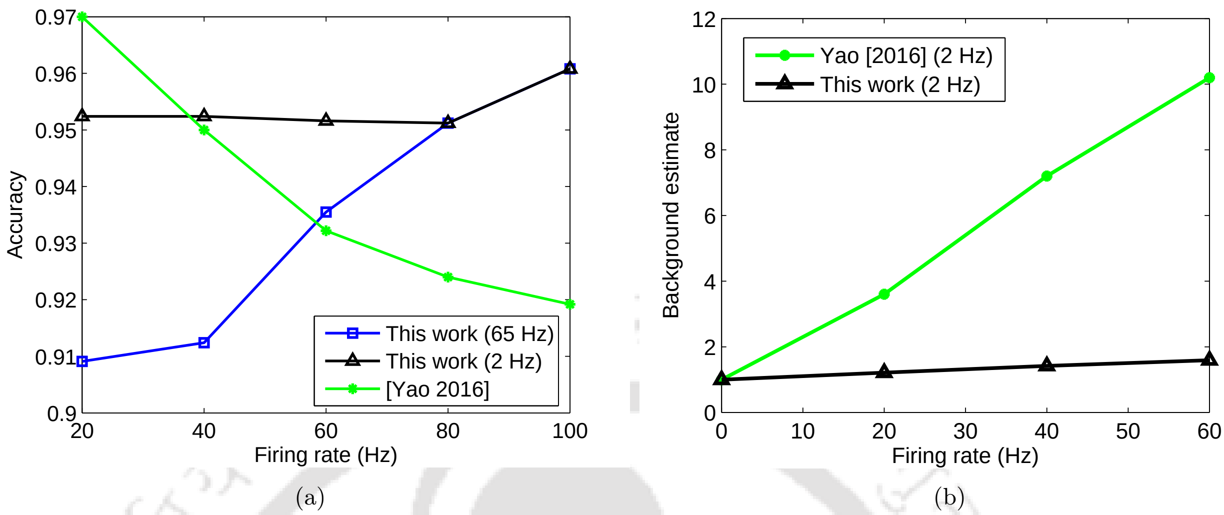
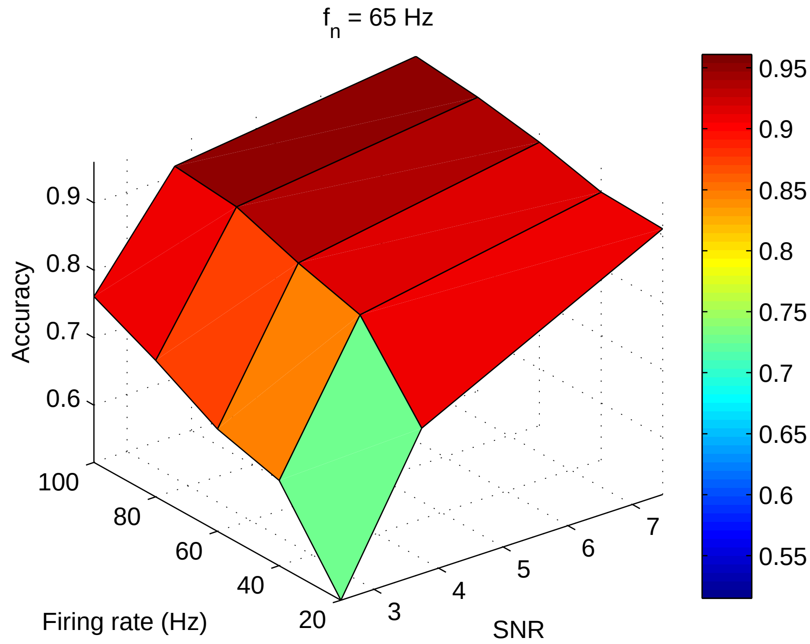


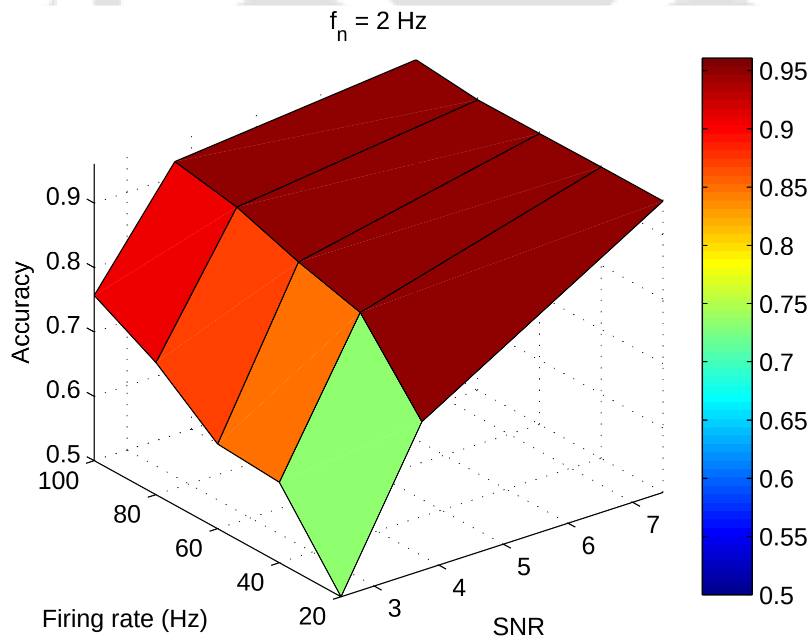
Figure 6.11: (a) Spike detection accuracy of the proposed method compared with SPD based technique [1] at different firing rates. (b) A comparison of background estimation of SET based method with Yao's method for the case of $\sigma_n = 0.1$. The proposed method shows only slight dependence and stays almost constant with increasing firing rates.

was less noisy and as a result there were less probability of spike misses. We have also compared with the supervised NEO algorithm as a reference method. The conventional NEO thresholding performed better for both FPR and FNR as there were less number of false detections even at higher noise levels. However, this NEO-based method does not provide adaptive threshold estimation.

To show robustness against the firing rate, we have evaluated the accuracy performance of SET thresholding method and also compared with the simulations of latest reported SPD based spike detector proposed by Yao in [1]. The detection accuracy was measured for the noise standard deviation of 0.1 with average firing rates varying from 20 to 100 Hz which is presented in Figure 6.11(a). The accuracy of the proposed method reaches more than 95% for firing rate greater than 80 Hz. For $f_n =$ 65 Hz, as the firing rate decreases from 80 Hz to 20 Hz, the accuracy of SET method falls to 91% while the accuracy for $f_n = 2$ Hz stays above 95 % even for lower range of firing rates. The reason for this is that the true positives decrease when the firing rate is low. However, at higher cutoff frequency, the number of false negatives comparatively increase causing accuracy to decrease at lower firing rates. For firing rates above 80 Hz, the true positives for both the cut-off frequencies increase significantly as compared to false detections. Hence, the accuracy improves with the firing rate and tends to approach unity according to (6.15). At these firing rates, the accuracy performance for both the curves apparently overlaps and shows increasing trend well above 100 Hz. The thresholding



(a)



(b)

Figure 6.12: 3D representation of detection accuracy of the proposed algorithm with different noise level and firing rate for (a) $f_n = 65$ Hz (b) $f_n = 2$ Hz. SNR varies from 3 to 7 while firing rate from 20 to 100 Hz.

Table 6.2: Design Comparison With Reported Spike Detectors

Parameters	[27]	[122]	[127]	[120]	[114]	This work
Technology (nm@V)	180 @1.8	180 @1.8	130 @1.2	130 @1.2	180 @±0.9	180 @0.8
Area/Ch. (mm^2)	0.03	0.07	0.41	0.021	N/A	0.018
Power/Ch. (μW)	1.5	0.78	85	0.05	47	5.1
Feature	NEO	NEO	EC-PC	NEO	N/A	ED
Adaptive threshold	Yes	No	No	Yes	Yes	Yes
Signal integrity	Poor	Complete	Poor	Poor	Poor	Complete
Domain	Analog	Analog	Analog	Digital	Analog	Analog

method of [1] provides better detection accuracy performance at lower firing rates. But there were more number of missed spikes as the firing rate increased. The improvement in performance of the proposed method can be quantified by comparing the background estimation of two algorithms for $\sigma_n = 0.1$. As depicted in Figure 6.11(b), our method shows comparatively less dependence on firing rates while the background estimate of Yao's method deviates by more than a factor of 10 at the firing rate of 60 Hz.

Figure 6.12 shows detection accuracy of our algorithm for different SNR and firing rate pairs at the LPF cut-off frequency of 65 Hz and 2 Hz. It is observed from both the figures that the performance shows consistent improvement for higher firing rates. The algorithm achieves overall accuracy of an average of 85 % for SNR values ranging from 4 to 7. The accuracy for $f_n = 2$ Hz in Figure 6.12(b) decreases rapidly for SNR < 3 while the performance in Figure 6.12(a) is slightly less sensitive to noise level. It suggests that the detection performance at $f_n = 2$ Hz is a better choice for detecting neural signals with densely distributed spikes as it shows minimum dependence on firing rate activity. At 65 Hz cut-off, the detection performance decreases only slightly at lower SNR levels indicating that SET method is more robust to background noise variations. We have also measured the dependence of threshold estimate on the rate of AP depolarization preceding each spike. We found that there was no observable effect on detection accuracy for interspike interval greater than 1 ms.

