Short-Term Plasticity:
A Neuromorphic Perspective

PhD thesis submitted for the degree of
DOCTOR OF ENGINEERING (Dr.-Ing.)

by

Harshawardhan Ramachandran

Supervisor:

Prof. Dr. Elisabetta Chicca

Reviewers:

Prof. Dr. Martin Paul Nawrot
Dr. Chiara Bartolozzi

Bielefeld University
Faculty of Technology
Universitätsstr. 25
33615 Bielefeld
Germany

April, 2018

Harshawardhan Ramachandran: *Short-term plasticity: a neuromorphic perspective* , the design of neuromorphic short-term plasticity circuits and the demonstration of the short-term plasticity in a cricket calling song recognition network using the neuromorphic hardware, © April 2018

SUPERVISOR:
Prof. Dr. Elisabetta Chicca

DECLARATION

I declare that this thesis entitled "Short-term plasticity: a neuromorphic perspective" is the outcome of my research at the Bielefeld University. This work contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute, except where due acknowledgement has been made in the text.

Bielefeld, April 2018

Harshawardhan
Ramachandran, April 12,
2018

PUBLICATIONS

1. Thomas Rost, Harshawardhan Ramachandran, Martin Paul Nawrot, and Elisabetta Chicca, “A neuromorphic approach to auditory pattern recognition in cricket phonotaxis”, in Circuit Theory and Design (ECCTD), 2013 European Conference on, pp. 1-4. IEEE, 2013.

2. Harshawardhan Ramachandran, Stefan Weber, Syed Ahmed Aamir, and Elisabetta Chicca, “Neuromorphic circuits for Short-Term Plasticity with recovery control”, in 2014 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 858-861. IEEE, 2014.

3. Moritz B. Milde, Olivier J.N. Bertrand, Harshawardhan Ramachandran, Martin Egelhaaf and Elisabetta Chicca, “Spiking elementary motion detector in neuromorphic systems”, Neural Computation, submitted.

4. Harshawardhan Ramachandran, Martin Paul Nawrot, and Elisabetta Chicca, “Short-term plasticity and adaptation as computational primitives for temporal filtering in small neural circuits”, in preparation.

ACKNOWLEDGEMENTS

I would like to acknowledge the following people who played special parts in the journey of my PhD.

I would like to thank Dr. Elisabetta Chicca, for providing me an opportunity to work in this project, for guiding me in the right directions during my research and for her valuable advices that shaped my approach to tackle the problems.

I would like to thank Dr. Martin Paul Nawrot, for sharing his valuable knowledge and supporting me with confidence in the areas of my non-expertise.

I would like to thank Dr. Giacomo Indiveri, for providing critic and valuable suggestions to my research.

I would to like to thank Dr. Emre O. Neftci, Dr. Sadique Sheik and Dr. Fabio Stefanini, for helping me to understand the working of the neuromorphic hardware.

I would to like to thank Dr. Thomas Rost and Dr. Gundula Meckenhäuser, for providing me the simulation and behavior data of cricket phonotaxis.

I would like to thank Dr. K. P. Soman, for inspiring me to do a PhD.

I would like to thank Lauryna Jurkutė, Christian Poth, Dr. Fernando Perez-Peña, and Dr. Stephen Nease, for encouraging me throughout my PhD.

I would like to thank Stefan Weber and Dr. Stephen Nease for the ideas about the design of neuromorphic feed-back Short-Term Plasticity (STP) circuits and STP based band-pass filter circuits.

I would like to thank Uta Scheidt, for collecting data from the fabricated STP circuits.

I would like to thank Lauryna Jurkutė, Tim Walther, Philipp Klein and Dr. Stephen Nease for proof-reading my thesis.

I would like to thank Dr. Sebastian Zehe, Frank Lucio Maldonado Huayaney, Neurobiology group and Active Sensing group for their vaulable feedback to prepare for my PhD defense.

I would like to thank Cordula Heidbrede and Daniel Wolf, for helping me in all my requests.

I would like to thank Ineta Jurkutė, Egidiya Jurkienė and Sigitas Jurkus, for giving me a great moral support.

I would like to thank Dr. Qiang Li, Stefan Weber, Ole Richter, Tim Walther, Marius Schirmer and Philipp Klein for guiding me in Germany.

I would like to thank all the nice people of the Welcome Centre, especially Monika Bokermann and Bodo Gülich, for supporting me with their valuable advice in all the times.

I would like to thank all the present and the past members of the Neuromorphic Behaving Systems group, for nice discussions during the breaks.

I would like to thank all the kind-hearted people in mensa, for providing me a good nourishment.

I would like to thank my parents, for motivating me throughout my career.

ABSTRACT

Short-Term Plasticity (STP) is the ability of the synapse to modify its strength for a short time. Despite several silicon implementations, STP remains one of the least explored topics in the neuromorphic computing research. One form of STP implementation called Short-Term Depression (STD) is available to use in our mixed-signal subthreshold neuromorphic chip. However, the design lacks an independent control over recovery rate of STD. This limitation prevents the circuit to execute a particular synaptic dynamics, such as a strong depression followed by a fast recovery. Another variant of STP called Short-Term Facilitation (STF) is possible to implement in our neuromorphic chip by operating the synapse circuit available, in a specific regime. This operation prevents the time-constant of the synapse to be tuned independently from that of STF implementation. We designed novel STD and STF circuits to solve these problems. The STP circuits are compact in design, but the responses of one of the STP circuits (STF) reached the steady-state values only for certain input frequencies. Therefore, we designed another set of STP circuits by adding a negative feedback loop to our previous design. All these STP circuits are designed and fabricated in a standard Complementary Metal Oxide Semiconductor (CMOS) 180 nm technology and characterized. Alongside the Very Large Scale Integration (VLSI) design, we also demonstrated the role of the STP in a network to recognize the calling songs of crickets. We chose this network due to its small size and the auditory neurons involved in recognition are electrophysiologically studied in the literature. Although several research groups proposed the connectivity of these neurons, the functional structure of the network remains unclear. Therefore, we modeled a spiking neural network using STF in our neuromorphic hardware based on the neurophysiological evidence. Our network model selects the attractive frequencies comparable to the observations in female crickets and gives an idea about the connectivity scheme. Overall, through this research on Short-Term Plasticity (STP), we contributed to two active research fields: neuromorphic computing and computational neuroscience.

CONTENTS

ACRONYMS

VLSI	Very Large Scale Integration
DPI	Differential-Pair Integrator
IF	Integrate-and-Fire
AER	Address Event Representation
PCB	Printed Circuit Board
AMS	Austria Micro Systems
CMOS	Complementary Metal Oxide Semiconductor
IF2DWTA	Integrate-and-Fire 2-Dimensional Winner-Take-All
IFSLWTA	Integrate-and-Fire Soft-Learning Winner-Take-All
BJT	Bipolar Junction Transistor
FET	Field Effect Transistor
JFET	Junction gate Field Effect Transistor
MOSFET	Metal-Oxide-Semiconductor Field-Effect Transistor
FPGA	Field-Programmable Gate Array
CMI	Current-Mirror Integrator
OTA	Operational Transconductance Amplifier
DVS	Dynamic Vision Sensor
EPSP	Excitatory Post-Synaptic Potential
IPSP	Inhibitory Post-Synaptic Potential
EPSC	Excitatory Post-Synaptic Current
IPSC	Inhibitory Post-Synaptic Current
SFA	Spike Frequency Adaptation
STP	Short-Term Plasticity
STD	Short-Term Depression
STF	Short-Term Facilitation

STSP	Short-Term Synaptic Potentiation
PTP	Post-Tetanic Potentiation
PPF	Paired-Pulse Facilitation
LTD	Long-Term Depression
LTP	Long-Term Potentiation
STDP	Spike-Timing Dependent Plasticity
NMDA	N-methyl-D-aspartate
LIF	Leaky Integrate-and-Fire
AN1	Ascending Neuron 1
AN2	Ascending Neuron 2
LN1	Local Neuron 1
LN2	Local Neuron 2
LN3	Local Neuron 3
LN4	Local Neuron 4
BN	Brain Neuron
BNC	Central Brain Neuron
BNC1	Central Brain Neurons 1
BNC2	Central Brain Neurons 2
PD	Pulse Duration
PI	Pulse Interval
PP	Pulse Period
CD	Chirp Duration
CI	Chirp Interval
CP	Chirp Period
ISI	Inter-Spike interval
SD	Standard Deviation
PIR	Post-Inhibitory Rebound
EMD	Elementary Motion Detection
PPR	Paired Pulse Ratio

1 INTRODUCTION

1.1 MOTIVATION OF THIS STUDY

Deep neural networks, the state-of-art in artificial intelligence have been proven to show high accuracy in solving classification problems. However, the number of computational resources utilized by these deep networks are significantly high. Recently, the bio-derived class of neural networks called spiking neural networks are gaining a lot of attention from the computing research community due to its energy efficiency trait. The main ingredient of spiking neural networks are spikes, whose sparse nature minimizes the computational power required to process them. Time is inherently represented in spiking networks, meaning that no additional resource is needed to compute the time. This aspect paved the way for the development of an energy-efficient spike-communication protocol called the Address Event Representation (AER), in which the neurons communicate to each other concerning 'spike-times' and 'neuron-addresses'. The spiking networks can efficiently model the time-varying dynamics of the bio-inspired systems. In recent years, more and more dedicated hardware for running spiking neural networks are being developed. This class of dedicated hardware is called the neuromorphic hardware, and their design/architecture are optimized to implement spike-based algorithms. In this research, we model one such spiking network that recognizes the artificial calling songs of crickets, in the neuromorphic hardware.

We aim to model a system in the hardware that genuinely implements the biophysical models in silicon. Several kinds of neuromorphic hardware exist. Amongst all, the mixed-signal sub-threshold neuromorphic hardware adequately captures the dynamics of the bio-inspired neuron models and operates in biologically realistic time-constants. The hardware runs asynchronously and executes the computation in entirely parallel fashion. This system uses low-power and can be integrated with event-based sensors to be used in real-time closed-loop robotic applications. This type of neuromorphic hardware can also be used to explore the properties of computational neuroscience models. Therefore, the mixed-signal sub-threshold neuromorphic hardware is the best suitable choice to implement our network of auditory pattern recognition in crickets.

The neuromorphic chips are designed to be served as general purpose hardware to emulate the spiking neuron models. However, the silicon models are unchangeable after fabrication. Therefore, the designed system must implement the necessary aspects of the neuron models as much as possible. It is feasible to capture the fine details of the biophysical models by modeling a small-scale system. Therefore, the small-scale design minimizes the risk involved to miss out any aspect in a large-scale system design. It is also important to mention here that the fabrication costs are high to design a custom chip. The chip design has to be updated accordingly because bio-physical models are regularly improving. In this case, small-scale systems are ideal to be designed in dedicated hardware concerning design costs and the chip can be redesigned faster compared to the large-scale hardware. The small-scale system also serves as a prototype for the large-scale design. Considering the advantages mentioned above, we model a small-scale system inspired by the auditory system of crickets in this research.

1.2 OBJECTIVES OF THIS RESEARCH

We aim to understand insects with neural structures that are several orders of magnitude smaller than the mammalian brain yet display a variety of complex behavior. For example, crickets are exciting for neuroscientists due to their acoustic-oriented behavior. Male crickets produce calling songs, and female crickets respond to these calling songs by approaching the males. This phenomenon is called a cricket phonotaxis. The studies on phonotaxis suggest that females are attracted to certain types of male calling songs with specific temporal features. The electrophysiological recordings of the auditory neurons of a cricket brain validate this proof of principle. However, the underlying neural network responsible for the recognition of the calling songs is not evident in the literature. Understanding these small systems can give an insight into the computations occurring in tiny brains. An elementary network of cricket phonotaxis is modeled in [96] based on the neurophysiological evidence. This network model laid the foundation for this research. The model consists of four neurons with Short-Term Plasticity (STP) synapses in between. STP is one of the short and quick learning mechanisms of the synapses of the brain which is used in speech recognition, motor control, etc. Considering the small size, we implemented this network in the existing mixed-signal sub-threshold neuromorphic hardware designed by Prof. E. Chicca and Prof. G. Indiveri, at the Institute of Neuroinformatics, University of Zürich and ETH Zürich. Neuromorphic systems aim at emulating the biophysical mechanisms of the neural elements in the silicon substrate. Calling song recognition network of crickets is compact to study through the neuromorphic chips. During the implementation, we discovered that we cannot implement specific temporal dynamics of STP with the circuit present in the neuromorphic chip. Therefore we designed a series of STP circuits to perform particular computations of the STP such as the detection of bursts of spikes. To summarise, we redesign the STP circuit that can be tuned to emulate specific temporal dynamics to detect bursts. We also demonstrate the STP by modeling the calling song recognition network of crickets in the neuromorphic hardware.

1.3 MAIN CONTRIBUTION OF THIS WORK

With this interdisciplinary research, we aimed to contribute to both the computational neuroscience and the neuromorphic engineering. We designed and fabricated four neuromorphic STP circuits that are capable of detecting bursts of spikes. The proposed STP circuits can be integrated with the existing mixed-signal subthreshold neuromorphic system. Due to its compact design, the STP blocks can be used in large synaptic arrays. A novel calling song recognition network of crickets is modeled using STP. The model selectively chooses the attractive stimuli comparable to the neurophysiological evidence. It also suggests the connection scheme of auditory neurons in cricket brain. This model can be exported to perform acoustic-based tasks in neuro-robots.

1.4 STRUCTURE OF THE THESIS

The structure of this thesis is defined as follows:

Chapter 1 gives the general introduction which includes the motivation, aims, and the contribution of this research.

Chapter 2 is a literature review on short-term plasticity from the computational neuroscience perspective. The biophysical mechanisms of the STP and the commonly used STP models along with examples of the computational roles of the STP in neural circuits are discussed.

Chapter 3 provides an overview of neuromorphic hardware circuits used in this research. A wide variety of topics, ranging from the basic operations of transistors to the complex neuromorphic synapse and neuron circuits are reviewed, and different types of neuromorphic hardware, in particular, the hardware used for this research are covered.

Chapter 4 is dedicated to the neuromorphic circuits of the STP. The existing STP circuits are examined, and the novel STP circuits are proposed. The design of these STP circuits are explained, and their operations are analyzed using simulations and hardware implementations.

Chapter 5 demonstrates the STP in a calling song recognition network of crickets. The network is modeled using the neuromorphic hardware, and the responses of the individual neurons are tuned based on the neurophysiological evidence. The network responses are presented along with the deviations resulting from device mismatch effects, across a group of same networks.

Chapter 6 is the conclusion of this research. Future directions of this study and outlook of the neuromorphic engineering are discussed.

2

SHORT-TERM SYNAPTIC PLASTICITY

2.1 INTRODUCTION

Plasticity in synapses postulates learning in biology. Synaptic plasticity helps to remember the history of activity between the neurons. Neural systems of animals in various developmental stages exhibit different forms of synaptic plasticity. For example, the calyx of Held synapse, located in the mammalian auditory brainstem shows a rapid type of synaptic learning mechanism called Short-Term Plasticity (STP). STP is a type of synaptic plasticity that lasts for a short period ranging from milliseconds to seconds and even minutes. STP relies only on the pre-synaptic activity that modifies the release of the neurotransmitters from the synaptic bouton. Structural changes of the synapse are instead involved in long-term plasticity. Pre- and post-synaptic activities influence long-term plasticity, which supports the formation of lifelong memory [13] as well as working memory [77]. In vivo [110] and in vitro [3] stimulations suggest long-term modifications of synaptic strength lasting for hours or longer. Both short- and long-term plasticities affect the strength of the synapse in two distinct ways: potentiation (or facilitation) strengthens the synapse, depression weakens the synapse. STP is the key element of this research. We built circuits that emulate the temporal dynamics of the STP (in chapter 4). We also demonstrated the STP in a small neural network, which recognizes the calling songs of crickets (in chapter 5). In this chapter, we discuss the two types of STP, its computational properties and the theoretical models of STP. We aim to provide an understanding of the STP, from the perspective of a neuroscientist.

One form of STP called the Short-Term Facilitation (STF) is an enhancement of synaptic efficacy for a short period in the order of tens to hundreds of milliseconds. STF has been reported in neuro-muscular junctions [6], hippo-campus [94], synapses between pyramidal cells, and bi-tufted inter-neurons [95]. Facilitation occurs due to the additive influx of calcium ions following the pre-synaptic spikes, thereby increasing the probability of release of neurotransmitters into the synapse cleft. Many types of synaptic enhancement exist, and they occur on several short timescales. For instance, a type of facilitation called augmentation increases the synaptic strength for a few seconds. Others include Post-Tetanic Potentiation (PTP) that strengthens the synapse for several seconds to minutes and Short-Term Synaptic Potentiation (STSP) that enhances the synapse for several minutes. According to [93], it is often unclear to distinguish augmentation from the PTP.