In Table 6.2, design comparison of this work with NEO based architectures [27, 120, 122] and a recently reported exponential component-polynomial component (EC-PC) based method [127] is presented. In terms of chip area, the present design while operating at a low supply voltage of 0.8

V occupies a small silicon area per channel. Overall, the proposed spike detector presents a high detection performance having a moderate power consumption with an area efficient design approach. The performance of the proposed design is at par with the existing design work, hence the design finds suitability for multichannel implantable neural data acquisition systems.

6.3 Conclusion

A circuit implementation of real-time adaptive threshold detection algorithm is presented. The detector is able to successfully detect spikes under high spike firing rate conditions. The detection strategy preserves neural waveform integrity with a linear phase delay filter. Our results indicate that the detector performance is insensitive to the choice of cut-off frequency especially at low SNR conditions. The detector also maintains the accuracy of background noise estimate for spike firing rate of more than 100 Hz. The proposed algorithm has been implemented in a compact analog CMOS integrated circuit. The total area occupied by the circuit is 0.018 mm^2, so it is suited for array implementation. The power consumption per channel remains within the tolerable limit of 800 W/mm^2 [29]. This provides opportunity for the technique to be adopted in multichannel implantable systems. An integrated multichannel device using such detector would also have room for the inclusion of other circuits (amplifiers, filters, etc.). Potential applications of the proposed design are neural prosthetics systems for brain-machine interface and low power online seizure detection and prediction for implantable devices.

7

Summary and Conclusions

Contents

7.1 Summary of Contributions and Discussions

The basic objective of this dissertation is to present a new method which can automatically access the threshold level for a real-time spike detection system. Any on-chip spike detecting algorithm must deal with the fact that spike firing rate along with amplitudes and shapes change with time with a slight movement of electrode array in the tissue [128]. Our objective is therefore, to reduce the detection errors caused by these changes on the performance of the spike detector. Furthermore, spike detection circuits designed for neural implants are heavily dictated by size and power limitations. Several software-based spike detection techniques developed during the past few decades are quite computationally intensive and power consuming [17,113,129]. In this work, an adaptive algorithm for unsupervised spike detection has been proposed which is simple to realize and whose circuit implementation occupies small area. CMOS circuits are designed to calculate the standard deviation such that the threshold value is adaptive to the noise level. The algorithm is robust to the spike firing rate, their amplitude and shape. Low power, area efficient analog circuits to implement the algorithm have also been discussed. The proposed algorithm is evaluated using synthetic neural signals based on real extracellular recordings. The test results show better performance with other known threshold-based spike detection methods. The hardware feasibility is demonstrated with a low cost and an area-efficient circuit implementation for use in multi-channel neural recording implants for BMI applications.

Since the real extracellular recordings are in μV range while the signals required for spike detection should be in mV range, therefore a neural amplifier is also designed to pre-condition the spike level from few μVs to 10s of mVs. The first part of the dissertation presents the design and analysis of an implantable neural amplifiers for recording weak potentials produced in the brain which includes high frequency spike and low frequency LFP signal. The later part of this design work proposes a band-programmable three-stage analog front-end (AFE) interface with the aim to reduce current supply and noise. It receives neural signals from microelectrodes, and delivers the preconditioned signal to the following block in the recording path, which is a spike detector circuit. Its main function is to separate spike and LFP signals which can reduce low frequency interference for a subsequent spike detection circuit. Cadence simulation and circuit analysis have proven that this circuit is suitable for the spike as well as LFP recording.

The first stage of AFE interface describes the design of a capacitively-coupled neural amplifier. To prove its design feasibility, two fully-integrated CMOS first-stage amplifiers for LFP recording with mid-band gain of around 37 dB and 25 dB have been demonstrated. The first prototype LFP amplifier based on fixed pseudo resistive-feedback version has the bandwidth ranging from 4 Hz to 210 Hz. The second prototype circuit is also an LFP amplifier which is an improvement over the first one. A biasing technique with a pseudo resistive feedback was utilized in the design of the neural amplifier to help minimize $1/f$ noise and reject the DC offset and polarization effects introduced by the measuring electrodes. The modified pseudo resistive-feedback version can achieve the high-pass cut-off frequency of 3 mHz Hz with a passband of 340 Hz whilst operating in the reduced supply voltage of 1.5 V. The experiments carried out endorse the effectiveness of the proposed scheme. These properties make it practical to cascade them to the post-amplifier (second and third stage of AFE) designed in section 4.3.2 to form a total 55 dB neural amplifier. Both amplifiers were designed with a 0.18 μm CMOS process. The quiescent power dissipation of the first-stage amplifiers is less than 10 μW under 1.5 V power supply.

7.2 Suggestions for Future Research

In this section we provide some possible directions for further research in this dissertation:

7.2.1 Performance improvement of Spike Detector

For further improving the performance, the Spike detector can be designed to support the detection of and biphasic spikes. To be more robust in the detection of biphasic spikes, two separate threshold levels can be used. An additional comparator circuit can be used: one for positive spikes and the other for negative spikes. Similar algorithm can be used for negative threshold thus creating the positive and negative thresholds symmetrically around the zero offset level. This approach would cover a wide range of input signal having both positive and negative phase of spikes. The present work is for a single positive threshold spike detection system. However, in practical systems, the major issue is the detection of biphasic spikes in neural signal for reducing the detection errors. The present work can be extended for detection of biphasic spike systems with the addition of single voltage comparator which may support detection of spikes with negative phase. The spike detector may include the core circuitry for calculating the upper and lower spike detection thresholds based on the standard deviation and mean of each neural channel as shown in Figure 7.1. The schematic can replace the comparator block in Figure 5.2 with $v_{in} = \psi(t)$. The calculated voltage v_{Th} act as a positive as well as negative threshold level which may be symmetrically set around the baseline of the neural signal. Having the

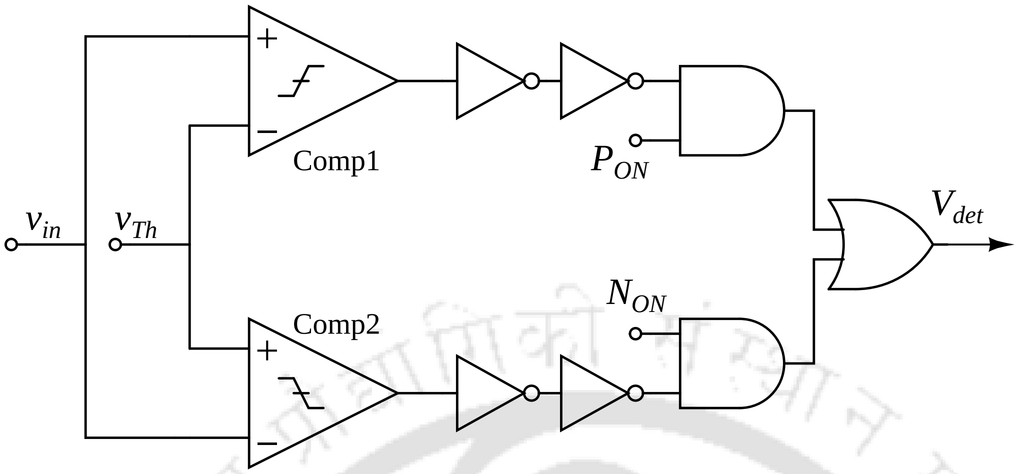


Figure 7.1: Proposed circuit schematic for positive and negative detection of spike.

flags $P_{ON} = N_{ON} = 1$ puts the spike detector in biphasic detection mode. It may then detect neural spikes by comparing the current sample to the previously calculated threshold. This would result in a two-fold performance improvement. However, this performance improvement will come at the cost of layout area as it requires addition of a comparator circuit.

7.2.2 Suppressing Noise Glitches in Spike Detection

The performance of spike detector largely depends upon the layout plan of the analog building blocks. Since the layout contains both the analog and digital signals therefore the floor plan of the spike detector chip should employ mixed-signal layout technique [27] to help minimize digital signals from disturbing the sensitive analog signals.

(i) The noisy clock generator should be placed far away from the input analog filter and preamplifier circuits. This helps to minimize the weak input signals from being disturbed by the switching noise of the clock generator.

(ii) A large decoupling capacitor should also be placed close to the preamplifier and analog filter so as to provide better on-chip supply decoupling for both circuits.

A

Voltage Threshold Comparator

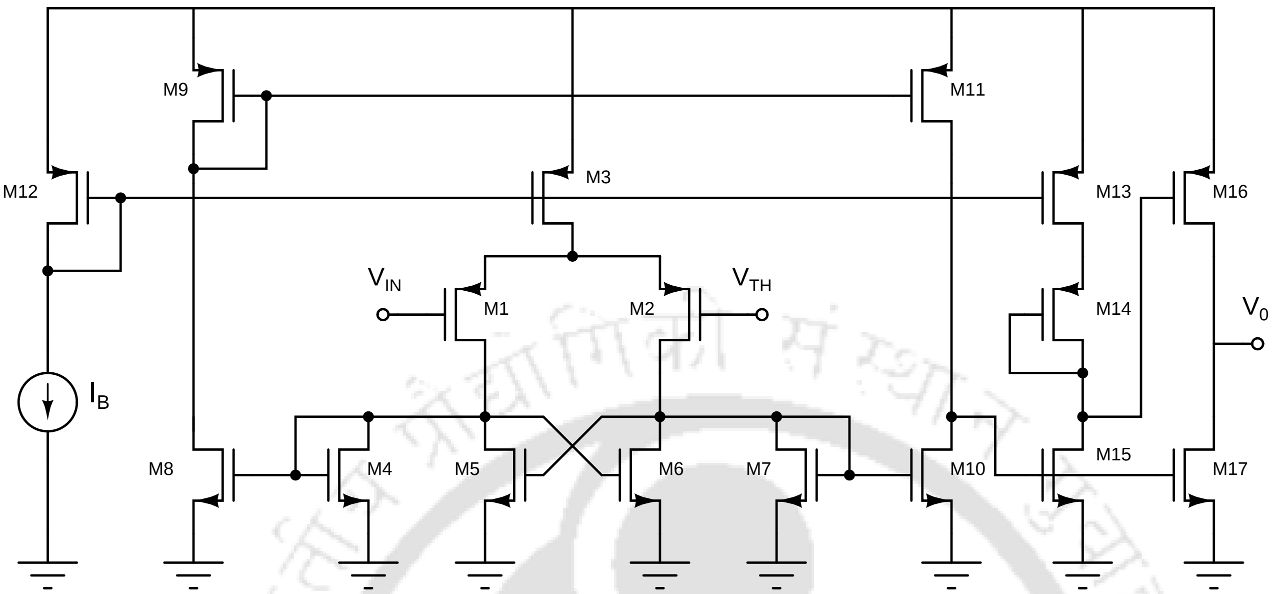


Figure A.1: Schematic of track and latch comparator.

Figure A.1 presents the threshold detection comparator design based on track and latch operation. The comparator determines the time at which the node V_{IN} equals the positive threshold voltage V_{TH}. It includes three stages – input differential stage, latch stage, and output buffer stage. The first stage consists of source-coupled differential pair M_l and M_2 that uses positive feedback to provide increased gain. This stage amplifies the input signal to improve the comparator sensitivity. It increases the minimum input signal with which the comparator can make a decision and isolates the input of the comparator from switching noise (often called kickback noise) coming from the positive feedback stage. This also serves to reduce the input referred latch offset voltage. The transconductance sets the gain of the stage, while the size of M_l and M_2 determines the input capacitance of the comparator where $g_{m1} = g_{m2}$. To maximize the unity gain frequency of the amplifier, the input devices use short gate lengths $W_1/L_1 = 6\mu m/0.7\mu m$. The tail current is nominally set to 1 μA. The positive feedback latch stage (M_{4-7}) is used to determine which of the input signals is larger and amplifies their difference. A controlled amount of positive feedback is used to effectively increase the input devices transconductance and hence the overall gain. The differential gain of positive feedback gain stage is given by

$$A_{FB} = \sqrt{\frac{\mu_n (W/L)_1}{\mu_p (W/L)_4}} \frac{1}{1-\alpha} \tag{A.1}$$

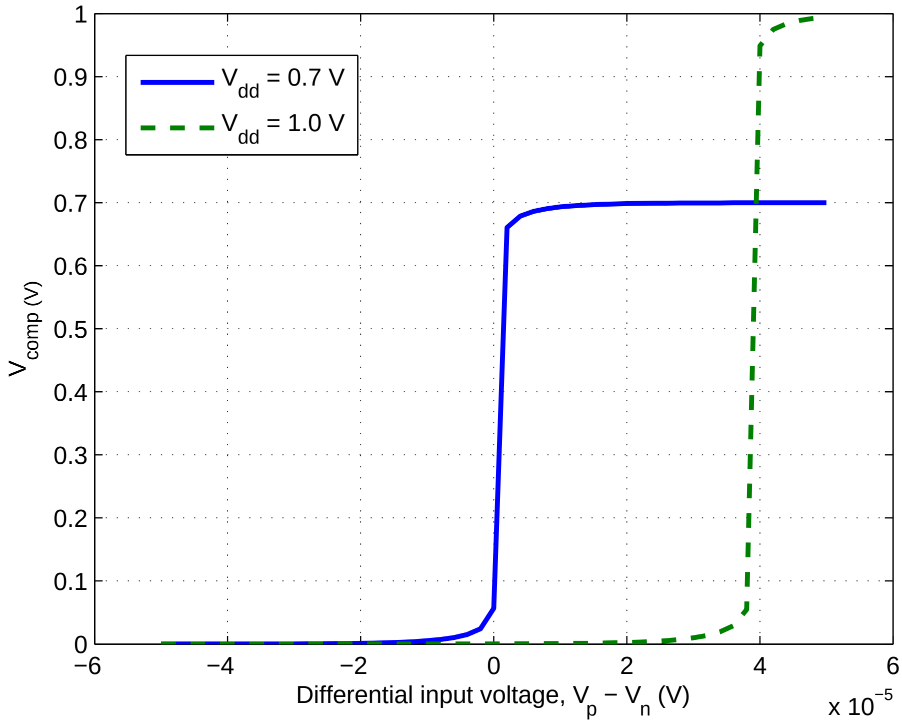


Figure A.2: DC sweep response of comparator.

where $\alpha = (W/L)_5/(W/L)_4$ is the positive feedback factor responsible for increasing the gain. A reasonable value of α is 0.75 which increases the gain by a factor of 4. The practical value of α is determined by the value of load device dimensions and is 0.9 due to mismatches because of the process variations. The single-ended output swing of the low-swing gain stage is roughly 500 mV. This dc level has to be shifted down with a diode-connected PMOS device to have a larger swing while still keeping the input devices in saturation.

The output buffer stage consists of a level converter followed by an inverter which gives the digital output. It converts the output of the latch stage to a full scale digital level output (logic 1 or logic 0). The inverter stage is of class AB type so that it should accept a differential input signal and not have slew-rate limitations. Class AB output stage also ensure low distortion. One advantage of the level converter shown is that because it is based on inverter, it has peak transition current that is significantly high than the static current (0.1 μA) in the buffer stage. This behavior allows for fast rise/fall times resulting in lower jitter sensitivity.

Input offset voltage originates from preamplifier and the latch. It results from the transistor mismatches such as threshold voltage V_{TH}, internal node capacitances and output load capacitance

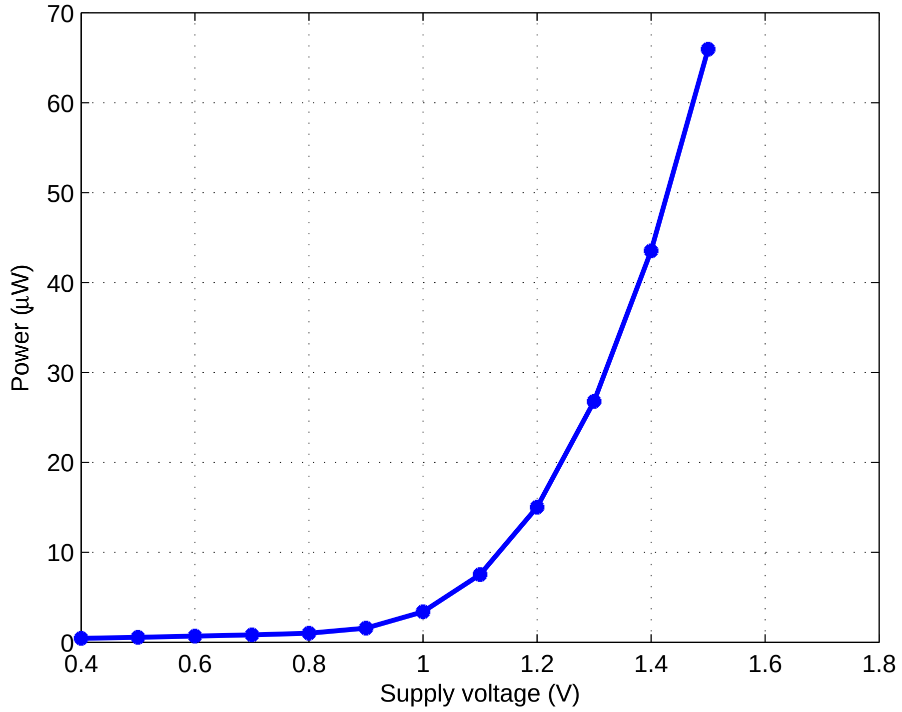


Figure A.3: Simulated power performance.