Another form of STP called the Short-Term Depression (STD) exists in the synapses between the pyramidal cells and the multi-polar inter-neurons [95], synapses in layer 2/3 of rat primary visual cortex [108], and neuro-muscular junctions [18]. STD is the short-time reduction in strength of a synapse due to the depletion of neurotransmitters caused by the pre-synaptic activity. In the pre-synaptic terminal, several sites (or pools) contain the neurotransmitters. They are reserved pool, readily releasable pool, and immediately releasable pool. Depletion of neurotransmitters in the readily releasable pool determines the STD.

Short-term depression and facilitation may coexist at the same synapse. The balance between the two depends on the number of the available vesicles of neurotransmitters, that is, the quantal content. High probability of release of neurotransmitter per action potential favors depression [114] (because the most readily releasable quanta are released first). The remaining quanta are less quickly released (due to the slow replenishment of quantal store). The low probability of neurotransmitter release per action potential favors facilitation [114]. Facilitation does not depend on the release of the neurotransmitter during the pre-synaptic spike. Only the entry of calcium after the first pre-synaptic spike causes facilitation. The residual calcium exists always after every pre-synaptic spike. Over repeated spike activity increases the amount of remaining calcium thereby favors facilitation. Both STD and STF turn the static synapse into a dynamic synapse. The adaptive strength of the STP ensures the synapses to display various temporal dynamics. Few of them will be discussed in the next sections of this chapter.

2.2 MODELS OF SHORT-TERM PLASTICITY

Several models based on the bio-physical mechanisms of Short-Term Plasticity (STP) have been proposed in the literature [108, 71, 107]. STP models from [108] and [71] will be briefly discussed in this section as they are commonly used in neuroscience research.

Abbott and his colleagues proposed a STP model in [108]. They fit the recordings from the excitatory synapses (in layer 2/3) of rat primary visual cortex with the STP model.

The parameters from the fits are then used to predict the responses of the STP model to arbitrary stimuli. This STP model provides a tool to understand the role of synaptic processes in the sensory responses of cortical neurons. In biology, STP is characterized by the change in the amplitude of the Excitatory Post-Synaptic Potential (EPSP) due to the modification in the synaptic strength. In the model, the change in the response amplitude A results from the product of an initial amplitude and dynamic variables representing facilitation and depression. The EPSP amplitude A is defined as:

$$A = A_0 \cdot F \cdot D \tag{1}$$

where A_0 is the initial value of EPSP amplitude. F is the facilitation variable which is $\geqslant 1$, and D is the depression variable which is $\leqslant 1$. The dynamic variables are updated for each incoming pulse by the following equations:

$$D \leftarrow D \cdot d \tag{2}$$

$$F \leftarrow F + f \tag{3}$$

where d and f are constant factors. They represent the amount of depression and facilitation per pre-synaptic action potential. The depression is updated multiplicatively, whereas the facilitation is updated additively, to limit the substantial effects of facilitation, especially during the high-input frequencies.

During Inter-Spike intervals (ISIs), depression and facilitation variables recover toward their initial values exponentially, as given by the following first-order differential equations:

$$\tau_D \cdot \frac{dD}{dt} = 1 - D \tag{4}$$

$$\tau_F \cdot \frac{dF}{dt} = 1 - F \tag{5}$$

where τ_D and τ_F are the time-constants of depression and facilitation.

For simplicity reasons, we considered only the two-compartment model with one depression variable D, as proposed in the original model in [108]. On the contrary, the four-compartment model has several depression variables D_1, D_2, and D_3 and different time-constants. This simple model can capture the main features of the short-term temporal dynamics that affect the strength of the synapse during and between input pulses. The model also predicts the complex stimulation patterns such as Excitatory Post-Synaptic Current (EPSC) responses to random stimulus trains, similar to those occurring in vivo. However, the model lacks the precision to predict the response to specific stimuli within the spike train (due to the increase in the error of the fits that follows the trial-to-trial variability in the data).

We look into another phenomenological model of STP proposed by Markram and his colleagues in [71]. The model was used to analyze the transmission of essential synaptic features to the post-synaptic neuron during STP. The following equations characterize the model:

$$\frac{dR}{dt} = \frac{1-R}{D} \tag{6}$$

$$\frac{du}{dt} = \frac{U-u}{F} \tag{7}$$

where STD and STF are represented as two independent variables R and u respectively. U corresponds to the utilization of synaptic efficacy which is determined by the probability of release of neurotransmitters. D represents the time-constant of depression, and F represents the time-constant of facilitation. This is also a simple model with only three parameters, U, D and F the values of which are $\leqslant 1$. The model is updated for every incoming pre-synaptic action potential by the following equations in the preserved order:

$$R \leftarrow R \cdot (1-u) \tag{8}$$

$$u \leftarrow u + U \cdot (1-u) \tag{9}$$

The parameter U determines the peak value of the first action potential. A small U favors facilitation and a significant U results in depression. Features that get transferred from the pre-synaptic neuron to the post-synaptic neuron during the STP are investigated using this model in [71]. This model suggests the change in the input frequencies is the most significant feature that gets transferred across the neurons. The above mentioned STP models are commonly used by the neuroscientists to demonstrate the computational role of STP in individual neurons as well as in large networks (which will be discussed in the following section). These models laid the foundation of designing STP circuits in silicon [15].

2.3 COMPUTATIONAL ROLE OF STP

Several scientific works explain the computational significance of STP at the level of single neurons as well as in large networks. This section will be useful in understanding the fundamental properties of the STP based on which our circuits and the network are designed. Few of the primary computational roles of STP synapses will be discussed in the following.

2.3.1 *Temporal filtering*

Synapses of a brain act as temporal filters to the incoming neuronal signals. STP implements one such filtering mechanism that affects the strength of the synapse. The strength of the synapse with STD is gradually decreased in response to a continuous stream of pre-synaptic spikes. This effect makes the post-synaptic neuron less responsive to a sustained stimuli. At the same time, the synaptic strength is recovered during the ISIs. An example, describing the filtering properties of STD presented in [107] is shown in Fig. 1.

In case of a low-frequency stimulus, sufficient time is available for STD to recover the synaptic strength back to its initial value. On the other hand, the synaptic strength is reduced quickly, to a high-frequency stimulus. Hence, STD tunes the synapse as a low-pass filter given that the high-frequency components of the stimulus are suppressed, and the low-frequency components are transmitted with the highest strength.

Opposite behavior is observed in the case of STF, during which the strength of the synapse increases in response to incoming spikes. Initially, the post-synaptic neuron does not respond to the incoming spikes, due to the weak synapse. Over repeated

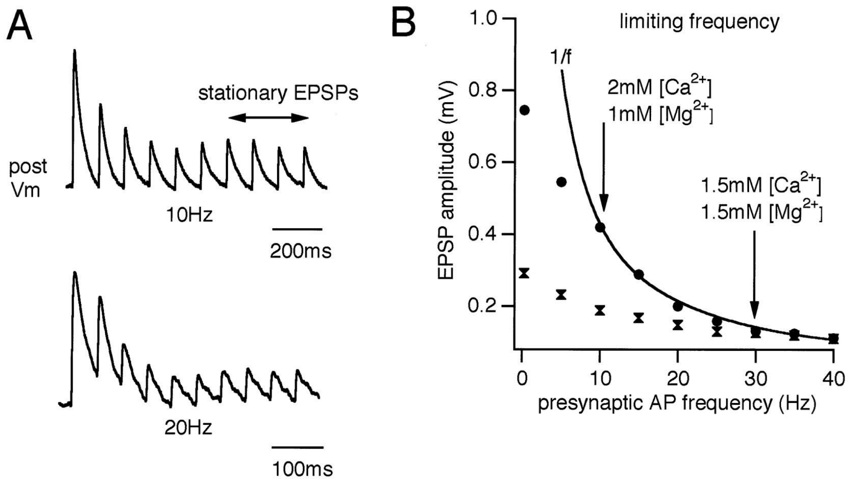


Figure 1: Temporal filtering properties of STD measured at the neocortical pyramidal neurons, presented in [107]. The EPSPs responses of the same neuron, averaged over 20 trials, to a 10 Hz (top) and 20 Hz (bottom) spike train stimuli are shown in (A). In both the frequencies, the amplitude of the EPSP is decreased and eventually reached a steady-state value (stationary EPSP) over repeated stimulation, due to the presence of STD. The EPSPs amplitude recovers towards its resting value during the ISIs. The magnitude of the stationary EPSP in response to the 20 Hz stimulus is smaller than that of 10 Hz stimulus. The EPSP amplitudes (stationary) plotted against the pre-synaptic stimulus frequencies are shown in (B). The solid line shows the inverse relationship of the EPSP amplitude to (stationary) the pre-synaptic input frequency. Filled 'O' marks denote the responses to the high concentration of calcium, and filled 'X' marks denote the responses to low calcium concentration (at the same synapse). The release probability is reduced by lowering calcium concentration (see filled 'O' and filled 'X'), which slows the rate of synaptic depression. The low-pass filter characteristics of the neuron towards its pre-synaptic input frequencies are visible from this plot.

stimulations, the number of input spikes is increased, and as a result, the synapse becomes strong, due to the presence of STF.

The strength of the synapse recovers back to its weak initial value for a low-frequency stimulus due to large ISIs. However, the synaptic strength is increased in response to the high-frequency stimuli. Therefore, STF tunes the synapse as a high-pass filter given that the low-frequency components of the input are suppressed, and the high-frequency components are transmitted with full strength. It is to be noted that the synaptic strengths are always limited by the highest and the lowest possible values.

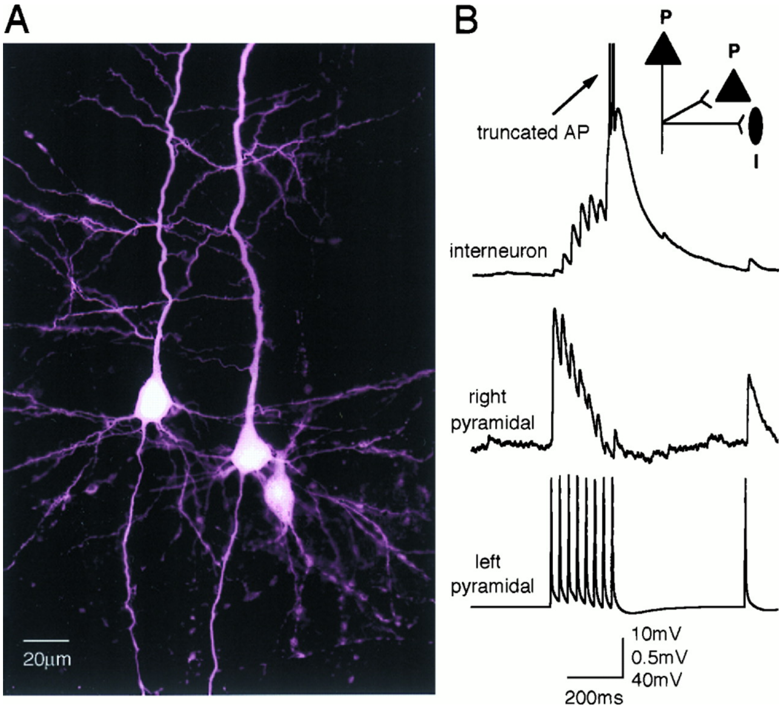


Figure 2: Selective communication between the neurons through STP using neuronal bursts, as shown in [71]. (A) displays the image of three biocytin-filled neurons pictured through the light microscope. Top right of the figure shows the connectivity diagram. The pyramidal neuron (left) is connected to the pyramidal neuron (right) and the bipolar inter-neuron (right). (B) shows the single-trial responses of all three neurons to the same input spike train with 30 Hz frequency. The left pyramidal neuron projects to the right bipolar inter-neuron through the STF synapse. The EPSPs of the inter-neuron builds-up and spikes at the end of the input burst, as a result of the increase in synaptic strength by STF. The left pyramidal neuron projects to the right pyramidal neuron through the STD synapse. The right neuron marks the onset of the burst with a spike, because of the high initial synaptic strength. Eventually, the amplitude of the EPSP is decreased, and no spike is elicited during the burst due to STD. After a long ISI which followed after the burst, the right pyramidal neuron responds again with a spike to a single spike input, as the strength of the synapse is recovered back to its high initial strength.

2.3.2 *Burst detection*

Short-Term Plasticity (STP) in the synapse enables the post-synaptic neuron to detect bursts from the pre-synaptic neuron. 'Bursts' are strictly timed spikes with short ISIs. Following example explains the role of STP in identifying the neuronal bursts. Consider a burst of spikes stimulates a neuron through the STD synapse. Assume the strength of the STD synapse is high when the first spike of the burst arrives at the synapse. The incoming spikes reduce the synaptic strength (due to STD), because of insufficient time available to recover back to its original strength during a burst. In this way, the post-synaptic neuron can detect the onset of bursts due to the high initial synaptic strength.

On the contrary, imagine the post-synaptic neuron with the STF synapse is tuned for a specific frequency and duration of the burst, such that the neuron slowly builds up its Excitatory Post-Synaptic Potential (EPSP) during the burst without eliciting any spike. When the EPSP crosses the threshold, the neuron eventually spikes marking the end of the burst. Therefore, the STD and the STF makes the post-synaptic neuron to detect the onset and offset of the bursts. These burst detection properties enable the neurons to communicate to other neurons of the network [71] selectively.

An example to demonstrate the selective communication through STP is adapted from [71] and shown in Fig. 2. The light microscopic image of the three biocytin-filled neocortical neurons is shown in the left half of figure (A). The connectivity pattern of the three neurons is shown in the top right corner of figure (B), which shows that the left pyramidal neuron innervated the right pyramidal neuron as well as the bipolar inter-neuron (right). The synapse between the left pyramidal neuron and the right pyramidal neuron has STD. The synapse between the left pyramidal neuron and the bipolar inter-neuron has STF. The responses of all three neurons to the same spike train stimulus of 30 Hz frequency are shown in the bottom right part of the figure. When the left pyramidal neuron emits a mixture of spike bursts followed by a single spike with a large ISIs, the bipolar inter-neuron responds only to the bursts (due to the high-pass filter property of the STF). The right pyramidal neuron with the STD synapse responds to both the single spike input and the bursts (due to the high initial synaptic strength). In this way, a single neuron can communicate in different ways to other neurons through the STP synapses and the neuronal bursts.

2.3.3 *Gain control*

STD implements a gain control mechanism in the synapses. An example of this principle is demonstrated in [2] using an integrate-and-fire model. The setup consists of two neuron groups connected to one post-synaptic neuron. This neuron receives a low-frequency stimulus (10 Hz) through 100 synapses from one neuron group and a high-frequency stimulus (100 Hz) through another 100 synapses from the other group. A random spike train stimulus is presented to the network. Three experiments are performed by modulating the input frequency as shown in each column of the Fig. 3.

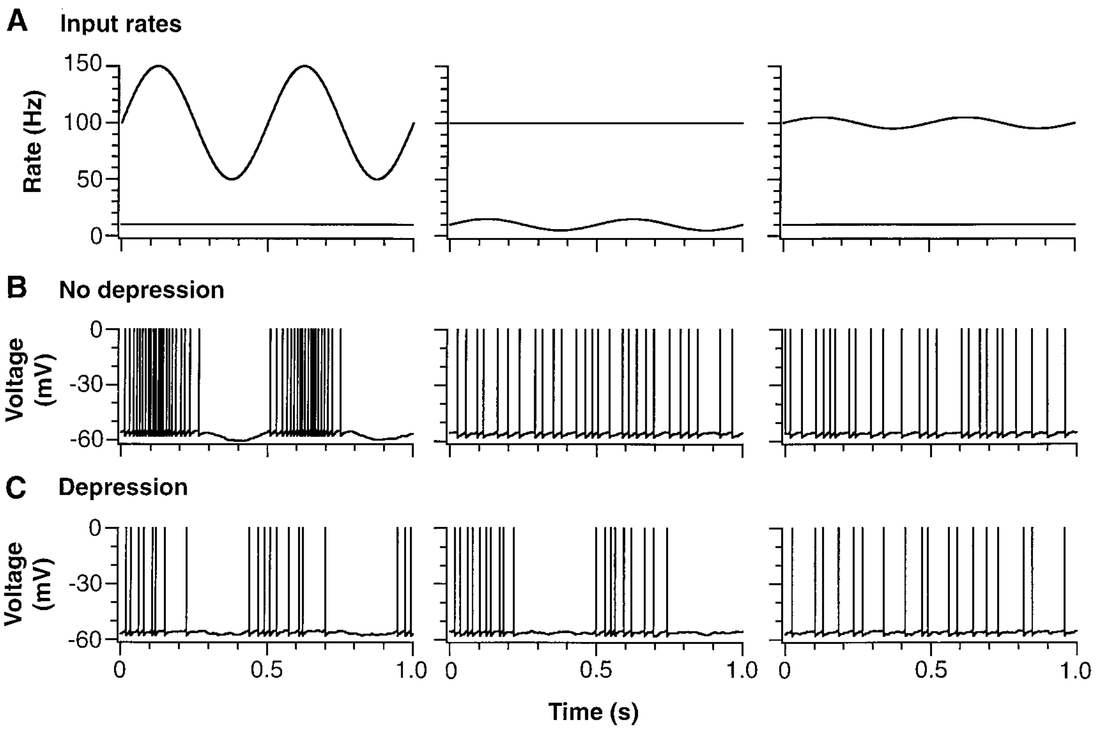


Figure 3: The gain control established by STP, as demonstrated in [2]. (A) shows three input frequency modulations. (B) shows the post-synaptic neuron output without STD synapses. (C) shows the post-synaptic neuron output with STD synapses. Refer to the text for details about the network. Large high-frequency modulations are shown in the left, large low-frequency modulations in the middle and small high-frequency modulations in the right. Neuron without STD in (B) is unable to differentiate between the large low-frequency modulations (middle) and the small high-frequency modulations (right). Neuron with STD in (C) captures the large percentage modulations for the low-frequency stimulus (middle).

- Scenario-1: The high-frequency stimulus is modulated by 50% (i.e., 100 ± 50 Hz) without changing its mean-frequency over time.

- Scenario-2: The low-frequency stimulus is modulated by 50% (i.e., 10 ± 5 Hz).

- Scenario-3: The high-frequency stimulus is modulated by 5% (i.e., 100 ± 5 Hz).

Two different cases are considered for each scenario of this experiment. In the first case (Fig. 3 middle), the neuron has STD synapses and in the second case (Fig. 3 bottom), the neuron does not have STD in its synapses.

Let us start by discussing the case-1. Since there is no depression in the synapses, the strength of the synapses remains unchanged.

- Scenario-1: The post-synaptic neuron can capture the significant modulations of the high-frequency inputs in its output.
- Scenario-2: Large modulations of the low-frequency inputs do not affect the output.
- Scenario-3: The responses to small modulations of the high-frequency inputs look similar to the ones from scenario-2.

Let us proceed to the case-2. In this case, the synaptic weights adapt due to STD.

- Scenario-1: The post-synaptic neuron shows significant high-frequency modulations in its response. Meanwhile, the spike count drops due to STD. However, the synapses are tuned in such a way, that the synaptic weights are restored back before the input modulation completes its cycle.
- Scenario-2: STD amplifies the output of large low-frequency modulations with a high gain and suppresses the domination of high-frequency modulations with a low gain.
- Scenario-3: Unlike the case-1, responses to small high-frequency modulations are distinguishable from the responses to large low-frequency input modulations.

It is important to note in both the cases, the output of the post-synaptic neuron is a result of a combination of the low-frequency and the high-frequency inputs. However, with the presence of STD (case-2), the post-synaptic neuron can capture the modulations both in low- and high-frequencies. This way, STD controls the gain in large networks.

2.3.4 *Direction selectivity*

The role of STD in direction selection is demonstrated in [19] and their implementation is discussed here. A small network of the visual cortex is modeled using the STD synapses. The network is shown in Fig. 4(A). Each circle represents a subset of afferent neurons in the ON-OFF receptive field. The ON afferent neurons stimulate the post-synaptic neuron called the V1 cell when the central region of the receptive field alone is exposed to luminance without the outer surrounding region. The OFF affer-

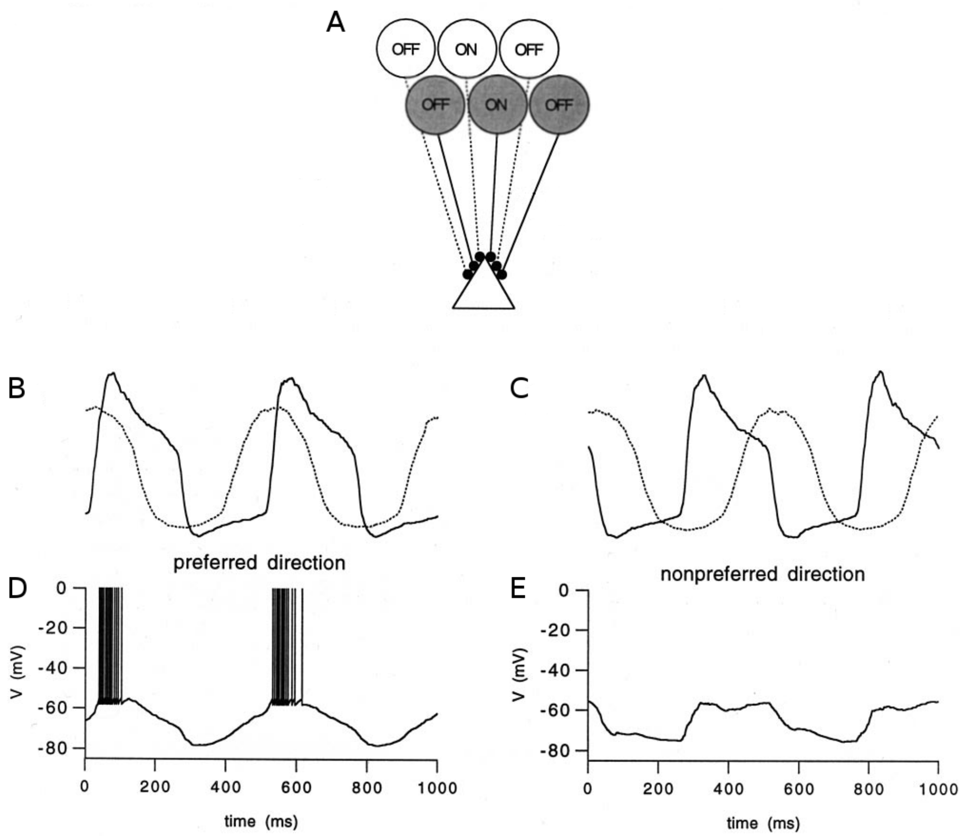


Figure 4: Demonstration of the role of STP in direction selectivity in vision, presented in [19]. (A) The network model of simple cells in the primary visual cortex. The top row represents afferent neurons without the STD, and the bottom row represents afferent neurons with STD synapses. The ON (or the OFF cells) in each row stimulate the V1 cell when the central region (or the surrounding region) of the receptive field is exposed to luminance. The EPSPs of the V1 cell is shown by presenting a sinusoidal signal on each row of the network separately (B and C) and on both the rows of the network simultaneously (D and E). EPSP of the V1 cell, when stimulated in a preferred direction, is shown in (B) and non-preferred direction in (C). Solid curves in (B and C) represents the EPSP of the V1 cell when stimulated through the STD synapses, and the dotted lines (B and C) denote the EPSPs when stimulated through the non-STD synapses. (B) The EPSPs are in phase when stimulated in a preferred direction. (D) The spiking behavior of the V1 cell when stimulated in a preferred direction. (C) The membrane potentials are out of phase when stimulated in a non-preferred direction. (E) The non-spiking behavior of the V1 cell when stimulated in a non-preferred direction.

ent neurons stimulate the same V1 cell only when the surrounding region around the central region of the receptive field is exposed to luminance without the center. The ON and OFF afferent neurons are arranged in two rows, all converging into the V1 cell. In the top row, the synapses between the afferent neurons and the V1 cell have no STD, while in the bottom row, the synapses have STD.

A sinusoidal luminance signal is presented separately to stimulate the two rows of the network. The network is arranged in a spatially distinct manner, such that the stimulus reaches the non-STD afferent neurons row first when it comes from one direction, and the stimulus hits the STD afferent neurons when it comes from the other direction (see Fig. 4(A) for clarity). This arrangement provides a spatially distinguishable response, and STD offers a temporal variability in the output response.

By stimulating the non-STD afferents separately, the V1 cell shows oscillations in its EPSP (see Fig. 4(B) and (C) dotted lines).

When stimulated the STD afferents alone, the V1 cell shows saw-tooth-like waveforms (with a phase-advance) in its EPSP (see Fig. 4(B) and (C) solid lines).

Let us discuss the outcome of presenting the stimulus to both the rows simultaneously. Two directions are possible in this scenario: Either the signal hits the non-STD afferents first (non-preferred direction) or the other (preferred direction).

By presenting the stimulus in a non-preferred direction, even when both the rows responds, the V1 cell fails to evoke a spike, because the afferent outputs are out of phase (see Fig. 4(D)). However, when both the rows are stimulated in the other direction, the two afferent outputs are in phase, thanks to the phase-advancement by STD (see Fig. 4(E)). Therefore, the V1 cell responds with a maximum number of spikes for the preferred direction.

The output of the V1 cell depends on where the signal reaches first. Hence, STD can be used to implement direction selection in networks.

2.3.5 *Encoding sound intensity*

[69] presents the evidence of the STP in synapses of the auditory nerve in the auditory brainstem of the chick. Fig. 5 shows the averaged EPSP responses in response to eight input pulses that are provided for six different pulse frequencies [69]. The

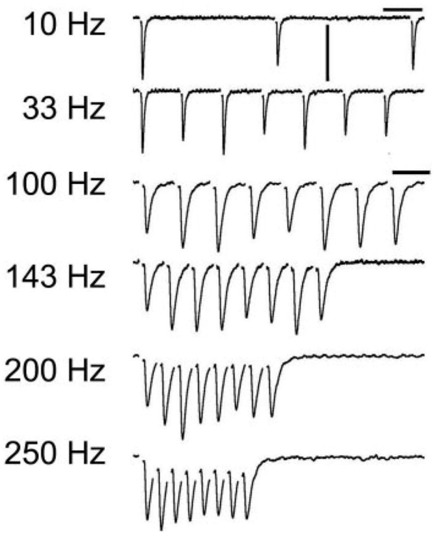


Figure 5: Evidence of STP in the synapses of an auditory nerve in the auditory brainstem of the chick, as presented in [69]. The traces show the average responses of EPSPs to input trains of 8 pulses provided at six different pulse frequencies. The EPSP responses to 10 Hz and 33 Hz stimuli indicate the presence of the STD, as the maximum amplitude of the second EPSP is smaller compared to the first EPSP. In the responses shown from the 100 Hz stimulus to the 250 Hz stimulus, the magnitude of the second EPSP is higher compared to the first EPSP, which shows the presence of the STF with a faster time-constant than the STD. The amplitudes of the EPSP responses to the high-frequency stimuli start decreasing after receiving a certain number of input spikes. This fall in amplitude indicates the presence of the STD with a slow time-constant. The rise and the fall of the EPSP amplitudes in response to the high-frequency stimulus are postulated due to the interplay between the STD and the STF in the same synapse.

EPSPs in response to a low-frequency stimulus shows the presence of the STD, which is evident from the maximum amplitude of the second spikes of the 10 Hz and the 33 Hz stimuli. During intermediate frequencies starting from the second spike of the 100 Hz stimulus up to the 250 Hz stimulus, the effect of the STF is visible in the increase in their maximum amplitudes. At the same time, the maximum amplitudes of the EPSPs start decreasing after the enhancement of a few spikes in response to the high-frequency stimuli due to the STD. The competition between the STF and the STD at the same synapse results in the band-pass filter response of the neuron. Therefore, the presence of STP is evident in the audition, which plays a role in selecting particular frequencies that encode the preferred sound intensities. This kind of temporal band-pass filter is modeled using STP in an auditory network of female crickets to recognize the male calling songs. More details of this model and implementation can be found in see Chapter 5.

2.4 CONCLUSION

So far, we discussed various computational roles of the STP such as temporal filtering, detecting bursts and controlling gain in this chapter. These temporal filtering properties are useful in shaping the network activity. We learned that the synapses with STP are relevant in sensory processing and higher-order cortical processing. The examples we discussed in this chapter justify the importance of modeling STP synapses in a single neuron as well as in large networks. The temporal filtering property of STP is the crucial element of this research. We used STP as a temporal filter in a small neural network, that selects the attractive stimuli (see Chapter 5 for more details). We also characterized the temporal filtering properties of the neuromorphic STP circuits that we designed (refer Chapter 4 for further information).

3

SUB-THRESHOLD NEUROMORPHIC HARDWARE

3.1 INTRODUCTION

Neuromorphic engineering is a term coined by Carver Mead in the late 80's, which describes the use of the Very Large Scale Integration (VLSI) technology to implement the neural computations [72]. VLSI is a process of manufacturing integrated circuits or chips using many transistors. A transistor is a semiconductor device which acts as a voltage controlled current source, depending on the operation region (for example, it acts as a resistor in the ohmic region, see Sec. 3.3 for further explanation). The idea of neuromorphic engineering originated from building silicon neuron circuits by exploiting an equivalence between neuroscience and electronics. The ionic conductance of a biological neuron depends exponentially on the membrane potential of the neuron. Similarly, when the transistor is operating in the sub-threshold regime, the amount of flow of charge carriers in the channel of the transistor is exponentially dependent on the applied gate voltage of the transistor. The definition of the term 'neuromorphic' has been changed over the last two decades. Now any dedicated analog, digital, or mixed-signal hardware that emulates or simulates the computations of neurobiology is referred to as neuromorphic hardware. In the recent years, neuromorphic hardware has gained a lot of attention from the electronics community considering the power efficiency, processing speed, and scalability factor. Mixed-signal (analog/digital) neuromorphic platforms such as the sub-threshold neuromorphic system and 'BrainScales' perform parallel asynchronous computations. Therefore the speed of the operation does not scale with the network size. The power consumption of the sub-threshold hardware is low because the transistors are operated in a sub-threshold regime, during which the magnitude of the currents is in the order of nano- or pico-Ampere. The digital neuromorphic hardware also has its design optimised for power efficiency. For example, IBM's 'TrueNorth' neuromorphic chip is capable of classifying images at 6000 frames per second [32] per watt in comparison to NVIDIA's Tesla P4 which classifies images at 160 frames per second per watt. Therefore, the term 'neuromorphic' refers to the silicon implementation of the powerful and parallel computing elements of the brain. Neuromorphic systems offer a platform to emulate the neural networks directly on the hardware. The size of the network does not influence the speed of this neuromorphic hardware. In contrast to the Von-Neumann architecture of the conventional digital systems, memories are co-localized in the neuromorphic design of the analog systems. This feature makes the neuromorphic systems best-suited for computationally intensive tasks which involve extensive (write and) read operations (to and) from memory. For example, updating the synaptic weights in a network. There are different types of brain-like neural computing systems proposed in the literature, and we will discuss a few of them.

The Stanford University desinged a real-time neuromorphic system known as 'Neurogrid' [12]. Their chip is made of analog sub-threshold circuits that consume a small amount of power. A quadratic integrate-and-fire model is used to implement silicon neurons. The chip has an inbuilt router that communicates to other chips through spike packets. Neurogrid is aimed to be used in neuro-prosthesis and robotic applications.

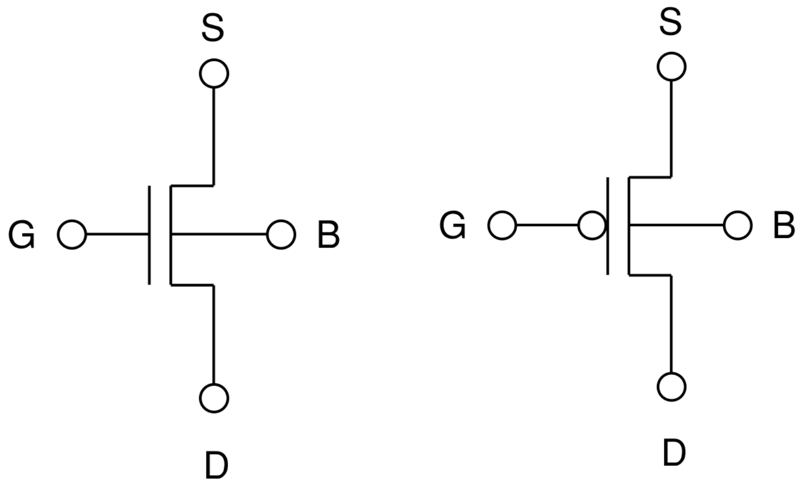
The Institute of Neuroinformatics, University of Zürich and ETH Zürich designed two variants of full-custom mixed-signal neuromorphic chips called 'Dynap-se' for constructing spiking neural networks with dense connections and 'Dynap-le' optimized for online learning [51]. Currently, both the chips are prototypes, and they comprise low-power analog sub-threshold circuits that operate in real-time. These neuromorphic systems aim to reproduce the computations of the brain from the biophysics of the real neurons to the silicon neurons. The applications of this hardware cover a broad scope of possibilities, ranging from brain-machine interfaces to robotic applications. Recently, a neuromorphic processor called 'ROLLS' is designed in ETH Zürich using 180 nm CMOS process [88]. The chip consists of 256 neurons and 128k synapses. The chip is re-configurable and supports online learning. The chip also supports the implementation of attractor neural networks. The analog neuromorphic hardware offers low-power consumption in contrast to their digital counterparts. The hardware offers real-time (or accelerated) parallel computations. However, the models are fixed in the analog hardware and it is prone to issues such as variability in the responses and reproducability of the parameters.