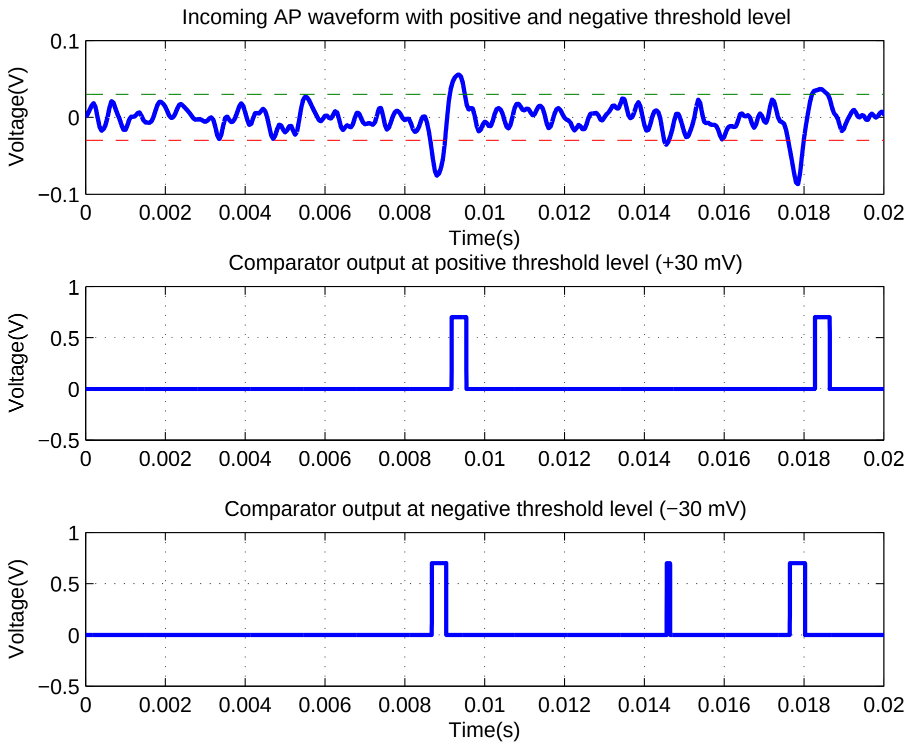


Figure A.4: Transient response of comparator to AP input for positive and negative threshold detection.

variations, which deteriorates the accuracy of these comparators. Offset voltage is simulated by sweeping the input signal difference ($V_p - V_n$) for two supply voltage values of 0.7 V and 1.0 V as shown in Figure A.2. The measured difference value were noted as 0.98 μV and 39.01 μV respectively at which the output of the comparator reaches to mid supply. The low values of offset voltages were due to a large preamplifier gain which trades with the speed of comparator. The measured propagation delay was 1.2 μsec which is sufficient for measuring the changing levels of fast varying neural background noise. Figure A.3 shows an exponential rise in power dissipated by comparator as the supply voltage increases. The simulated power performance demonstrates that the comparator can yield the desired output over a wide range of supply voltage values (e.g., 0.4 V to more than 1.5 V). It consumes 821.7 nW at a near threshold supply of 0.7 V while a power of 1.001 μW at 0.8 V. However, the comparator is operated at an optimal voltage of 1 V (3.405 μW) so as to maintain the required bandwidth and to avoid distorted transient waveform.

The simulated transient response of the comparator for two different settings of threshold is shown in Figure A.4. The incoming action potential waveform is applied to the input of the comparator with a near threshold of 0.7 V supply, which is compared with the positive and negative detection threshold voltage of $\pm$30 mV. This gives a digitized output of 0.7 V peak pulse value as the input crosses the threshold mark.

B

Analog Multiplier

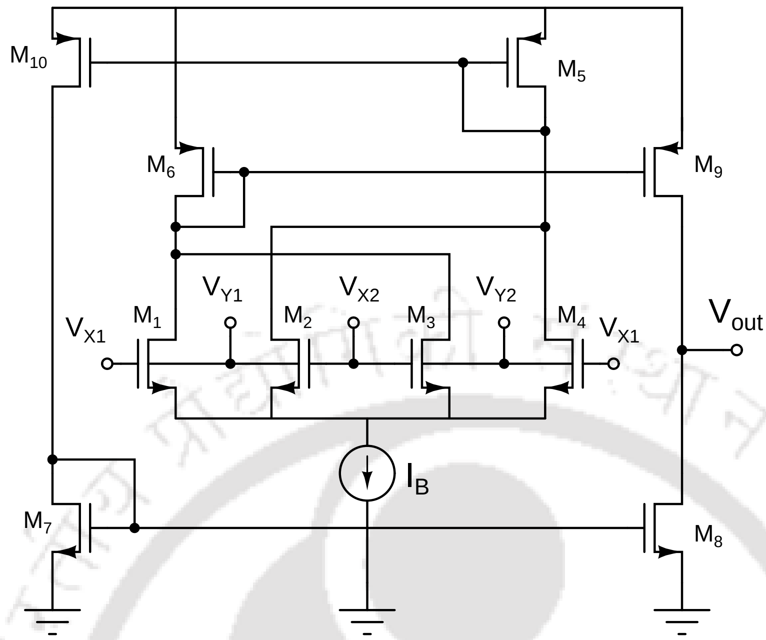


Figure B.1: CMOS implementation of a four quadrant Gilbert cell multiplier with bulk input.

The core of differential multiplier consists of two cross-coupled N-type differential pairs as shown in Figure B.1. The analog multiplication is achieved by driving the gate and bulk terminals of the device based on the design suggested in [123]. All the four input transistors are operating in subthreshold region. For normal operation of MOS transistors $M_{1,2,3,4}$ source-bulk and drain-bulk junction should be reverse-biased and $V_X = v_{x1} - v_{x2} \leqslant V_T/\kappa$ and $V_Y = v_{y1} - v_{y2} \leqslant V_T/(1-\kappa)$. By using above conditions, the output current i_0 can be shown to be the product of two inputs.

$$i_0 = \frac{I_B.\kappa^2}{4.V_T^2}.v_X.v_Y \tag{B.1}$$

The factor κ is the reciprocal of the slope factor n ($\kappa \approx 0.6$ in bulk CMOS processes). V_T is the thermal voltage. MOS multiplier working in sub threshold region has the advantage that the current levels are typically orders of magnitude lower than devices biased above threshold. This allows the present design methodology to achieve a very low power dissipation of 0.401 μW while operating on a 1 V power supply. The design presents an approximately 700 mV of linear range as depicted in Figure B.2 and can be further enhanced by using linearization techniques at subthreshold [32]. The linear range can be extended by various methods viz. source degeneration, gate degeneration and bump linearization. In these multipliers transconductance is directly proportional to the drain current. However these linearization techniques were not adopted in the present work since the biopotential

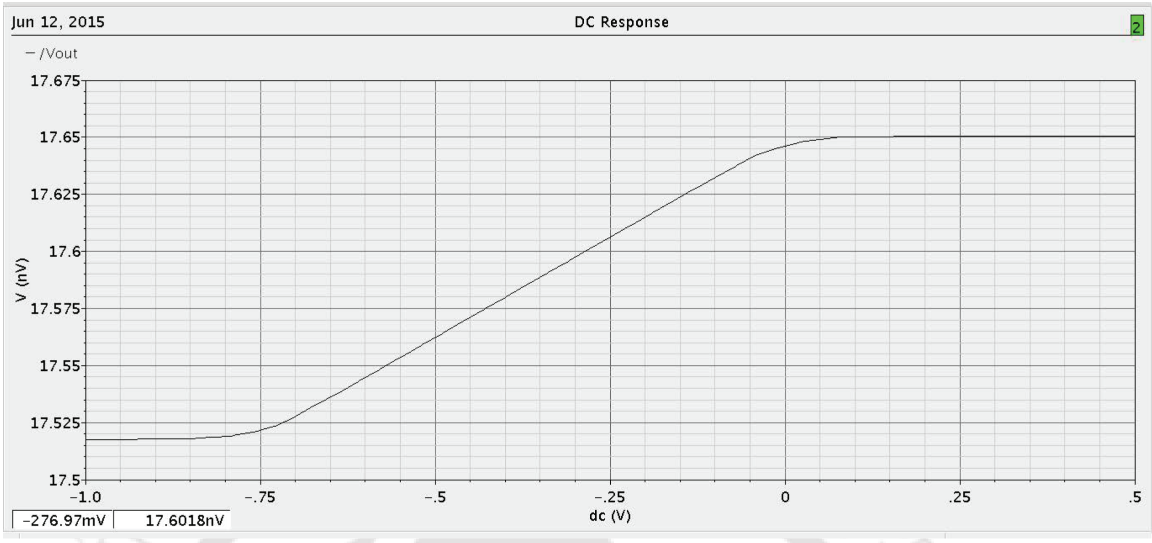
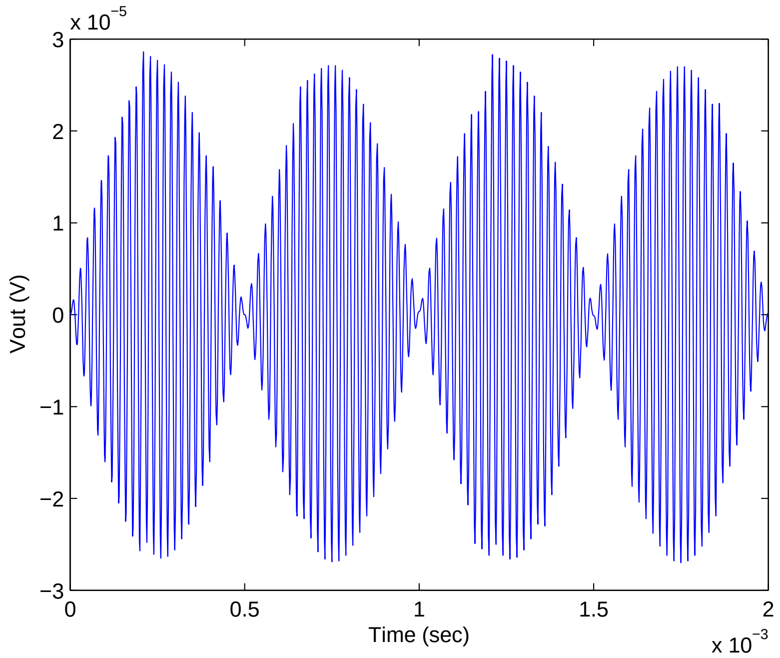