The Heidelberg University developed a multi-scale wafer system called 'BrainScaleS' [99]. BrainScaleS is made of real analog circuits that operate transistors in the above-threshold regime. The wafer-scale system delivers a speed of 10,000 times faster than the real-time. The basic communication is implemented within the wafer, and the wafer-to-wafer communication is implemented through the Field-Programmable Gate Array (FPGA). FPGA is an integrated circuit which consists of programmable logic gates and I/O circuitry. BrainScaleS is designed with the aim to understand time-consuming factors of biological systems such as long-term training tasks. The models of learning synapses are continually evolving due to the recent advancements in neuroscience. Therefore, an on-chip plasticity processor is designed in the latest revision of the BrainScales called the 'HICANN-DLS', in which the learning rules are programmable [37]. The chip consists of 2k synapses and 64 neuron blocks and operates at a speed-up factor of 1000 compared to the biological real-time. Currently this chip is a prototype, and in the long run, it is aimed to be scaled-up to implement large-scale networks.

The University of Manchester developed a massively parallel digital computing machine known as 'SpiNNaker' [38]. SpiNNaker is a multi-core system made of ARM core processors that provide a real-time simulation environment for running synapse and neuron software models. The routing between the cores is based on the packet-switched Address Event Representation (AER) protocol, where the spikes are sent as packets. This system is developed with the goal of modeling large-scale spiking neural networks. The digital neuromorphic hardware offers several advantages such as the flexibility of the neural models and the portability of the parameters. However, the hardware has limitations for successful real-time operations.

Despite the university research groups, the silicon industries are also developing dedicated hardware to implement spiking neural networks, considering the promising outcomes of the neuromorphic computing research. IBM launched a fully digital neuromorphic chip called 'TrueNorth' [76]. Their chip consists of 5.4 million transistors fabricated in 28 nm technology made up of 4096 cores. Each core consists of 256 neurons and 256 synapses. TrueNorth is power-efficient and is used in real-time cognitive applications such as processing high-dimensional visual data.

The selection of the neuromorphic hardware varies with the target application. In this research, we aimed to model the synaptic computations at the level of single neuron in silicon. Given the small size of our network and the need for the biologically realistic time-constants (50-200 ms [89]), our best choice for this research is the sub-threshold mixed-signal hardware developed by Prof. E. Chicca and Prof. G. Indiveri at the Institute of Neuroinformatics, University of Zürich and ETH Zürich. We will start with a basic understanding of the transistors and the building blocks of sub-threshold neuromorphic circuits. Knowledge of these circuits is helpful in uncovering operation of STP circuits proposed in the next chapter. The working principles of silicon synapses and silicon neurons present in our neuromorphic hardware are explained. These synapses and neurons are used in constructing the calling song recognition network (see Chap-



(a) nMOS

(b) pMOS

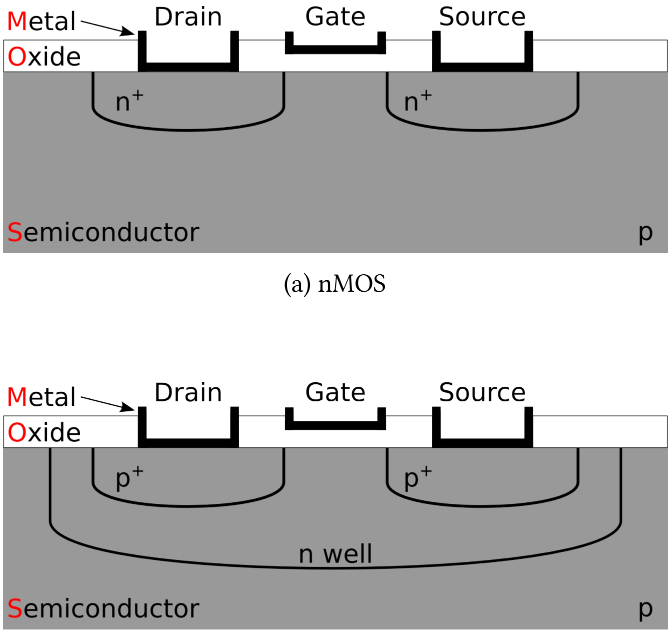
Figure 6: Transistor symbols presented in [64], showing the four terminals: source (S), drain (D), gate (G) and bulk (B). The bubble is used to denote the hole as a majority charge carrier in the pMOS.

ter 5 for more details). The setup of our neuromorphic hardware is covered towards the end of the chapter.

3.2 MOSFET

Transistors form the basic building block of modern electronics. Invented in 1947 at Bell Laboratories, transistors revolutionized the field of electronics. In modern technology, transistors are present in almost all electronic devices ranging from calculators to mobile phones we use every day. Transistors are commonly used in digital circuits to construct logic circuits and switches. The process of the fabrication has been improved a lot over the years, allowing to produce smaller devices. The semiconductor fabrication process defines the technology node based on the size of the smallest transistors that are commercially available. The size of the next technology node is expected to be 10 nm by 2017 in contrast to the 10 $\mathrm{\mu m}$ sized node in the 70's.

Based on the structure, transistors are classified into Bipolar Junction Transistors (BJTs) and Field Effect Transistors (FETs). BJTs use both electron and hole charge carriers in their channels, whereas FETs are uni-polar transistors that are operated by a single-carrier-type in their channels. Junction gate Field Effect Transistors (JFETs) are purely voltage-controlled devices without any need for bias currents to turn them ON. Metal-Oxide-Semiconductor Field-Effect Transistors (MOSFETs) need a minimal current to



(a) nMOS

(b) pMOS

Figure 7: Cross-section view of the nMOS (a) and the pMOS (b) transistors (source: Neuromorphic Engineering I lecture by Dr. Elisabetta Chicca). The nMOS transistor has the n-type diffusion for the source and the drain. The pMOS transistor has the p-type diffusion for the source and the drain. In both the nMOS and pMOS transistors the semiconductor substrate is p-type and a layer of oxide are present between the substrate and the gate. The n-well is implanted on the substrate of the pMOS transistor which serves as bulk for the pMOS.

turn ON, and they can source high currents to the load. MOSFETs is one of the transistor types that are most commonly used in the VLSI design. The word metal (M) in the MOSFET is given because the gate of the transistor is used to be made from the aluminum in the earlier days, whereas now the gates of the transistors are made of poly-silicon. Traditionally, silicon-oxide (O) is deposited on the surface of the semiconductor (S) substrate to isolate the gate from the channel. The term (FET) corresponds to the field-effect transistor. An electric field is applied to alter the conductivity of the channel in the substrate of the MOSFETs. MOSFETs are divided into n-type and p-type based on their majority charge carriers. Fig. 6 shows the symbols of nMOS (a) and pMOS (b). MOSFETs consist of four terminals: source, drain, gate, and bulk. In this chapter, we describe the transistor design based on the single-tub process, because the transistors of the neuromorphic chips we use are manufactured using the same process. In this fabrication process, a separate n-tub or n-well is placed within the p-substrate to implant a pMOS.

The cross-section view of the nMOS and the pMOS transistors are shown in the Fig. 7. In case of both the nMOS and the pMOS transistors, polycrystalline silicon doped with a p-type semiconductor material forms the substrate of the transistors. The gate of the nMOS and the pMOS transistors do not conduct any charge because of the insulator oxide deposited between the gate and the substrate. The bulk is a reference terminal for the transistors.

In the case of an nMOS transistor, the source is an n-type doped material on the p-substrate that serves as a source of electrons which is at a lower potential compared to the gate voltage of the transistor. The drain of the nMOS transistor is at a higher potential than the source and serves as the drain for the electrons. In nMOS, the bulk is the substrate, connected to the Ground. The movement of electrons results in a current flow across the nMOS transistor from its source to the drain through the channel.

The flow of holes (electron-holes) results in a current of pMOS transistors. The source of a pMOS is at a higher potential than the gate and serves as a source of holes. The drain of a pMOS is at a lower potential than the source and serves as the drain to holes. In pMOS, the bulk is a separate 'n'-well (see Fig. 7: top), connected to the power supply (VDD), to the source, or to any arbitrary voltage. According to the standard layout design rules provided by the manufacturer, a proper n-well placed for the pMOS occupies an ample space in the silicon and makes the pMOS design costly [64]. Nevertheless, pMOS is used as widely as nMOS transistors in the analog circuit designs.

3.3 SUB-THRESHOLD CHARACTERISTICS OF A TRANSISTOR

Both the nMOS and pMOS transistors can be operated in two regimes: the sub-threshold and the above-threshold, based on the applied gate voltage. Transistors are used in the above-threshold regime in standard digital electronics as well as in conventional analog electronics such as Operational Transconductance Amplifiers (OTAs). In the digital domain, the transistors operating in the sub-threshold regime are considered to be OFF. Transistors are operated in the above-threshold regime, during which the currents are in the range of micro to milli-Ampere based on the fabrication process.

The operation of a transistor is shown in Fig. 8. When a positive voltage is applied to the gate of an nMOS, the holes in the channel are repelled towards the substrate

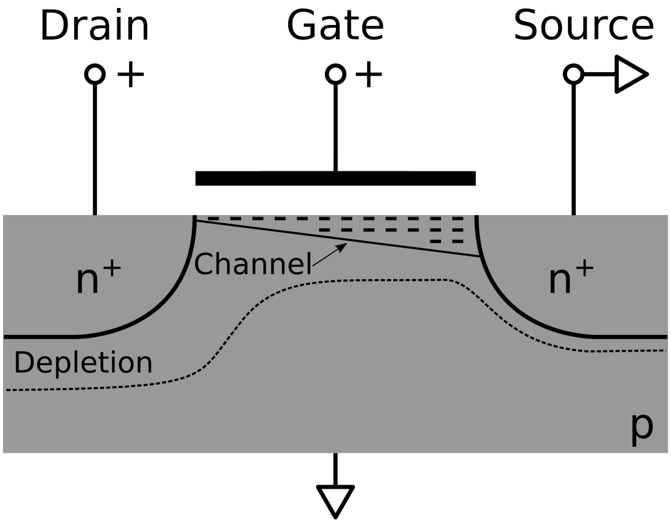


Figure 8: Cross-section of an nMOS transistor (source: Neuromorphic Engineering I lecture by Dr. Elisabetta Chicca). The n-type source and drain diffusion are visible in the p-substrate. The poly-silicon represented a black bar on top of the substrate is used as the gate. The oxide between the gate and the substrate is shown as white space. A depletion region is formed on the p-substrate below the gate between the source and the drain when a positive voltage is applied to the gate of an nMOS. Due to this positive voltage, the holes in the channel are repelled, and a depletion layer of negatively charged ions is formed. A thin inversion layer of free electrons called a channel is formed between the gate and the depletion layer when the gate voltage exceeds a certain threshold.

and a depletion layer of negatively charged ions is formed between the source and the drain below the gate. When the gate voltage exceeds a certain threshold, a thin inversion layer (or channel) of free electrons is also formed between the gate and the depletion layer. Hence, the above-threshold regime is also called a strong inversion mode. When the electric field is applied to the transistor, the drift-currents are created due to the movement of the free electrons. The drift current is the major component of the current source in the above-threshold regime.

In case of the sub-threshold regime, the inversion layer (or channel) is not formed when the gate voltage stays below the threshold. Hence, the sub-threshold regime is also called a weak-inversion mode. The diffusion of charges from source to drain results in diffusion currents. In the sub-threshold regime, the diffusion current is the major component of the current flow. The sub-threshold currents are tiny compared to the above-threshold currents and are in the order of pico- or nano-Ampere. The sub-threshold currents are in similar orders of magnitude (pico-Ampere), as measured in electro-physiology [52]. Small currents result in less power consumption of circuits which makes the sub-threshold transistors ideal choice for developing neuro-

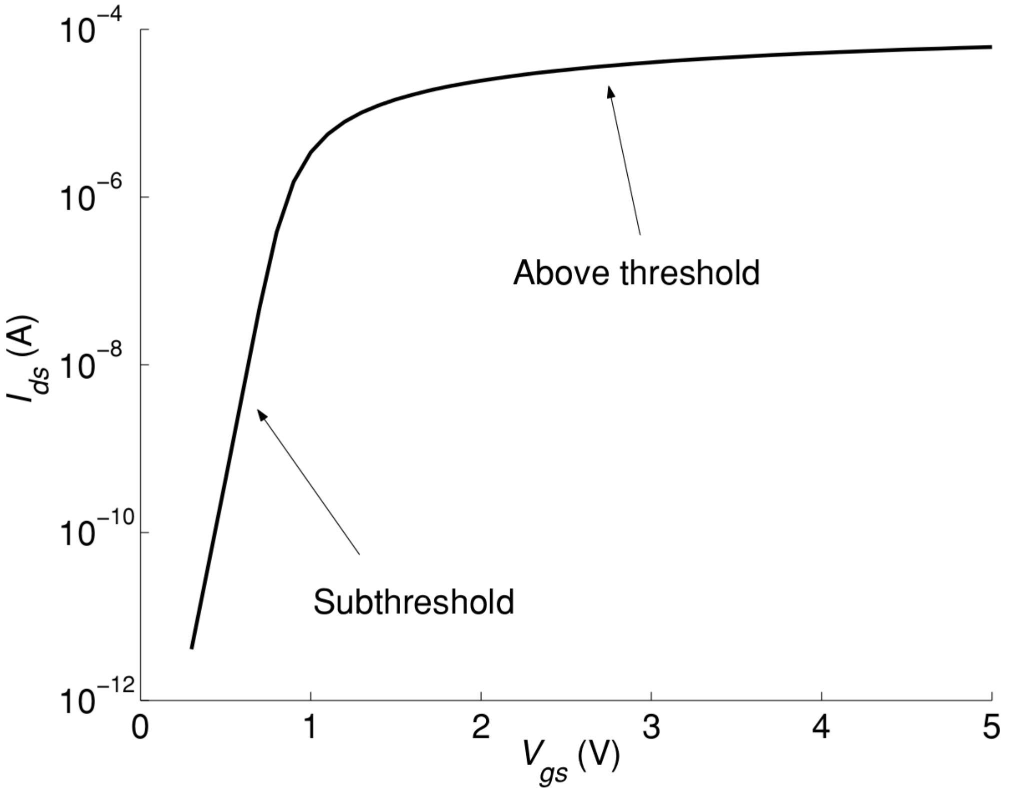


Figure 9: Sub-threshold Current-Voltage (IV) characteristics of the transistor presented in [64]. For a given gate-source voltage V_{gs}, transistor current raises linearly, when V_{ds} is $\leqslant U_T$, and gets saturated, when V_{ds} is $\geqslant 4U_T$, by changing V_{ds}.

morphic chips. These low-power chips are potentially useful in power-hungry computations: for example, real-time object recognition in mobile robots. We also used the sub-threshold transistors in our design of the STP neuromorphic circuits (see Chapter 4).

Current-Voltage (IV) characteristics of a transistor, for the gate-source voltage (gate voltage with respect to the source voltage) V_{gs} sweep is shown in Fig. 9. Both subthreshold and the above-threshold regimes are visible in the plot. In both the regimes, the transistor is operating in the saturation region, which will be discussed in the following section. Current-Voltage (IV) characteristics of a sub-threshold transistor is shown in Fig. 10. There are two regions of sub-threshold operation called linear (or ohmic) and saturation, depending on the applied drain-source voltage (drain voltage with respect to the source voltage) V_{ds}. Let us look into the behaviour of the transistor current in all these regions in the following.

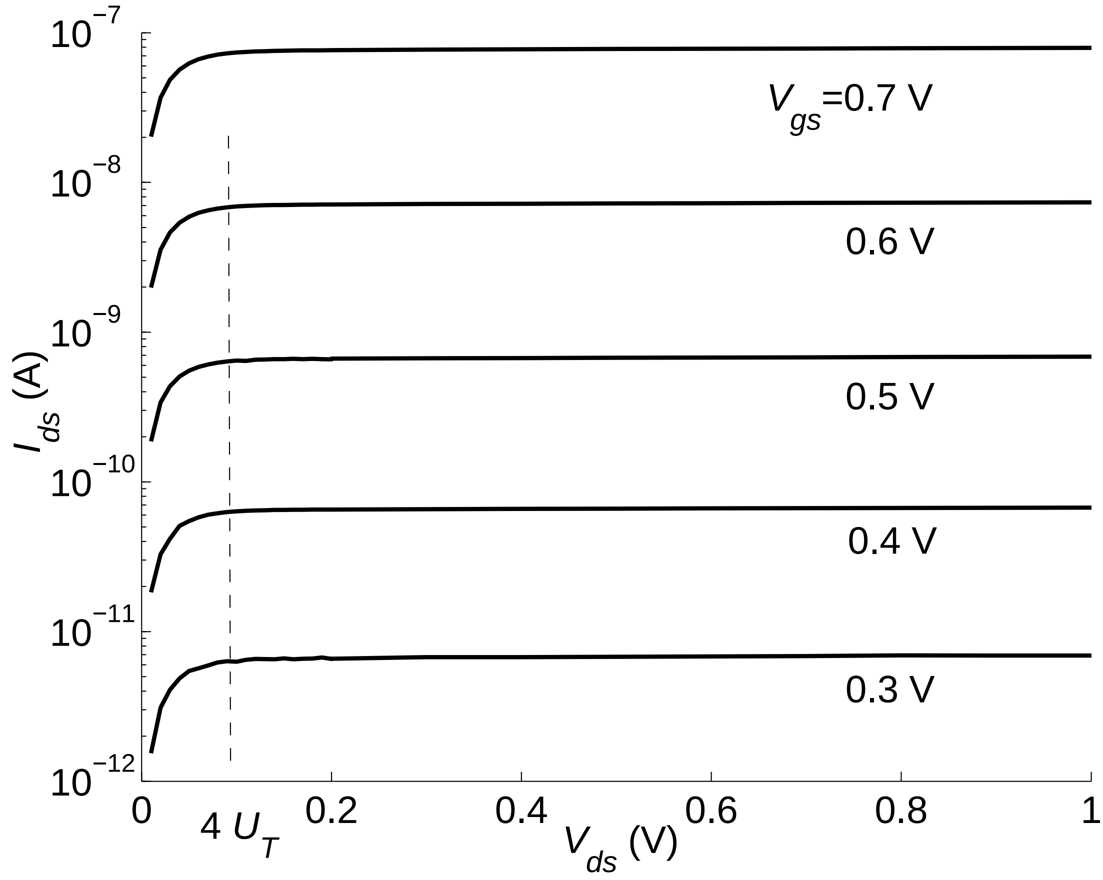


Figure 10: IV characteristics of the sub-threshold nMOS transistor for various gate voltages, presented in [64]. The current increases exponentially for every 100 mV increase in gate-source voltage. Hence, the log y scale is used that spans the entire sub-threshold transistor currents, ranging from pico-Ampere to hundreds of micro-Ampere. Therefore linear increase denotes the exponential rise, and the saturation region denotes the linear rise in currents. It can be seen from the plots, that the transitions from the linear region to the saturation regions are independent of the applied gate-source voltages.

3.3.1 *Sub-threshold ohmic operation*

For small V_{ds}, the current I_{ds} is approximately linear with V_{ds} as shown in Fig. 10. Therefore, it is called the ohmic or linear region of the sub-threshold transistor. The following equations describe the current-voltage (IV) characteristics of the nMOS and the pMOS operating in the sub-threshold linear region.

nMOS

$$I_{ds} = I_{n0} e^{\kappa_n V_g / U_T} \left(e^{-V_s/U_T} - e^{-V_d/U_T} \right) \tag{10}$$

$$I_{ds} = I_{n0} e^{(\kappa_n V_g - V_s)/U_T} \left(1 - e^{-V_{ds}/U_T}\right) \tag{11}$$

For $V_{ds} \leqslant 4U_T$,

$$I_{ds} \simeq I_{n0} e^{(\kappa_n V_g - V_s)/U_T} \left(V_{ds}/U_T\right) \tag{12}$$

pMOS

$$I_{sd} = I_{p0} e^{\kappa_p (V_{dd} - V_g)/U_T} \left(e^{-(V_{dd} - V_s)/U_T} - e^{-(V_{dd} - V_d)/U_T}\right) \tag{13}$$

$$I_{sd} = I_{p0} e^{(-\kappa_p V_g + V_s)/U_T} \left(1 - e^{V_{ds}/U_T}\right) \tag{14}$$

For $V_{sd} \leqslant 4U_T$,

$$I_{sd} \simeq -I_{p0} e^{(-\kappa_p V_g + V_s)/U_T} \left(V_{ds}/U_T\right) \tag{15}$$

where

- I_{n0} and I_{p0} denote the transistor dark-currents. The dark-current comes from the random generation of electrons and holes in the depletion region. The dark-current contributes to leakage and serves as the source of noise in the transistor.

- κ_n and κ_p denote the capacitive coupling ratio that determines the transistor sub-threshold slope factor. κ can be calculated from the slope of the log of IV characteristics of the sub-threshold transistor operating in the saturation region. There is a basic analog circuit called the source follower (explained in next section), the output of which depends on this slope factor.

- U_T the thermal voltage.

- V_g the gate voltage, V_s the source voltage, and V_d the drain voltage.

- V_{dd} is the power supply voltage provided to the bulk.

3.3.2 *Sub-threshold saturation operation*

For $V_{ds} \geqslant 4U_T$, the concentration of electrons at the drain end of the channel becomes negligible concerning the concentration at the source end. The diffusion current becomes independent of the drain voltage and depends only on the source voltage. It is called the saturation operation of the sub-threshold transistor as shown in Fig. 10. An ideal sub-threshold transistor operates in a saturation region. The equations of the saturated sub-threshold transistors are as follows.

nMOS

$$I_{ds} = I_{n0} e^{(\kappa_n V_g - V_s)/U_T} \left(1 - e^{-V_{ds}/U_T}\right) \tag{16}$$

For $V_{ds} > 4U_T$,

$$I_{ds} = I_{n0} e^{(\kappa_n V_g - V_s)/U_T} \tag{17}$$

pMOS

$$I_{sd} = I_{p0} e^{(-\kappa_p V_g + V_s)/U_T} \left(1 - e^{V_{ds}/U_T}\right) \tag{18}$$

For $V_{sd} > 4U_T$,

$$I_{sd} = I_{p0} e^{(-\kappa_p V_g + V_s)/U_T} \tag{19}$$

where

- I_{n0} and I_{p0} denote the transistor dark currents.
- κ_n and κ_p denote the capacitive coupling ratio.
- U_T the thermal voltage.
- V_g the gate voltage, V_s the source voltage, and V_d the drain voltage.

Ideally, in this region, the currents are exponential to the applied gate voltage. Therefore, this exponential characteristic is useful to model the biologically realistic temporal dynamics, which are also exponential. Mostly we operate all the circuits in our neuromorphic chip (mixed-signal hardware) in this region of the transistor. Therefore, the parameter choices are significant for the ideal operation of these circuits. The following section provides a better understanding of the subthreshold operation of these circuits.

3.4 BASIC BUILDING BLOCKS OF ANALOG VLSI CIRCUITS

Complex neuromorphic circuits, such as silicon synapse and silicon neuron circuits, are constructed based on analog building blocks. Few of these basic building blocks will be discussed in this section. The STP circuits we designed in this research are built based on these elements. These circuits are also used in developing general analog circuits outside the neuromorphic domain.

3.4.1 *Diode-connected transistor*

The schematic of the diode-connected transistor is shown in Fig. 11. The transistor itself is a voltage controlled current source. The current is an exponential function of the gate voltage when the transistor is operated in the sub-threshold saturation region. A sub-threshold transistor can be used as an exponential current-to-voltage converter by merely fixing the source voltage and providing the input current through the drain. However, the gate is isolated from the channel, and the input current does not have any effect on the gate. Therefore, the gate is shorted to the drain terminal, to see the change in the gate voltage for the given input current. In this configuration, the input node stays in a positive feedback loop with the gate terminal. As long as a sufficient amount of the current flows, the diode-connected transistor always operates in the saturation region (as the drain is reverse-biased with respect to the channel). The

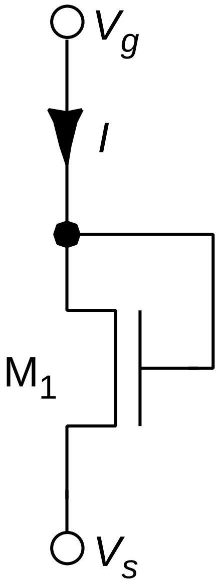


Figure 11: Schematic of an n-type diode-connected transistor, presented in [64]. The positive feedback loop from the drain to the gate is needed to operate the diode-connected transistor in the saturation region.

transistor operates as a diode, allowing the current to flow from drain to source, at the same time, blocking the current flow in the opposite direction. The diode-connected transistor is used in building the current-mirror circuit (see next subsection).

3.4.2 *Current mirror*

The schematic of a current-mirror circuit is shown in Fig. 12. The current-mirror is a two transistor circuit that mirrors the input current into a scaled output current. The circuit consists of two transistors of the same type, where one of them is a diode-connected transistor, and the gates of the two transistors are shorted. The input is the current supplied through the drain of the diode-connected transistor. The output is the scaled version of the input current available at the drain of the other transistor. There are two configurations of the current mirror. In the first configuration, the source voltages are fixed. The transistor dimensions determine the linear scaling factor of the output currents. In the second configuration, the dimensions of both the transistors are identical. The source voltages are varied to determine the current-mirror gain which is exponential in this case. Based on the required type of gain, either the transistors width-and-length ratios or the source voltages are varied. In case of linear gain configuration, the transistor dimensions are fixed after fabrication. Therefore the gain is also non-modifiable. The current-mirror circuits are used in providing a positive-feedback loop, for example, in opamps or in silicon neuron circuits (to generate spikes

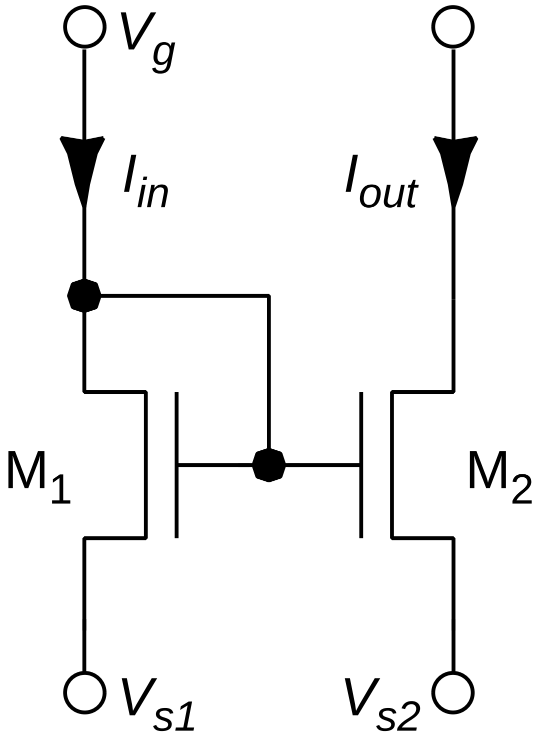


Figure 12: Schematic of an n-type current-mirror circuit, presented in [64]. The input current I_{in} is supplied through the drain of the diode-connected transistor. The gates of two transistors M_1 and M_2 are shorted. Therefore the input current is mirrored at the drain of the output transistor with a scaling factor. The scaling factor is decided either by the difference between the source voltages or the width-and-length ratios, depending on the current mirror configuration.

at a low-power). Current-mirrors can also be used to amplify small sub-threshold currents.

3.4.3 *Source follower*

The schematic of a n-type source-follower circuit is shown in Fig. 13. The source follower is another type of the two-transistor circuit, the output voltage of which follows the input voltage with the gain. Two $nMOS$ transistors are connected in series as shown in the figure. The bias voltage V_{bn} sets the total current flowing through the circuit. The input voltage V_{in} is provided to the input transistor. Output voltage V_{out} is measured at the connecting node of the transistors. The output voltage is the difference between the input and the bias voltages, multiplied by a scalar κ. The source follower can be used to provide an adjustable offset to a voltage, for example, a threshold in the silicon neuron circuit. It is also used in the neuromorphic dynamic vision

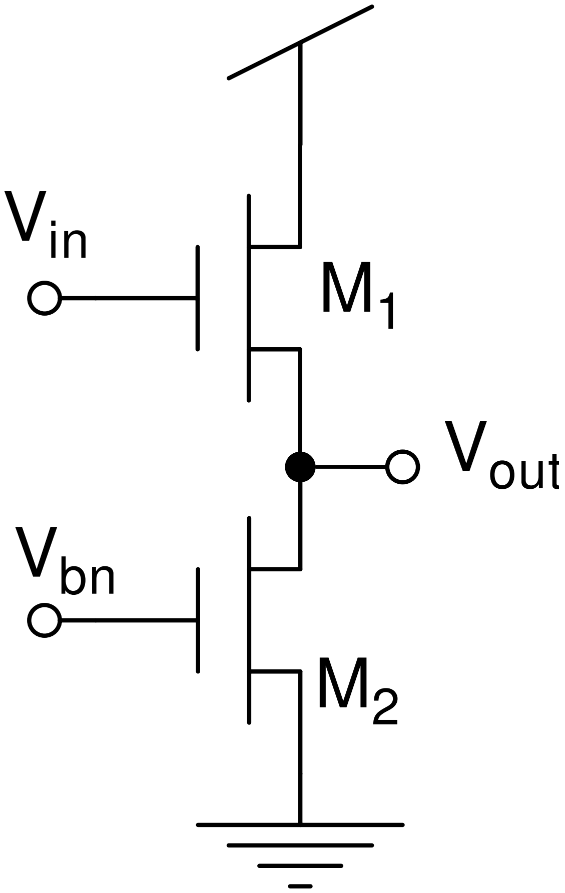


Figure 13: Schematic of an n-type source-follower circuit, presented in [64]. V_{in} is the input voltage and V_{bn} is the bias voltage that sets the total current flowing in the branch of the source follower. The output voltage V_{out} follows the input voltage V_{in} (source) with a gain κ_n.

sensor together with the photo-diode. In this research, we used the source-follower to provide a negative-feedback loop in the STP circuits (refer Sec. 4.4 for more details).

3.4.4 *Differential pair*

The schematic of a differential-pair circuit is shown in Fig. 14. The circuit consists of three transistors (in two branches), the output currents of which depend on the difference of two input voltages. The design of the circuit follows the same structure as the source follower, with an additional transistor connected to the source follower's common voltage node. The bias voltage V_b sets the whole current through the circuit. By Kirchoff's law, the sum of the output currents flowing through two input transistors is equal to the bias current. For a differential input voltage $\delta V = V_1 - V_2$, the resulting currents I_1 and I_2 are sigmoidal in shape. This non-linearity is useful in implementing several neural functions, for example, the activation function of a node in artificial neu-

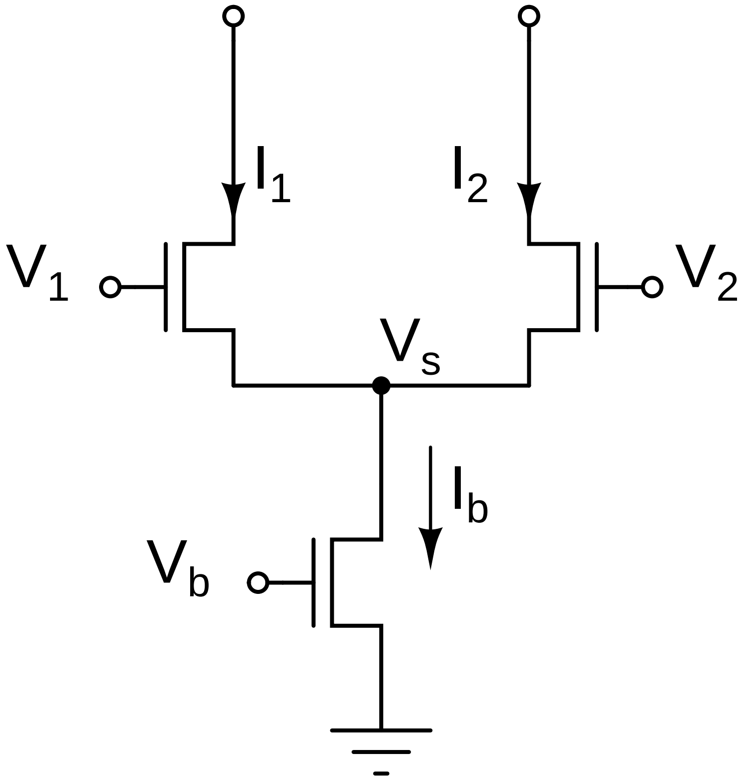


Figure 14: Schematic of an n-type differential-pair circuit, presented in [64]. The differential pair shares the same basic structure as the source follower, except that the bias current I_b from M_b is shared by two input transistors M_1 and M_2. When a differential voltage $V_1 - V_2$ is applied to the input transistors, the resulting output currents I_1 and I_2 will be sigmoidal in shape.

ral networks. The differential-pair circuit is used to build amplifiers, silicon synapses, and neuron circuits.