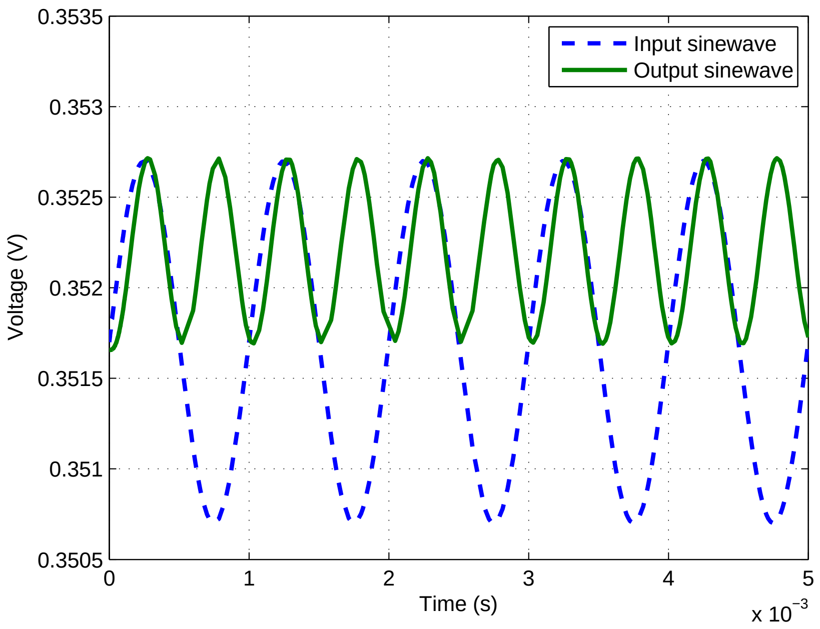
Figure B.2: Simulated DC transfer function.

signal being dealt here lies well within 200 mV range and hence our multiplier characteristics suffices for the same.

Figure B.3(a) shows a 2 *msec* of transient analysis of multiplier with a 50 KHz sinusoidal signal applied across differential input V_X and a 1KHz signal at the back gate differential input V_Y. The output waveform resembles a analog modulated waveform with a ±28 mV peak amplitude operating at 200 nA tail current. The analog multiplier can also be utilized for squaring the signal often useful for nonlinear processing in biomedical applications. Figure B.3(b) shows the simulated time response of the analog multiplier for a frequency doubling operation. The squaring operation can extract a measure of the energy of signal equivalent to the product of the square of the signal amplitude and that of the frequency [130]. The operation is however sensitive to noise since a clean signal can be corrupted by an additive zero mean Gaussian noise therefore a proper linear filter has to be inserted before applying squaring algorithm. The output of squarer is fed to a first-order low-pass $G_m - C$ filter. The filter is implemented with an OTA based on two asymmetric differential pair topology. The OTA is biased in weak-inversion region so as to choose high-pass corner frequency below 100 Hz. This attenuates any high-frequency oscillation emanating from the squarer block leaving only an integrated output. This is equivalent to the averaged-out DC waveform in real-time, thus adapting to time-varying level of noisy background voltage. The power consumption of the first order low pass filter is 434.3 pW when run on 1 V of power supply.



(a) Amplitude modulation.



(b) Frequency doubling operation.

Figure B.3: Simulated output waveform of multiplier as amplitude modulator and frequency doubler

Bibliography

[1] E. Yao, Y. Chen, and A. Basu, "A 0.7 v, 40 nw compact, current-mode neural spike detector in 65 nm cmos," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 2, pp. 309–318, April 2016.

[2] C. T. Nordhausen, E. M. Maynard, and R. A. Normann, "Single unit recording capabilities of a 100 microelectrode array," *Brain Research*, vol. 726, no. 1, pp. 129 – 140, 1996.

[3] G. Guitchounts, J. E. Markowitz, W. A. Liberti, and T. J. Gardner, "A carbon-fiber electrode array for long-term neural recording," *Journal of Neural Engineering*, vol. 10, no. 4, p. 046016, 2013.

[4] I. Fried, U. Rutishauser, M. Cerf, and G. Kreiman, *Single Neuron Studies of the Human Brain: Probing Cognition*. MIT Press, 2014.

[5] R. Brette and A. Destexhe, *Handbook of Neural Activity Measurement*. Cambridge Univ. Press, 2012.

[6] J. L. Collinger, B. Wodlinger, J. E. Downey, W. Wang, E. C. Tyler-Kabara, D. J. Weber, A. J. McMorland, M. Velliste, M. L. Boninger, and A. B. Schwartz, "High-performance neuroprosthetic control by an individual with tetraplegia," *The Lancet*, vol. 381, no. 9866, pp. 557–564, Feb. 2013. [Online]. Available: https://doi.org/10.1016/S0140-6736(12)61816-9

[7] B. Cessac, H. Paugam-Moisy, and T. Vieville, "Overview of facts and issues about neural coding by spikes," *Journal of Physiology-Paris*, vol. 104, no. 1, pp. 5 – 18, 2010, computational Neuroscience, from Multiple Levels to Multi-level. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0928425709000849

[8] K. A. Ludwig, J. D. Uram, J. Yang, D. C. Martin, and D. R. Kipke, "Chronic neural recordings using silicon microelectrode arrays electrochemically deposited with a poly(3,4-ethylenedioxythiophene) (pedot) film," *Journal of Neural Engineering*, vol. 3, no. 1, p. 59, 2006. [Online]. Available: http://stacks.iop.org/1741-2552/3/i=1/a=007

[9] K. Najafi, "Solid-state microsensors for cortical nerve recordings," *IEEE Engineering in Medicine and Biology Magazine*, vol. 13, no. 3, pp. 375–387, June 1994.

[10] A. Hierlemann, U. Frey, S. Hafizovic, and F. Heer, "Growing cells atop microelectronic chips: Interfacing electrogenic cells in vitro with cmos-based microelectrode arrays," *Proceedings of the IEEE*, vol. 99, no. 2, pp. 252–284, Feb 2011.

[11] R. Harrison, P. Watkins, R. Kier, R. Lovejoy, D. Black, B. Greger, and F. Solzbacher, "A low-power integrated circuit for a wireless 100-electrode neural recording system," *Solid-State Circuits, IEEE Journal of*, vol. 42, no. 1, pp. 123–133, Jan 2007.

[12] W. Wattanapanitch, M. Fee, and R. Sarpeshkar, "An energy-efficient micropower neural recording amplifier," *Biomedical Circuits and Systems, IEEE Transactions on*, vol. 1, no. 2, pp. 136–147, June 2007.

[13] W. Bialek, F. Rieke, R. de Ruyter van Steveninck, and D. Warland, "Reading a neural code," *Science*, vol. 252, no. 5014, pp. 1854–1857, 1991.

[14] R. R. Harrison, "The design of integrated circuits to observe brain activity," *Proceedings of the IEEE*, vol. 96, no. 7, pp. 1203–1216, July 2008.

[15] R. Olsson and K. Wise, "A three-dimensional neural recording microsystem with implantable data compression circuitry," *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 12, pp. 2796–2804, Dec 2005.

[16] R. R. Harrison, "A low-power integrated circuit for adaptive detection of action potentials in noisy signals," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439)*, vol. 4, Sept 2003, pp. 3325–3328 Vol.4.

[17] I. Obeid and P. Wolf, "Evaluation of spike-detection algorithms fora brain-machine interface application," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 6, pp. 905–911, June 2004.

[18] Y. Perelman and R. Ginosar, "An integrated system for multichannel neuronal recording with spike/lfp separation, integrated a/d conversion and threshold detection," *Biomedical Engineering, IEEE Transactions on*, vol. 54, no. 1, pp. 130–137, Jan 2007.

[19] K. Oweiss, K. Thomson, and D. Anderson, "A systems approach for real-time data compression in advanced brain-machine interfaces," in *Conference Proceedings. 2nd International IEEE EMBS Conference on Neural Engineering, 2005.*, March 2005, pp. 62–65.

[20] Z. S. Zumsteg, C. Kemere, S. O'Driscoll, G. Santhanam, R. E. Ahmed, K. V. Shenoy, and T. H. Meng, "Power feasibility of implantable digital spike sorting circuits for neural prosthetic systems," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 13, no. 3, pp. 272–279, Sept 2005.

[21] S. Kim, S. Han, J. Cha, L. Liu, L. Yao, Y. Gao, and M. Je, "A sub-μw/ch analog front-end for δ-neural recording with spike-driven data compression," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, no. 1, pp. 1–14, Feb 2019.

[22] M. Zamani, D. Jiang, and A. Demosthenous, "An adaptive neural spike processor with embedded active learning for improved unsupervised sorting accuracy," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, no. 3, pp. 665–676, June 2018.

[23] K. Guillory and R. Normann, "A 100-channel system for real time detection and storage of extracellular spike waveforms," *Journal of Neuroscience Methods*, vol. 91, no. 1, pp. 21 – 29, 1999.

[24] P. Watkins, R. Kier, R. Lovejoy, D. Black, and R. Harrison, "Signal amplification, detection and transmission in a wireless 100-electrode neural recording system," in *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, May 2006, pp. 4 pp.–2196.

[25] M. Steyaert and W. Sansen, "A micropower low-noise monolithic instrumentation amplifier for medical purposes," *Solid-State Circuits, IEEE Journal of*, vol. 22, no. 6, pp. 1163–1168, Dec 1987.

[26] J. Holleman, A. Mishra, C. Diorio, and B. Otis, "A micro-power neural spike detector and feature extractor in .13 um CMOS," in *2008 IEEE Custom Integrated Circuits Conference*, Sept 2008, pp. 333–336.

[27] E. Koutsos, S. E. Paraskevopoulou, and T. G. Constandinou, "A 1.5 uw neo-based spike detector with adaptive-threshold for calibration-free multichannel neural interfaces," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*, May 2013, pp. 1922–1925.

[28] A. Rodrguez-Prez, M. Delgado-Restituto, and F. Medeiro, "A 515 nw, 0-18 db programmable gain analog-to-digital converter for in-channel neural recording interfaces," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 8, no. 3, pp. 358–370, June 2014.

[29] D. Han, Y. Zheng, R. Rajkumar, G. Dawe, and M. Je, "A 0.45 V 100-channel neural-recording IC with sub-u W/channel consumption in 0.18 um CMOS," *Biomedical Circuits and Systems, IEEE Transactions on*, vol. 7, no. 6, pp. 735–746, Dec 2013.

[30] H. Tanimoto, M. Koyama, and Y. Yoshida, "Realization of a 1-V active filter using a linearization technique employing plurality of emitter-coupled pairs," *Solid-State Circuits, IEEE Journal of*, vol. 26, no. 7, pp. 937–945, 1991.

[31] K.-C. Kuo and A. Leuciuc, "A linear MOS transconductor using source degeneration and adaptive biasing," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 48, no. 10, pp. 937–943, 2001.

[32] A. Pesavento and C. Koch, "A wide linear range four quadrant multiplier in subthreshold CMOS," in *Circuits and Systems, 1999. ISCAS '99. Proceedings of the 1999 IEEE International Symposium on*, vol. 2, Jul 1999, pp. 240–243 vol.2.

[33] S. Dwivedi and A. Gogoi, "Improvement of linearity in a CMOS transconductor by employing a squaring circuit," in *Emerging Research Areas and 2013 International Conference on Microelectronics, Communications and Renewable Energy (AICERA/ICMiCR), 2013 Annual International Conference on*, June 2013, pp. 1–3.

[34] B. Gilbert, "The multi-tanh principle: a tutorial overview," *Solid-State Circuits, IEEE Journal of*, vol. 33, no. 1, pp. 2–17, 1998.

[35] E. Vittoz and J. Fellrath, "Cmos analog integrated circuits based on weak inversion operations," *IEEE Journal of Solid-State Circuits*, vol. 12, no. 3, pp. 224–231, Jun 1977.

[36] F. Silveira, D. Flandre, and P. G. A. Jespers, "A gm/id based methodology for the design of cmos analog circuits and its application to the synthesis of a silicon-on-insulator micropower ota," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 9, pp. 1314–1319, Sep 1996.

[37] D. M. Binkley, C. E. Hopper, S. D. Tucker, B. C. Moss, J. M. Rochelle, and D. P. Foty, "A cad methodology for optimizing transistor current and sizing in analog cmos design," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 22, no. 2, pp. 225–237, Feb 2003.

[38] Y. Tsividis and G. Masetti, "Problems in precision modeling of the mos transistor for analog applications," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 3, no. 1, pp. 72–79, January 1984.

[39] Y. Tsividis, "Moderate inversion in MOS devices," *Solid-State Electronics*, vol. 25, no. 11, pp. 1099 – 1104, 1982.

[40] C. C. McAndrew and P. A. Layman, "Mosfet effective channel length, threshold voltage, and series resistance determination by robust optimization," *IEEE Transactions on Electron Devices*, vol. 39, no. 10, pp. 2298–2311, Oct 1992.

[41] Y. S. Chauhan, S. Venugopalan, N. Paydavosi, P. Kushwaha, S. Jandhyala, J. P. Duarte, S. Agnihotri, C. Yadav, H. Agarwal, A. Niknejad, and C. C. Hu, "Bsim compact mosfet models for spice simulation," in *Proceedings of the 20th International Conference Mixed Design of Integrated Circuits and Systems - MIXDES 2013*, June 2013, pp. 23–28.

[42] Y. Tsividis, *Operation And Modeling in MOS Transistor*, Second, Ed. McGraw-Hill, 1999.

[43] C. Enz, F. Krummenacher, and E. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integrated Circuits and Signal Processing*, vol. 8, no. 1, pp. 83–114, 1995. [Online]. Available: http://dx.doi.org/10.1007/BF01239381

[44] E. Seevinck, E. A. Vittoz, M. du Plessi, T. H. Joubert, and W. Beetge, "Cmos translinear circuits for minimum supply voltage," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 47, no. 12, pp. 1560–1564, Dec 2000.

[45] P. M. V. Peteghem, I. Verbauwhede, and W. M. C. Sansen, "Micropower high-performance sc building block for integrated low-level signal processing," *IEEE Journal of Solid-State Circuits*, vol. 20, no. 4, pp. 837–844, Aug 1985.

[46] M. Degrauwe, E. Vittoz, and I. Verbauwhede, "A micropower cmos-instrumentation amplifier," in *Solid-State Circuits Conference, 1984. ESSCIRC '84. Tenth European*, Sept 1984, pp. 31–34.

[47] H. Daoud, S. B. Salem, S. Zouari, and M. Loulou, "Folded cascode ota design for wide band applications," in *International Conference on Design and Test of Integrated Systems in Nanoscale Technology, 2006. DTIS 2006.*, Sept 2006, pp. 437–440.

[48] A. Mourabit, G.-N. Lu, and P. Pittet, "Wide-linear-range subthreshold OTA for low-power, low-voltage, and low-frequency applications," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 52, no. 8, pp. 1481–1488, 2005.

[49] F. A. P. Baruqui and A. Petraglia, "Linearly tunable CMOS OTA with constant dynamic range using source-degenerated current mirrors," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 53, no. 9, pp. 797–801, 2006.

[50] H. Le-Thai, H.-H. Nguyen, H.-N. Nguyen, H.-S. Cho, J.-S. Lee, and S.-G. Lee, "An if bandpass filter based on a low distortion transconductor," *Solid-State Circuits, IEEE Journal of*, vol. 45, no. 11, pp. 2250–2261, 2010.

[51] B. Gilbert, "The multi-tanh principle: a tutorial overview," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 1, pp. 2–17, Jan 1998.

[52] L. H. C. Ferreira, T. Pimenta, and R. Moreno, "An ultra-low-voltage ultra-low-power CMOS miller OTA with rail-to-rail input/output swing," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 54, no. 10, pp. 843–847, Oct 2007.

[53] C. Enz and E. Vittoz, "CMOS low-power analog circuit design," in *Designing Low Power Digital Systems, Emerging Technologies (1996)*, 1996, pp. 79–133.

[54] J. Silva-Martinez, M. Steyaert, and W. Sansen, "Design techniques for high-performance full-CMOS OTA-RC continuous-time filters," *Solid-State Circuits, IEEE Journal of*, vol. 27, no. 7, pp. 993–1001, 1992.

[55] A. Gerosa and A. Neviani, "Enhancing output voltage swing in low-voltage micro-power OTA using self-cascode," *Electronics Letters*, vol. 39, no. 8, pp. 638–639, April 2003.

[56] S.-Y. Lee and C.-J. Cheng, "Systematic design and modeling of a OTA-C filter for portable ECG detection," *Biomedical Circuits and Systems, IEEE Transactions on*, vol. 3, no. 1, pp. 53–64, Feb 2009.

[57] F. Zhang, J. Holleman, and B. Otis, "Design of ultra-low power biopotential amplifiers for biosignal acquisition applications," *Biomedical Circuits and Systems, IEEE Transactions on*, vol. 6, no. 4, pp. 344–355, Aug 2012.

[58] W. M. Chen, H. Chiueh, T. J. Chen, C. L. Ho, C. Jeng, M. D. Ker, C. Y. Lin, Y. C. Huang, C. W. Chou, T. Y. Fan, M. S. Cheng, Y. L. Hsin, S. F. Liang, Y. L. Wang, F. Z. Shaw, Y. H. Huang, C. H. Yang, and C. Y. Wu, "A fully integrated 8-channel closed-loop neural-prosthetic cmos soc for real-time epileptic seizure control," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 1, pp. 232–247, Jan 2014.

[59] K. Abdelhalim, H. M. Jafari, L. Kokarovtseva, J. L. P. Velazquez, and R. Genov, "64-channel uwb wireless neural vector analyzer soc with a closed-loop phase synchrony-triggered neurostimulator," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 10, pp. 2494–2510, Oct 2013.

[60] H. Kassiri, K. Abdelhalim, and R. Genov, "Low-distortion super-gohm subthreshold-mos resistors for cmos neural amplifiers," in *2013 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Oct 2013, pp. 270–273.