All these essential analog building blocks we discussed until now, are used in the design of the vital neuromorphic circuits, which will be detailed in the following sections.

3.5 SILICON SYNAPSE

A synapse is a junction through which a neuron communicates with other neurons. The synapse is a fundamental processing unit of neural computation. An average mammalian neuron spans around 1000 synaptic connections with the post-synaptic neuron, at the same time, a pre-synaptic neuron receives input spikes through approximately 10,000 synaptic links (for example, the Purkinje cell of the cerebellum [56]). Two types of synaptic connections exist. They are electrical and chemical synapses. The electrical synapses are also called gap junctions which are useful for short-range transmission of spikes between the neurons. The chemical synapses are used in long-range communications between the neurons and are capable of producing complex synaptic behavior such as amplifying the signals with a high gain. Silicon synapses presented here implement the conductance-based transmission model of the chemical synapses [27]. These synapses integrate the incoming spikes spatiotemporally from the pre-synaptic neurons. Simplified synaptic models describe the Excitatory Post-Synaptic Current (EPSC) as a step increase during the onset of the pre-synaptic spike. The EPSC decays exponentially after the pre-synaptic spike. This rise-and-fall dynamics of the EPSC has been implemented in many silicon synapses. Few of these synapse circuits designed using sub-threshold transistors will be discussed in this section.

Several analog synapses are proposed in the neuromorphic literature. One of the oldest sub-threshold transistor based ‘pulsed current-source’ synapse circuit is proposed by Carver Mead in [72]. The circuit consists of two transistors and outputs the synaptic current. The post-synaptic neuron connected to this synapse integrates the current and shows a rise in its membrane potential. The advantage of this circuit is compact design. However, the circuit cannot integrate the input pulses into continuous currents. Another drawback is, when the input spike trains with same mean rate arrive at different times to this synapse, it is not possible for the post-synaptic neuron to distinguish between the two input spike trains. Nevertheless, this synapse is widely used in [79], [39], and [22] due to its compactness.

Another compact synapse circuit called ‘reset-and-discharge’ was proposed by [61] in the early 1990s. For a given input voltage pulse, the synaptic current is produced. The duration of the output current can be modified by a tunable recovery. However, the circuit response depends on the last input spike only. Therefore, the linear summation of the post-synaptic currents is not possible at the neuron, which is an essential property in the theoretical analysis of neural networks.

A modified version of the ‘reset-and-discharge’ synapse called the ‘linear charge-and-discharge’ synapse circuit is proposed in [5]. It is one of the commonly used synapse circuits in neuromorphic chips. However, the EPSC of the circuit saturates exponentially during which the input frequencies cannot be encoded by the synapse.

The third variant of the linear charge-and-discharge synapse circuit called the Current-Mirror Integrator (CMI) synapse circuit is proposed in [14]. The circuit has a diode-connected transistor that determines the recovery rate of the EPSC. This circuit is also widely used in [47, 66]. However, this circuit can produce large EPSCs amplitudes only for long EPSC durations.

The fourth variant of the 'linear charge-and-discharge' synapse is the 'log-domain integrator' synapse circuit proposed in [5]. One of the transistors of this circuit has its bulk connected to its source terminal instead of V_{dd} (usual case scenario). Therefore, the voltage is logarithmic to the channel currents (hence the name). The circuit has linear filtering properties and can be used to implement large arrays of synapses. These currents can be summed up in a single neuron to elicit a spike. However, this synapse occupies a large silicon area due to the separate bulk connection. Another limitation is the input pulse-width which is usually short, resulting in small currents. This problem can be tackled by having an additional pulse-extender circuit, and it makes the area of the synapse even bigger. All the problems mentioned above are solved by the new design of the synapse circuit called the Differential-Pair Integrator (DPI) synapse. We will discuss this synapse in detail as this circuit implements the silicon synapse in the neuromorphic chip, we used to carry out the experiments in this research.

The DPI synapse circuit was proposed in [9]. The schematic of the DPI synapse is shown in Fig. 15. The synapse consists of four $nMOS$ transistors, two $pMOS$ transistors, and a capacitor. The transistors M_w, M_{thr} and M_τ are arranged in a differential pair configuration along with a diode-connected transistor M_{in}. Spikes from the pre-synaptic neurons (or event-based sensors) are sent as digital input pulses to the gate of the transistor M_{pre}. M_3 acts a digital switch that turns on the path for the capacitor to charge. The weight of the synapse is determined by the currents I_w and I_{thr} (shown in blue) flowing through the transistors M_w and M_{thr} that charge the capacitor C_{syn} within the duration of the input pulse. The time-constant of the synapse is determined by the current I_τ (shown in blue) through the transistor M_τ, which discharges the capacitor in between the input pulses. The voltage across the capacitor controls the gate of the output transistor M_{post} which sources EPSC to the following post-synaptic neuron. Therefore, the currents through the weight and the threshold transistors determine the amplitude of the EPSC (within the input pulse duration) and the current through the time-constant or recovery transistor determines the speed of recovery of the EPSC (in between the pulses). The EPSC shows exponential dynamics (shown in red) in response to the digital input pulse (shown in red). It is worthwhile to mention here that the EPSCs observed in the biological synapses show exponential temporal dynamics as well [27]. Therefore, the DPI synapse captures the biologically motivated computational properties of the synapse.

The EPSC responses of the DPI synapse presented in [9] are shown in Fig. 16. In this experiment, the DPI synapse is stimulated with a digital input pulse. Then, the EPSC responses are recorded for two values of the weight voltage (left) and two values of

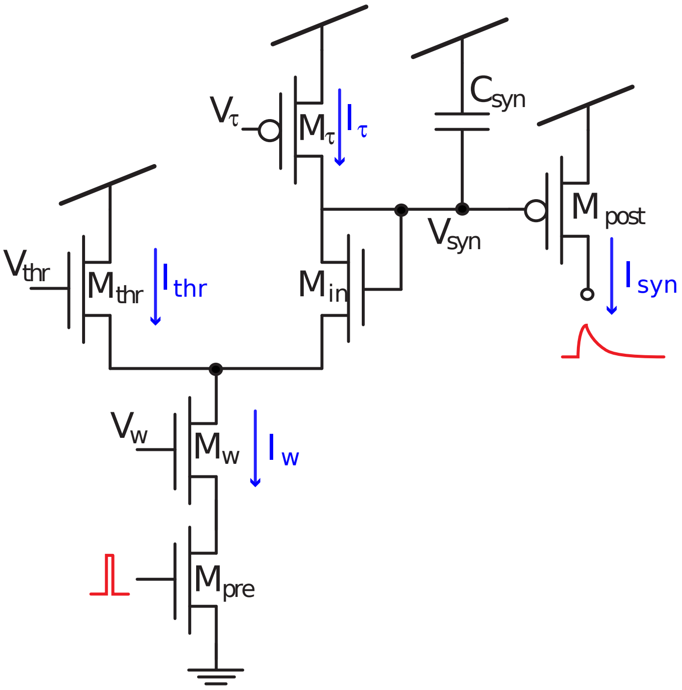


Figure 15: Schematic of the neuromorphic Differential-Pair Integrator (DPI) synapse circuit, proposed in [9]. The circuit consists of six transistors and one capacitor. The circuit is arranged in a differential pair with a diode-connected transistor configuration. Digital input pulses (red) are provided through the gate of the transistor M_{pre}. The capacitor C_{syn} is charged by the currents I_w and I_{thr} (blue) through the transistors M_w and M_{thr}, for the duration of the input pulse. The capacitor is discharged by the current I_τ (blue) through the transistor M_τ in between the pulses. The voltage across the capacitor controls the gate of the transistor M_{post} which determines the output EPSC I_{syn} (blue). The dynamics of the EPSC (red) is exponential in response to the digital input pulse (red), exhibiting low-pass filter characteristics.

the threshold voltage (right). The time-constant voltage is fixed, and the experiment is repeated over ten trials. The mean (black lines) and the Standard Deviations (SDs) (grey shaded region) are plotted in the figure. The lower curves in dotted line (left and right) shows the EPSC for fixed bias voltages. In the left-plot, weight voltage is increased, that is marked by a large amplitude of the EPSC (continuous line). In the right plot, a similar increase in the EPSC amplitude is observed by decreasing the threshold voltage. The threshold voltage of the DPI synapse can be used as a global parameter. For instance, an external homeostasis circuit can be used to modify the weight (EPSC amplitude) of the synapse [8].

We have seen that the EPSC responses (in Fig. 16) show the exponential dynamics similar to the synaptic currents observed in the neurobiology. Next, we characterize the EPSC by the following equation.

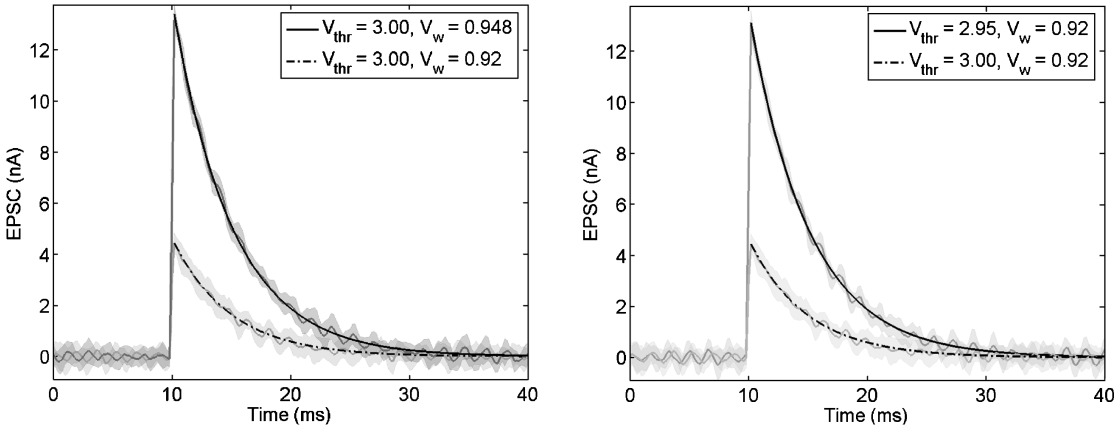


Figure 16: EPSCs of the neuromorphic DPI synapse circuit, proposed in [9], in response to one input pulse. Responses to two variants of the weight voltage (left) and two variants of the threshold voltage (right) are shown. The time-constant voltage is fixed to the experiment. The curves denote the mean (black lines) and the Standard Deviation (SD) (grey shaded region) over ten repeated trials of the experiment. The lower curve (in dotted line - left, right) shows the responses before changing the bias voltages. The plots show that the EPSC amplitude (continuous line) can be increased by either increasing the weight voltage (left) or the threshold voltage (right).

During a spike:

$$I_{syn}(t) = \frac{I_{thr} I_w}{I_\tau}\left(1 - e^{-\frac{(t-t_0)}{\tau}}\right) + I_{syn}(t_0) e^{-\frac{(t-t_0)}{\tau}} \tag{20}$$

In between the spikes:

$$I_{syn}(t) = I_{syn}(t_0) e^{-\frac{(t-t_0)}{\tau}} \tag{21}$$

It is clear from the equations above that during the spike I_{thr} and I_w scales up the EPSC amplitude and I_τ scales it down. It is also evident that the EPSC recovers exponentially in between the spikes.

Several computations observed in biology can be implemented in the DPI synapse, through additional blocks of circuitry that can be attached to it [9]. Few examples of these computational blocks are discussed in the following.

The N-methyl-D-aspartate (NMDA) receptor is a glutamate receptor and ion channel protein present in the nerve cells. This channel is activated only when the membrane potential is depolarized above a certain threshold by binding glutamate and glycine to it. This behavior can be phenomenologically implemented by attaching a p-type differential-pair circuit to the output transistor of the DPI synapse. A bias called

NMDA is provided to the gate of the transistor at one end of the differential-pair, the drain of which is connected to the membrane potential node of the neuron. The membrane potential is provided to the gate of the transistor at the other end of the differential-pair, the drain of which is grounded. Therefore, the circuit implements a threshold mechanism that is activated only when the membrane potential crosses the threshold set by the NMDA parameter. As a result, more synaptic current flows into the membrane potential of the neuron, thereby implementing the voltage-gated NMDA mechanism.

Another example is the Short-Term Depression (STD) block, which is connected to the gate of the weight transistor of the DPI synapse [9]. The details of this circuit will be explained in the next chapter. The STP circuits we modeled in this research (see next chapter) can also be externally attached and used to alter the weight of the DPI synapse.

The long-term plasticity circuit block can be attached to the DPI synapse [23]. But the explanation of this block is out of the scope of this research. In the currently available sub-threshold neuromorphic chips, a complementary version (pMOS replaced by nMOS transistor and vice versa) of the DPI synapse is used to implement the inhibitory synapse.

Above all, the DPI synapse itself can implement a type of Short-Term Plasticity (STP) called the Short-Term Facilitation (STF) (see Chapter 2 for more details), without any extra circuitry. The parameters of the synapse such as the weight and the time-constant voltages are tuned to be small values. The EPSC of the synapse displays the STF dynamics if $I_{syn} << I_{thr}$ [23]. During this condition, the amplitude of I_{syn} increases in response to every input spike, as long as the condition is valid. As I_{syn} increases, eventually this condition is crossed and I_{syn} becomes $>> I_{thr}$. Given this condition, the DPI synapse operates as a first-order low-pass filter. The STF implementation of the DPI synapse is crucial for this research, and it is used to model the calling song recognition network (refer Chapter 5).

The design of the DPI synapse (excitatory) allows a simple modification in its design to implement the inhibitory synapse. For instance, by adding a current mirror to the output transistor. The mirrored EPSC flows in the opposite direction to the actual EPSC. When connected to the same node at the neuron, then by Kirchoff's law, the resulting current will be the difference between them. Therefore, the Inhibitory Post-Synaptic Current (IPSC) can also be generated with the DPI synapse.

The DPI synapse is highly re-configurable and is used as the central synaptic element in designing various synaptic components inspired by the biology. The DPI synapse circuit has linear filtering properties towards incoming spikes. It is also capable of multiplexing the incoming spikes that arrive at different times from different neurons. The size of the synapse layout is an essential factor in designing neuromorphic chips. The synapse size and count determine most of the chip area because a large number of synapses connect multiple neurons in the chip. However, the DPI synapse is compact, and the only space consuming component is its capacitor. The size of the capacitor is chosen based on the time-constant to be modeled. This synapse is aimed to operate with a biologically relevant time-constant (50-200 ms [89]), which makes the design bulky. Nevertheless, considering the advantages mentioned above, DPI synapse has been fabricated in large arrays in several neuromorphic chips.

3.6 SILICON NEURON

A neuron is the fundamental unit of the nervous system. Typically, a neuron consists of the dendrites, the soma, and the axon. Each region of the neuron has a distinct function such as to receive, to generate and to transmit the action potential. The dendrites branch out to receive inputs from many pre-synaptic neurons. The action potential (or the spike) is generated at the 'axon-hillock' and conveyed through the axon to the synapse of the following neurons. This spiking behavior of biological neurons has been emulated in silicon to build neuromorphic chips. Silicon neurons mimic the electrophysiological characteristics of the biological neurons by exploiting the physics of the silicon substrate. There are several types of neuron circuits proposed in the neuromorphic literature that implement the behavior of the spiking neuron models ranging from an elaborate conductance-based or Hodgkin-Huxley neuron model to a simple Integrate-and-Fire (IF) model.

The IF neuron is one of the simplest models that represents the electrical properties of the neurons in terms of a resistor and a capacitor. The input currents from the synapses are integrated by the resitor-capacitor circuitry and the output voltage of the capacitor with the threshold. The neuron fires when the capacitor is charged above a certain threshold. Otherwise, the input currents decay through the circuit. The summation property of the neurons is taken into account in the IF model, in which the membrane capacitor integrates the input synaptic currents. The neurons designed in the neuromorphic hardware we used in this research, are based on the leaky IF model. Therefore, the literature discussed in this section is restricted to the silicon neuron implementations of the IF model.