[61] W. Wattanapanitch and R. Sarpeshkar, "A low-power 32-channel digitally programmable neural recording integrated circuit," *Biomedical Circuits and Systems, IEEE Transactions on*, vol. 5, no. 6, pp. 592–602, Dec 2011.

[62] R. Harrison, "A low-power, low-noise CMOS amplifier for neural recording applications," in *Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on*, vol. 5, 2002, pp. V–197–V–200 vol.5.

[63] G. Perlin, A. Sodagar, and K. Wise, "Neural recording front-end designs for fully implantable neuroscience applications and neural prosthetic microsystems," in *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE*, Aug 2006, pp. 2982–2985.

[64] R. R. Harrison, P. T. Watkins, R. J. Kier, R. O. Lovejoy, D. J. Black, B. Greger, and F. Solzbacher, "A low-power integrated circuit for a wireless 100-electrode neural recording system," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 1, pp. 123–133, Jan 2007.

[65] J. Kao, S. Stavisky, D. Sussillo, P. Nuyujukian, and K. Shenoy, "Information systems opportunities in brain machine interface decoders," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 666–682, May 2014.

[66] T. Horiuchi, T. Swindell, D. Sander, and P. Abshier, "A low-power CMOS neural amplifier with amplitude measurements for spike sorting," in *Circuits and Systems, 2004. ISCAS '04. Proceedings of the 2004 International Symposium on*, vol. 4, May 2004, pp. IV–29–32 Vol.4.

[67] K. A. Ng and Y. P. Xu, "A compact, low input capacitance neural recording amplifier with cin/gain of 20fF.V/V," in *Biomedical Circuits and Systems Conference (BioCAS), 2012 IEEE*, Nov 2012, pp. 328–331.

[68] H. Scherberger, M. R. Jarvis, and R. A. Andersen, "Cortical local field potential encodes movement intentions in the posterior parietal cortex," *Neuron*, vol. 46, no. 2, pp. 347 – 354, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0896627305001996

[69] D. Heldman, W. Wang, S. Chan, and D. Moran, "Local field potential spectral tuning in motor cortex during reaching," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 14, no. 2, pp. 180–183, June 2006.

[70] F. Wood, M. Black, C. Vargas-Irwin, M. Fellows, and J. Donoghue, "On the variability of manual spike sorting," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 6, pp. 912–918, June 2004.

[71] Y. Kajikawa and C. Schroeder, "How local is the local field potential?" *Neuron*, vol. 72, no. 5, pp. 847 – 858, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S089662731100883X

[72] A. Arieli, A. Sterkin, A. Grinvald, and A. Aertsen, "Dynamics of ongoing activity: Explanation of the large variability in evoked cortical responses," *Science*, vol. 273, no. 5283, pp. 1868–1871, 1996. [Online]. Available: http://www.sciencemag.org/content/273/5283/1868.abstract

[73] K. G. Oweiss, "Multiresolution analysis of multichannel neural recordings in the context of signal detection, estimation, classification and noise suppression," Ph.D. dissertation, 2002, copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2016-05-10. [Online]. Available: https://search.proquest.com/docview/276527451?accountid=16262

[74] D. Y. Barsakcioglu, Y. Liu, P. Bhunjun, J. Navajas, A. Eftekhar, A. Jackson, R. Q. Quiroga, and T. G. Constandinou, "An analogue front-end model for developing neural spike sorting systems," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 8, no. 2, pp. 216–227, April 2014.

[75] M. Ballini, J. Muller, P. Livi, Y. Chen, U. Frey, A. Stettler, A. Shadmani, V. Viswam, I. Lloyd Jones, D. Jackel, M. Radivojevic, M. Lewandowska, W. Gong, M. Fiscella, D. Bakkum, F. Heer, and A. Hierlemann, "A 1024-channel CMOS microelectrode array with 26,400 electrodes for recording and stimulation of electrogenic cells in vitro," *Solid-State Circuits, IEEE Journal of*, vol. 49, no. 11, pp. 2705–2719, Nov 2014.

[76] L. B. Leene, M. Maslik, P. Feng, K. M. Szostak, F. Mazza, and T. G. Constandinou, "Autonomous soc for neural local field potential recording in mm-scale wireless implants," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018, pp. 1–5.

[77] G. Schalk, "Can electrocorticography (ecog) support robust and powerful brain-computer interfaces?" *Frontiers in Neuroengineering*, vol. 3, pp. 9–, May 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2903308/

[78] B. Litt, A. D'Alessandro, R. Esteller, J. Echauz, and G. Vachtsevanos, "Translating seizure detection, prediction and brain stimulation into implantable devices for epilepsy," in *First International IEEE EMBS Conference on Neural Engineering, 2003. Conference Proceedings.*, March 2003, pp. 485–488.

[79] M. Islam, M. Khan, H. Deng, and K. Mamun, "Decoding movements from human deep brain local field potentials using radial basis function neural network," in *Computer-Based Medical Systems (CBMS), 2014 IEEE 27th International Symposium on*, May 2014, pp. 105–108.

[80] K. Michmizos, D. Sakas, and K. Nikita, "Prediction of the timing and the rhythm of the parkinsonian subthalamic nucleus neural spikes using the local field potentials," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 16, no. 2, pp. 190–197, March 2012.

[81] H. Kassiri, A. Bagheri, N. Soltani, K. Abdelhalim, H. M. Jafari, M. T. Salam, J. L. P. Velazquez, and R. Genov, "Battery-less tri-band-radio neuro-monitor and responsive neurostimulator for diagnostics and treatment of neurological disorders," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 5, pp. 1274–1289, May 2016.

[82] A. M. Sodagar, G. E. Perlin, Y. Yao, K. Najafi, and K. D. Wise, "An implantable 64-channel wireless microsystem for single-unit neural recording," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 9, pp. 2591–2604, Sept 2009.

[83] M. Yin and M. Ghovanloo, "A low-noise preamplifier with adjustable gain and bandwidth for biopotential recording applications," in *2007 IEEE International Symposium on Circuits and Systems*, May 2007, pp. 321–324.

[84] X. Zou, X. Xu, L. Yao, and Y. Lian, "A 1-v 450-nw fully integrated programmable biomedical sensor interface chip," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 4, pp. 1067–1077, April 2009.

[85] R. D. Flint, C. Ethier, E. R. Oby, L. E. Miller, and M. W. Slutzky, "Local field potentials allow accurate decoding of muscle activity," *Journal of Neurophysiology*, vol. 108, no. 1, pp. 18–24, 2012, pMID: 22496527. [Online]. Available: https://doi.org/10.1152/jn.00832.2011

[86] S. D. Stavisky, J. C. Kao, P. Nuyujukian, S. I. Ryu, and K. V. Shenoy, "A high performing brain-machine interface driven by low-frequency local field potentials alone and together with spikes," *Journal of Neural Engineering*, vol. 12, no. 3, p. 036009, 2015. [Online]. Available: http://stacks.iop.org/1741-2552/12/i=3/a=036009

[87] A. Khorasani, R. Foodeh, V. Shalchyan, and M. R. Daliri, "Brain control of an external device by extracting the highest force-related contents of local field potentials in freely moving rats," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 1, pp. 18–25, Jan 2018.

[88] R. H. Olsson, D. L. Buhl, A. M. Sirota, G. Buzsaki, and K. D. Wise, "Band-tunable and multiplexed integrated circuits for simultaneous recording and stimulation with microelectrode arrays," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 7, pp. 1303–1311, July 2005.

[89] T. Denison, K. Consoer, W. Santa, A. T. Avestruz, J. Cooley, and A. Kelly, "A 2 956uw 100 nv/rthz chopper-stabilized instrumentation amplifier for chronic measurement of neural field potentials," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 12, pp. 2934–2945, Dec 2007.

[90] R. Muller, S. Gambini, and J. M. Rabaey, "A 0.013 $hboxmm^2$, 5 $muhboxW$, dc-coupled neural signal acquisition ic with 0.5 v supply," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 1, pp. 232–243, Jan 2012.

[91] R. Muller, H. Le, W. Li, P. Ledochowitsch, S. Gambini, T. Bjorninen, A. Koralek, J. M. Carmena, M. M. Maharbiz, E. Alon, and J. M. Rabaey, "A minimally invasive 64-channel wireless ecog implant," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 1, pp. 344–359, Jan 2015.

[92] W. Jiang, V. Hokhikyan, H. Chandrakumar, V. Karkare, and D. Markovi, "A u50-mv linear-input-range vco-based neural-recording front-end with digital nonlinearity correction," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 173–184, Jan 2017.

[93] S. Y. Park, J. Cho, K. Lee, and E. Yoon, "Dynamic power reduction in scalable neural recording interface using spatiotemporal correlation and temporal sparsity of neural signals," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 4, pp. 1102–1114, April 2018.

[94] K. Fujiwara, M. Miyajima, T. Yamakawa, E. Abe, Y. Suzuki, Y. Sawada, M. Kano, T. Maehara, K. Ohta, T. Sasai-Sakuma, T. Sasano, M. Matsuura, and E. Matsushima, "Epileptic seizure prediction based on multivariate statistical process control of heart rate variability features," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 6, pp. 1321–1332, June 2016.

[95] H. Adeli, Z. Zhou, and N. Dadmehr, "Analysis of eeg records in an epileptic patient using wavelet transform," *Journal of Neuroscience Methods*, vol. 123, no. 1, pp. 69 – 87, 2003.