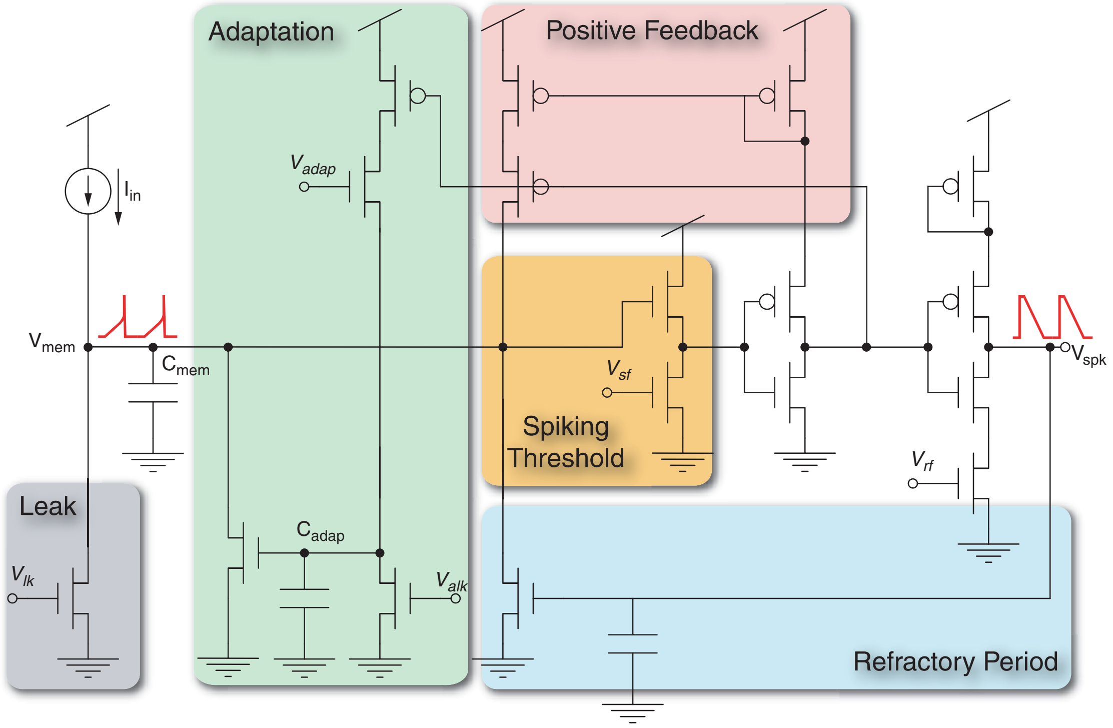


Figure 17: Schematics of the low-power silicon neuron circuit proposed in [49] is shown. The circuit consists of a leak, adaptation, threshold, positive-feedback, and refractory period blocks (color-coded). The input current I_{in} is injected into the neuron from the synapses. As a result, the membrane potential builds up across the capacitor C_{mem}. The leak is controlled by the leak transistor with the gate voltage V_{lk}. When the membrane potential crosses the threshold, set by the source follower bias V_{sf}, the positive-feedback loop is activated. The feedback loop consists of the current-mirror circuit that injects the current back into the neuron. As a result, the first inverter switches fast, and only a small amount of power is consumed. The refractory period is set by the voltage V_{rf} which determines the rate of discharge of the capacitor in the refractory period block. The current through the transistor, the gate of which is connected to this capacitor discharges the membrane capacitor after every spike, implementing the refractory period behavior. When the first inverter switches to the low output during the positive-feedback of spike generation, the adaptation block is turned on. The adaptation capacitor C_{adap} is charged at the rate set by the adaptation voltage V_{adap} and the adaptation leak voltage V_{alk}. When the first inverter switches to high output, the C_{adap} is discharged at the rate set by the V_{alk}. The build-up of charge across the C_{adap} provides an additional leak to the neuron circuit. This charge builds-up over time (after many output spikes), resulting in Spike Frequency Adaptation (SFA). A pMOS transistor is also present (not shown) to externally inject current to the neuron circuit.

The classical neuron circuit, called the 'Axon-Hillock' circuit, was proposed by Carver Mead in the late 80's [70]. The circuit integrates the incoming current and emits a

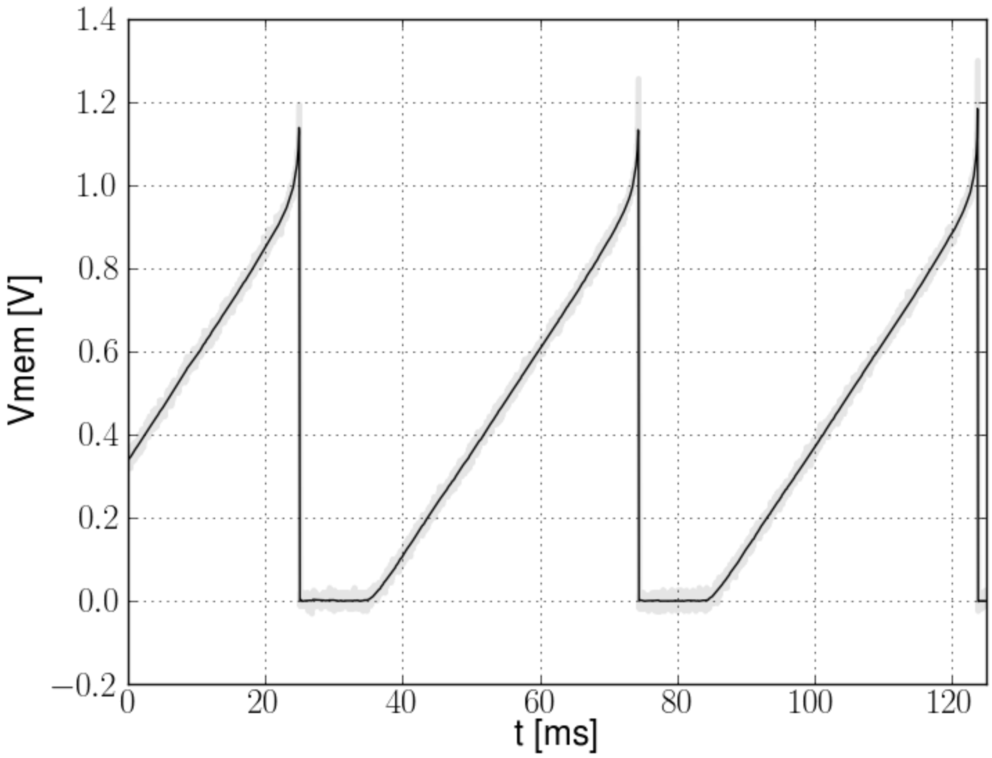


Figure 18: Sample membrane potential plot of the low-power silicon neuron proposed in [49]. A constant input current is injected into the neuron, and the spikes result from integrating the input current by the capacitor.

spike when its membrane voltage crosses the threshold. The design of the circuit is highly compact, as it consists of just six transistors and two capacitors. Four of those transistors are used in two inverters of the circuit. The inverters switch their output voltages when the input voltage crosses the inverter threshold, emulating the spiking threshold behavior. However, the circuits dissipate high power, as the switching time of the inverters is long, and both the transistors of the inverter are ON (draw currents) while switching.

Another sub-threshold integrate-and-fire neuron, called 'Tau-cell', was proposed in [31]. The tau-cell is a current mode circuit. Hence the membrane potential state variable is represented as a current. This circuit operates in low-power and is also used to implement the Mihalas-Niebur as well as the Izhikevich neuron models. However, the circuit lacks a prominent property of neuron, which is the Spike Frequency Adaptation (SFA).

A 'Log domain integrate-and-fire' neuron was proposed in [5]. The circuit The circuit operates in low-power and implements a low-pass filter. This circuit can also perform various types of spiking behavior such as regular spikes, bursts, and SFA. There are 2 pMOS transistors in the circuit, the bulks of which are connected to the membrane potential. Supplying an arbitrary voltage to the n-well increases the size of the layout design. Therefore, fabrication of the chip becomes expensive with this design.

A 'Low-power integrate-and-fire' silicon neuron circuit was proposed in [49]. This neuron implements the SFA, the threshold function, and the refractory period. This circuit is fabricated in the sub-threshold mixed-signal neuromorphic hardware. A variant of this neuron circuit, called the Differential-Pair Integrator (DPI) neuron with the SFA, the refractory period and the threshold functionalities (discussed in [50]) is fabricated in the other variant of the sub-threshold mixed-signal neuromorphic hardware we use. Despite the presence of the differential-pair circuitry in the DPI neuron, this neuron operates similarly to the low-power neuron. The 'low-power integrate-and-fire' neuron circuit is used in this research to model the calling song recognition network of crickets (Chapter 5). Therefore, we will discuss only the low-power IF neuron's operation in this section.

These synapses receive digital input spikes either from several pre-synaptic silicon neurons or digital events from any event-based neuromorphic sensors (for, eg. silicon retina [62]) (as discussed in the previous section). These neurons integrate the input EPSCs (and the IPSCs) from the DPI synapses and emit a spike if the membrane potential crosses the threshold as shown in the Fig. 18.

The schematics of the low-power IF neuron circuit is shown in Fig. 17. The leaky behavior of the membrane potential is implemented in the design. The current through the 'leak' transistor (with the gate voltage V_{leak}) discharges the membrane capacitor slowly and ensures the neuron to be leaky.

The neuron fires when the membrane potential crosses the threshold. A 'source follower' circuit with the bias voltage V_{thr} is used to implement the thresholding function. The spike is generated by activating a positive feedback loop. A 'current-mirror' circuit provides a positive feedback, resulting in a faster rise of the membrane potential [49]. This positive feedback loop ensures the neuron low-power compared to its predecessor Axon-Hillock. In the low-power circuit, the inverters switch faster due to the positive feedback currents.

The refractory period after the spike is also implemented. The current through the transistor (with a bias voltage V_{ref}) discharges the refractory period capacitor. Then, the refractory period capacitor is connected to the gate of an nMOS transistor, and the current through this transistor discharges the membrane capacitor after every spike. Therefore, a high V_{ref} voltage shortens the refractory period.

The Spike Frequency Adaptation (SFA) is implemented in the design. The adaptation is turned on when the first inverter switches to low output during the positive-feedback of spike generation. The adaptation capacitor C_{adap} is charged by the sum of the currents through the transistors with the gate voltages V_{adap} and V_{alk}. When the first inverter output reaches high, the C_{adap} is discharged by the current through the transistor with the gate voltage V_{alk}. The build-up of the charge across the C_{adap} provides an additional leak to the neuron circuit during the spike generation. Over continuous generation of spikes, the charge across the adaptation capacitor builds-up and results in Spike Frequency Adaptation (SFA).

A pMOS transistor is available in the design (not shown in the figure) to inject currents external to the synaptic currents. These external currents are useful in testing the neuron independent from the synapse.

The low-power neuron is compact, able to show various neural dynamics, and is capable of capturing the sub-threshold oscillations [59]. Considering the fact, that the neuron consumes significantly low-power, the earlier version of the same circuit (proposed in [46]) has been integrated into the design of the neuromorphic event-based sensors [7]. This low-power neuron has been replicated into several units to form large arrays of neurons in the neuromorphic chip. More details of this neuron array will be discussed in the next section. In this research, the silicon neuron is used to model the auditory neurons of crickets to recognize their calling songs (refer Chapter 5 for more details).

3.7 NEURON ARRAY

Arrays of the above mentioned synapse and neurons circuits are fabricated in the neuromorphic chips. The spike outputs of these arrays are discussed in time and frequency domains in this section.

Let us start by discussing the time-domain responses of the neurons of the neuromorphic chip, presented in [49]. Raster plots provide the information about the neurons that spiked for a given time duration. The raster plots of the above-discussed neuron array are also presented in [49], which is shown in Fig. 19. Four raster subplots are

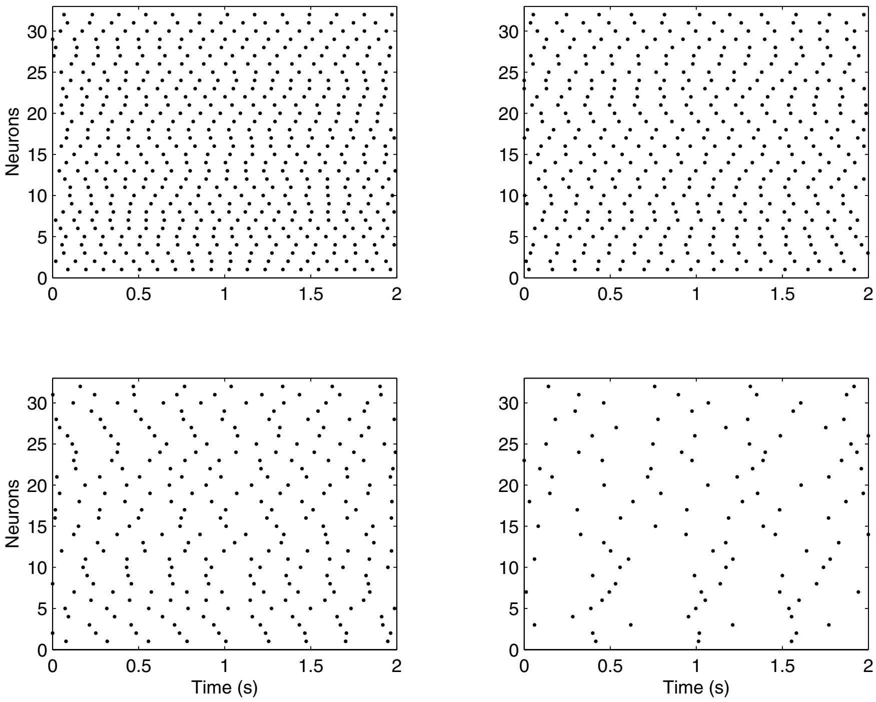


Figure 19: Raster plots presented in [49], obtained for the same four values of the refractory period voltages chosen in Fig. 20. In this figure, a particular amount of the bias voltage V_{gs} of the injection transistor is picked. The effect of the refractory period voltage reducing the total number of spikes from the neurons is visible from the four subplots. Within each subplot, the neurons of the array spike at different times due to the device mismatch from the fabrication process.

drawn for four values of refractory period voltages of the Fig. 20. A particular value of the injection voltage (V_{gs}) is used to inject the same amount of current in all the four subplots. The effect of refractory period is evident from these subplots, which reduces the total number of output spikes from all the neurons. However, within each subplot, the spike-times of the neurons are inconsistent across the array for a given gate voltage. These deviations in the responses among the neurons occur due to the effects of device mismatch.

Device mismatch is an inherent property of the transistors resulting from the fabrication process. According to [84], “Mismatch is the process that causes time-independent random variations in physical quantities of identically designed devices”. The mismatch in transistors generates deviations from the expected behavior, especially in sub-threshold analog circuits. These variations can be controlled by efficient chip design techniques. However, the device variations still exist after fabrication. Many researchers in the neuromorphic community consider the device mismatch a feature. For instance, in [104], the device mismatch is used to model the axonal delays of the neurons. In our research, we used the device mismatch to model different profiles of

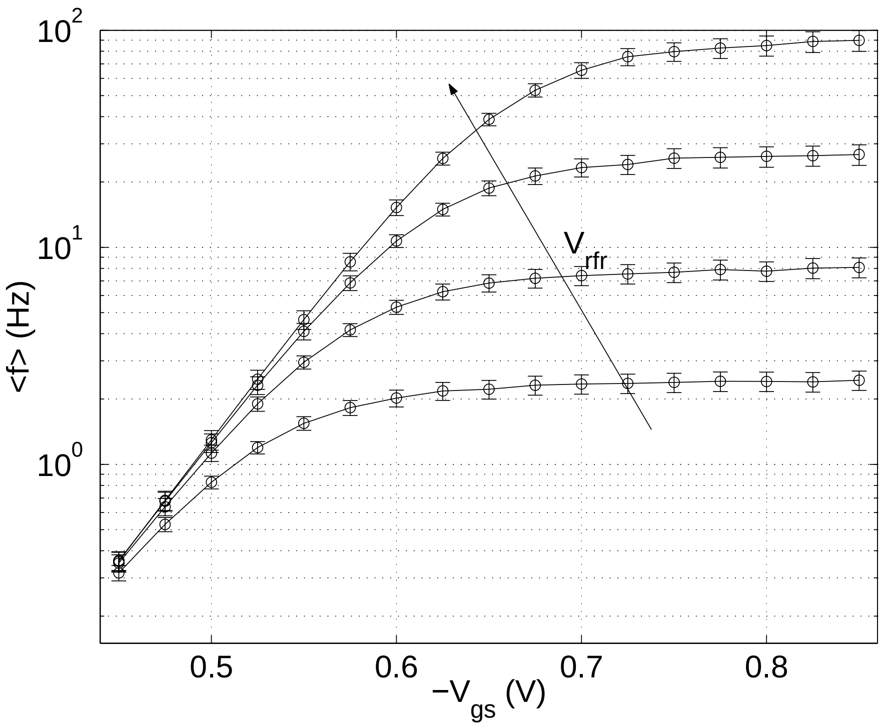


Figure 20: Frequency-Current (F-I) curves presented in [49], obtained from the neuron array of the IFSLWTA neuromorphic chip for four values of refractory period voltage. Each curve in the figure shows the firing rates of the neurons computed for increasing amounts of injection current through the $pMOS$ transistor set by the bias voltage V_{gs} (negative because of $pMOS$). The circles represent the mean firing rates, and the error-bars denote the SDs. The linear increase of the firing rates with the input currents and reaching steady-state values due to the refractory period are visible in the F-I curves.

the band-pass filters, in the network to recognize cricket calling songs. For more details refer Chapter 5. This mismatch property can be used to test the robustness of the networks (designed using these systems) to noise. In this aspect, silicon neurons share the common question with the biological neurons, "do all neurons behave the same?". The answer is "No". One way to compensate the mismatch effects is demonstrated in [81], in which the authors built a recurrent neural network and selectively change the connectivity profile through the AER scheme to normalize the network response. Another approach is to pick the neurons from the chip, whose spikes are aligned in time (as close as possible) with the given parameter sets. Despite these methods to tolerate the mismatch, tuning the parameters of individual neurons to obtain similar responses remains a challenge.