[96] M. S. Lewicki, "A review of methods for spike sorting: the detection and classification of neural action potentials," *Network: Computation in Neural Systems*, vol. 9, no. 4, pp. R53–R78, 1998, pMID: 10221571.

[97] M. S. Fee, P. P. Mitra, and D. Kleinfeld, "Variability of extracellular spike waveforms of cortical neurons," *Journal of Neurophysiology*, vol. 76, no. 6, pp. 3823–3833, 1996, pMID: 8985880. [Online]. Available: https://doi.org/10.1152/jn.1996.76.6.3823

[98] A. Eftekhar, E. P. Sivylla, and G. C. Timothy, “Towards a next generation neural interface: Optimizing power, bandwidth and data quality,” in *2010 Biomedical Circuits and Systems Conference (BioCAS)*, Nov 2010, pp. 122–125.

[99] B. Gosselin, A. E. Ayoub, J. F. Roy, M. Sawan, F. Lepore, A. Chaudhuri, and D. Guitton, “A mixed-signal multichip neural recording interface with bandwidth reduction,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 3, no. 3, pp. 129–141, June 2009.

[100] A. Sodagar, K. Wise, and K. Najafi, “A fully integrated mixed-signal neural processor for implantable multichannel cortical recording,” *Biomedical Engineering, IEEE Transactions on*, vol. 54, no. 6, pp. 1075–1088, June 2007.

[101] I. Bankman, K. O. Johnson, and W. Schneider, “Optimal detection, classification, and superposition resolution in neural waveform recordings,” *Biomedical Engineering, IEEE Transactions on*, vol. 40, no. 8, pp. 836–841, Aug 1993.

[102] M. Abeles and M. H. Goldstein, “Multispike train analysis,” *Proceedings of the IEEE*, vol. 65, no. 5, pp. 762–773, May 1977.

[103] C.-W. Ko, Y.-D. Lin, H.-W. Chung, and G.-J. Jan, “An eeg spike detection algorithm using artificial neural network with multi-channel correlation,” in *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol.20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No.98CH36286)*, vol. 4, Oct 1998, pp. 2070–2073 vol.4.

[104] T. Kalayci and O. Ozdamar, “Wavelet preprocessing for automated neural network detection of eeg spikes,” *IEEE Engineering in Medicine and Biology Magazine*, vol. 14, no. 2, pp. 160–166, Mar 1995.

[105] Y. Oomori, F. Yano, S. P. Ninomija, H. Kodama, T. Fujino, and J. A. Wada, “A computer-assisted spike detecting system using normalized method for epilepsy diagnosis,” in *Proceedings of the Twenty-First Annual Hawaii International Conference on System Sciences, 1988. Vol.IV. Applications Track*, vol. 4, Jan 1988, pp. 51–54.

[106] L. sen Pon, M. Sun, and R. J. Sclabassi, “The bi-directional spike detection in eeg using mathematical morphology and wavelet transform,” in *6th International Conference on Signal Processing, 2002.*, vol. 2, Aug 2002, pp. 1512–1515 vol.2.

[107] S. Kim, J. McNames, and K. Burchiel, “Action potential detection with automatic template matching,” in *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1, Sept 2004, pp. 41–44.

[108] M. Adjouadi, D. Sanchez, M. Cabrerizo, M. Ayala, P. Jayakar, I. Yaylali, and A. Barreto, “Interictal spike detection using the walsh transform,” *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 5, pp. 868–872, May 2004.

[109] Z. Nenadic and J. W. Burdick, “Spike detection using the continuous wavelet transform,” *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 1, pp. 74–87, Jan 2005.

[110] D. Barsakcioglu, Y. Liu, P. Bhunjun, J. Navajas, A. Eftekhar, A. Jackson, R. Quian Quiroga, and T. Constandinou, “An analogue front-end model for developing neural spike sorting systems,” *Biomedical Circuits and Systems, IEEE Transactions on*, vol. 8, no. 2, pp. 216–227, April 2014.

[111] T. Borghi, A. Bonfanti, R. Gusmeroli, G. Zambra, and A. Spinelli, “A power-efficient analog integrated circuit for amplification and detection of neural signals,” in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, Aug 2008, pp. 4911–4915.

[112] R. Chandra and L. M. Optican, “Detection, classification, and superposition resolution of action potentials in multiunit single-channel recordings by an on-line real-time neural network,” *IEEE Transactions on Biomedical Engineering*, vol. 44, no. 5, pp. 403–412, May 1997.

[113] R. Olsson and K. Wise, “A three-dimensional neural recording microsystem with implantable data compression circuitry,” *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 12, pp. 2796–2804, Dec 2005.

[114] S. Barati and A. M. Sodagar, “Discrete-time automatic spike detection circuit for neural recording implants,” *Electronics Letters*, vol. 47, no. 5, pp. 306 –307, March 2011.

[115] ——, "Adaptive spike detection method based on capacitor arrays dedicated to implantable neural recording microsystems," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, Aug 2010, pp. 6445–6448.

[116] R. Q. Quiroga, Z. Nadasdy, and Y. Ben-Shaul, "Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering," *Neural Computation*, vol. 16, no. 8, pp. 1661–1687, Aug. 2004.

[117] Y. Yang, C. S. Boling, A. M. Kamboh, and A. J. Mason, "Adaptive threshold neural spike detector using stationary wavelet transform in cmos," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 23, no. 6, pp. 946–955, Nov 2015.

[118] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[119] K. H. Kim and S. J. Kim, "Neural spike sorting under nearly 0-db signal-to-noise ratio using nonlinear energy operator and artificial neural-network classifier," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 10, pp. 1406–1411, Oct 2000.

[120] Y. Yang and A. J. Mason, "Hardware efficient automatic thresholding for neo-based neural spike detection," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 4, pp. 826–833, April 2017.

[121] Y. G. Li, Q. Ma, M. R. Haider, and Y. Massoud, "Ultra-low-power high sensitivity spike detectors based on modified nonlinear energy operator," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*, May 2013, pp. 137–140.

[122] B. Gosselin and M. Sawan, "An ultra low-power cmos automatic action potential detector," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 17, no. 4, pp. 346–353, Aug 2009.

[123] D. Coue and G. Wilson, "A four-quadrant subthreshold mode multiplier for analog neural-network applications," *Neural Networks, IEEE Transactions on*, vol. 7, no. 5, pp. 1212–1219, Sep 1996.

[124] C.-M. Chang, C.-L. Hou, W.-Y. Chung, J.-W. Horng, and C.-K. Tu, "Analytical synthesis of high-order single-ended-input ota-grounded c all-pass and band-reject filter structures," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, no. 3, pp. 489–498, March 2006.

[125] J. Wild, Z. Prekopcsak, T. Sieger, D. Novak, and R. Jech, "Performance comparison of extracellular spike sorting algorithms for single-channel recordings," *Journal of Neuroscience Methods*, vol. 203, no. 2, pp. 369 – 376, 2012.

[126] Y. Yang, S. Boling, and A. J. Mason, "A hardware-efficient scalable spike sorting neural signal processor module for implantable high-channel-count brain machine interfaces," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 11, no. 4, pp. 743–754, Aug 2017.

[127] Y. Zhou, T. Wu, A. Rastegarnia, C. Guan, E. Keefer, and Z. Yang, "On the robustness of ec-pc spike detection method for online neural recording," *Journal of Neuroscience Methods*, vol. 235, pp. 316 – 330, 2014.

[128] G. Santhanam, M. D. Linderman, V. Gilja, A. Afshar, S. I. Ryu, T. H. Meng, and K. V. Shenoy, "Hermesb: A continuous neural recording system for freely behaving primates," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 11, pp. 2037–2050, Nov 2007.

[129] H.-L. Chan, M.-A. Lin, T. Wu, S.-T. Lee, Y.-T. Tsai, and P.-K. Chao, "Detection of neuronal spikes using an adaptive threshold based on the maxmin spread sorting method," *Journal of Neuroscience Methods*, vol. 172, no. 1, pp. 112 – 121, 2008.

[130] J. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, Apr 1990, pp. 381–384 vol.1.

List of Publications

Journal Publications

1. S. Dwivedi and A. K. Gogoi, A 0.8 V CMOS OTA and its application in realizing a neural recording amplifier, *Medical and Bioengineering (JOMB) , Journal of*, vol. 4, no. 3, pp. 227−234, June 2014.

2. S. Dwivedi and A. K. Gogoi, A novel adaptive real-time detection algorithm for an area-efficient CMOS spike detector circuit, *AEU - International Journal of Electronics and Communications*, vol. 88, pp. 87−97, 2018.

Conference and Workshop Publications

1. S. Dwivedi and A. Gogoi, Improvement of linearity in a CMOS transconductor by employing a squaring circuit, in *Emerging Research Areas and 2013 International Conference on Microelectronics, Communications and Renewable Energy (AICERA/ICMiCR), 2013 Annual International Conference on*, June 2013, pp. 1−3.

2. S. Dwivedi and A. Gogoi, Local field potential measurement with low-power area-efficient neural recording amplifier, in *Signal Processing, Informatics, Communication and Energy Systems (SPICES), 2015 IEEE International Conference on*, Feb 2015, pp. 1−5.