Next, we look into the frequency of the spikes from the neurons of the chip perform in response to the input currents. Let us start by discussing the frequency responses of the neurons. The Frequency-Current (F-I) curves are obtained by increasing the input currents and measuring the output spike frequencies of the neurons. The F-I curves obtained from the neuron array of the neuromorphic chip Integrate-and-Fire Soft-Learning Winner-Take-All (IFSLWTA), presented in [49] is shown in Fig. 20. Four F-I curves are plotted for the same four values of refractory period bias as mentioned above. The circles represent the mean firing rates of the neurons, and the error-bars represent the SDs. The deviations in the responses occur due to the device mismatch as mentioned above. Note that the x-axis represents the gate voltage V_{gs} (negative because of the $pMOS$) of the 'current injection' $pMOS$ transistor. Therefore, the injected currents are exponential to the linear increase in the gate voltage as discussed earlier in this chapter. Also, note that the y-axis is in log-scale. Therefore, the linear rise of the curves (for $V_{gs} \leqslant 0.5V$) represents the exponential increase. The saturation of the curves (for $V_{gs} \geqslant 0.7V$) denotes the linear increase in a non-log scale.

These F-I curves suggest the implementation of the Leaky Integrate-and-Fire (LIF) neuron model in silicon: the firing rates increase linearly with the increase of the input currents and reach the steady-state values due to the refractory period.

3.8 NEUROMORPHIC ARCHITECTURE

So far, we have seen the responses of the arrays of neurons from the neuromorphic chip. In this section, we are going to discuss, how these neuron arrays are organized within the chip. The architecture of the neuromorphic chip Integrate-and-Fire Soft-Learning Winner-Take-All (IFSLWTA) presented in [49] is shown in Fig. 21.

The chip consists of 128 neurons with two excitatory, two inhibitory, and 28 learning synapses each. The neurons are arranged in one-dimensional columns and hence this chip is also reffered to as '1D' chip. Each neuron column is shown as a small central trapezoid. Each column consists of stacks of excitatory and inhibitory synapses, which are shown as squares, marked with 'E' and 'I' respectively. The row and the column encoders are shown as trapezoids in the left and the top. An AER digital input turns on the column and row encoders to stimulate an input pulse on the specific synapse address. The output spikes from the neurons are sent to the AER output circuits that generate digital address-events, which are shown in the bottom trapezoid. The AER protocol is explained in detail in the following paragraph. The parameters of the neurons and synapses are shared across the columns within the chip. The pin-outs from the chip are limited due to the design-cost constraints. The chip was designed with the aim of serving as a general purpose neuromorphic computing hardware. Therefore the biases

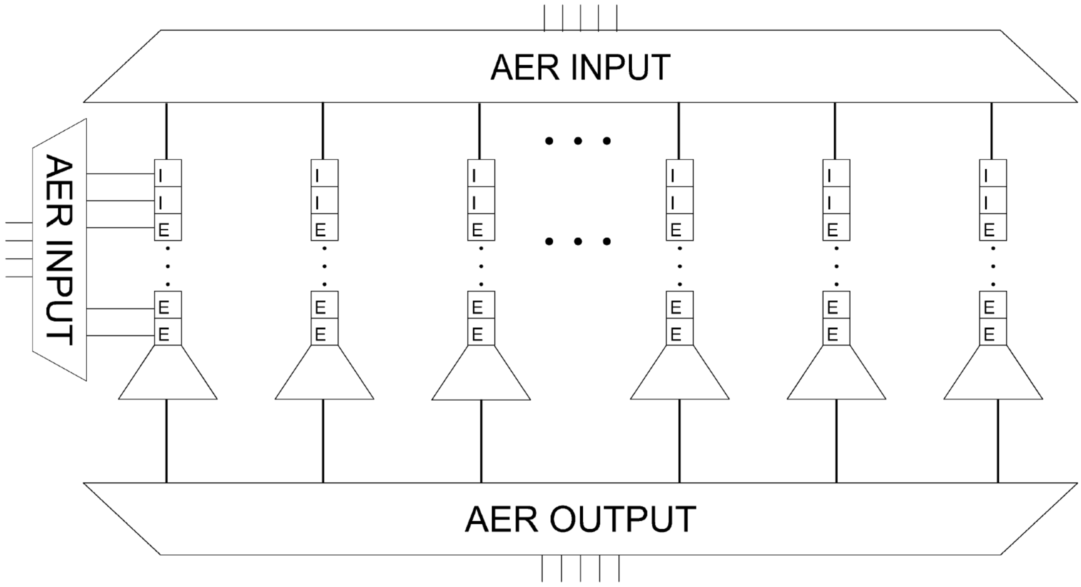


Figure 21: Block diagram of the neuromorphic IFSLWTA chip architecture presented in [49]. The chip contains a total of 128 neurons with two excitatory, two inhibitory synapses, and 128 learning synapses each. Each column in the figure represents the individual neuron block (shown as a small central trapezoid) which consists of stacks of excitatory and inhibitory synapses (shown as squares) marked with 'E' and 'I' respectively. The AER digital input turns on the column (top trapezoid) and row (left trapezoid) encoders to generate an input pulse on the specific synapse address. The output spikes from the neurons are sent to the AER output circuits (bottom trapezoid) that create digital address-events.

are shared across the synapses as well as the neuron arrays. However, they are independent between the excitatory, the inhibitory and the learning synapses within the columns, thereby offering flexibility in tuning the synapses. The post-synaptic neuron integrates the total current from all these synapses. However, note that the synapses that do not receive the events are not turned ON, meaning the chip operates in low-power.

The events are transmitted to (or from) the chip through an asynchronous handshaking protocol called AER. An example of the AER communication scheme is shown in Fig. 22. The output events from the source chip are shown in the left. The neuronal spikes from this chip are encoded into address events with the time-stamps (event = spiking neuron address, time of spike). These events are redirected asynchronously to the destination chip through the AER bus. The direction is based on the look-up table programmed in the mapper. The lookup table consists of the source and the destination addresses of the neurons. The mapper routes the events across the chips accordingly. Encoder and decoder are used for on-chip routing without specifying the neuron ad-

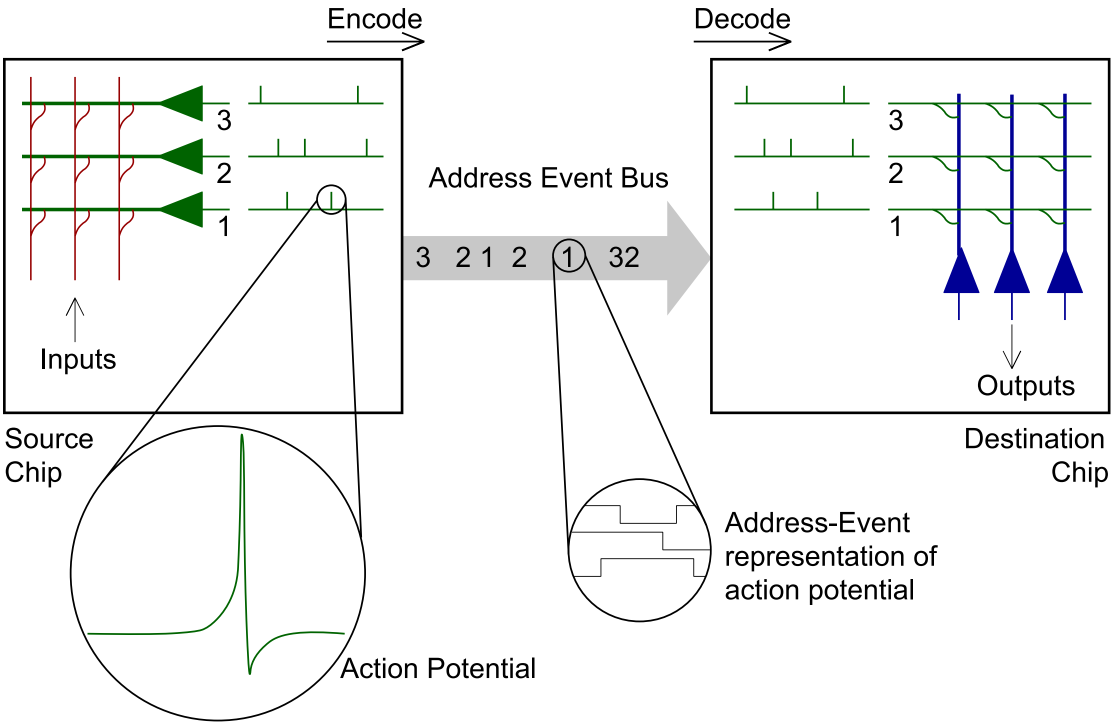


Figure 22: Block diagram of the AER communication protocol between two neuromorphic chips presented in [48]. The left part of the figure shows the output events of the source chip, the neuronal spikes of which are encoded into addresses with the time-stamps of spikes. These events are redirected to the destination chip through the AER bus with the help of the look-up table of the connectivity matrix programmed in the external mapper (not shown). The mapper is used to route the events across the chips through the look-up table. Encoder and decoder are used for the on-chip communication without the neuron addresses. The events are decoded in the destination chip into input pulses that reach the corresponding addresses of the synapses.

dresses. The events are finally decoded in the destination chip. These decoded events are used as input pulses to stimulate the corresponding addresses of the synapses. This asynchronous communication allows the real-time operation of the hardware, without depending on any external clock. So far, we discussed the external architecture of the setup.

Our setup consists of three neuromorphic chips designed by Prof. E. Chicca and Prof. G. Indiveri at the Institute of Neuroinformatics, University of Zürich and ETH Zürich, and connected as shown in Fig. 23 from [104]. Two are called Integrate-and-Fire 2-Dimensional Winner-Take-All (IF2DWTA) chips with 2048 neurons, and each neuron has two excitatory synapses and two inhibitory synapses. The third is the Integrate-and-Fire Soft-Learning Winner-Take-All (IFSLWTA) chip. The excitatory synapses of the above mentioned chips support Short-Term Plasticity (STP), which we will discuss

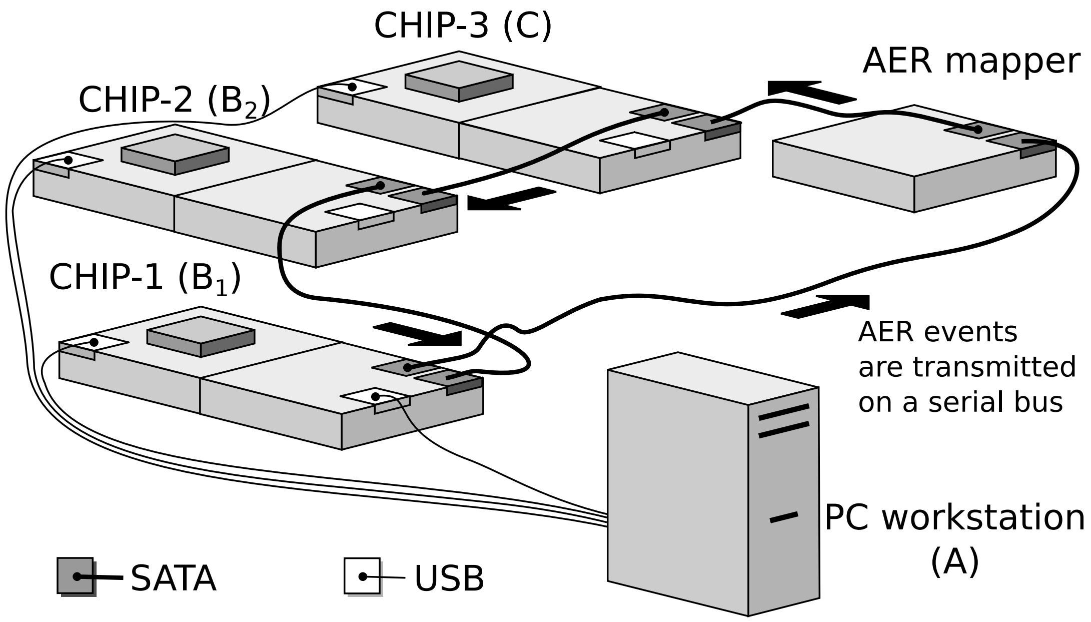


Figure 23: Schematic representation of our neuromorphic hardware setup presented in [104]. The hardware setup consists of three chips connected in a loop through SATA with the external AER mapper board. There are two IF2DWTA chips and one IFSLWTA chip. The IF2DWTA chip consists of 2048 neurons with two excitatory and two inhibitory synapses each. The IFSLWTA chip consists of 128 neurons with two excitatory, two inhibitory and 28 learning synapses each. Multiple chips are used to increase available neuron and independently tunable synapse count. The neurons and the populations are selected from the PC. The connectivity between the neurons is reconfigurable within the chip and across the chips. The connectivity matrix for connections between the chips is stored in the look-up table of the external mapper. One of the chips is connected to the PC through USB, through which the programmed spike events are sent from the PC. The output spikes are recorded from the chip through USB. Analog membrane voltage can be measured through an oscilloscope from the respective PCB of the chip and the neuron to be measured is programmable.

in detail in the next chapter. The term '2D' suggest the arrangement of the neural arrays within the chip. Both the chips are capable of implementing a 'Winner-Take-All' network, a computational paradigm of spiking neural network [80]. All these chips were fabricated in a standard Austria Micro Systems (AMS) 350 nm CMOS process.

Multiple chips are used in the setup to increase the number of neurons as well as independently tunable synapses available. The three chips are connected in a loop and this multi-chip architecture is analogous to have separate cortical regions of the brain. This architecture allows us to freely pick the neurons from any chip and connect them in any possible ways within the chip and across the chips. An external mapper is used to connect the neurons between the chips using a look-up table to build a spiking neural network. Encoder and decoder are used to communicate within the chip. One of the chips is connected to communicate to the PC through USB. The spike events are programmed in the PC and are sent from the computer as input spikes to the DPI synapse (one or many synapses) in the chip. In response to the input spikes, the synapse produces a current which is propagated to the neuron attached. The output spikes are recorded from the chip through USB. Analog membrane voltage can also be measured through an oscilloscope, directly from the respective PCB of the chip. The neuron to be measured is programmable.

In our research, we used two IF2DWTA chips of the above discussed multi-chip neuromorphic hardware setup to construct the calling song recognition network of crickets. The size of the network implemented using this architecture can be scaled-up with the latest CMOS technology. The new fabrication technologies offer more design area in silicon given the size of the transistors are small. Nevertheless, the multi-chip architecture is useful to design small-scale networks to emulate neo-cortical computations. The available hardware offers real-time and parallel operation of such network implementations.

3.9 CONCLUSION

In this chapter, we covered a wide range of literature beginning from a single transistor to existing neuromorphic systems. We also discussed the basic analog building blocks used in the design of silicon synapses and neurons. These building blocks are used in the design of our neuromorphic STP circuits. We will discuss in detail about these circuits in the next chapter. From the responses of the silicon synapses and neurons, we can deduce that it is possible to emulate the biologically realistic computations, thanks to the sub-threshold operation of a transistor. The exponential characteristics of the sub-threshold domain are useful in modeling the neural dynamics inspired by the biology. The neuromorphic chips designed using the sub-threshold transistors offer low-power consumption and real-time operation. Despite the presence of device mismatch effects, we showed in this research, that it is possible to design a network by tuning the responses at a single neuron level, using our sub-threshold neuromorphic hardware. We will discuss the implementation of this network in Chapter 5.

4

NEUROMORPHIC DESIGN OF SHORT-TERM PLASTICITY CIRCUITS

4.1 INTRODUCTION

We discussed in Chapter 1 that Short-Term Plasticity (STP) serves several functional roles at the synapse. STP dynamically influences the effect of a pre-synaptic neuron on its post-synaptic targets. The resulting synaptic filtering properties lead to a plethora of computational primitives in neural systems such as burst detection, transient enhancement, adaptation to sustained stimulus, synchrony detection and many more [2]. Specialized circuit designs are required to embed these useful computational primitives into large-scale neuromorphic hardware. Relatively few attempts have been made in this direction. Rasche and Hahnloser proposed a short-term adaptation circuit in 2001 [91]. It is an analog circuit, and it is most commonly used in the sub-threshold neuromorphic chips. The circuit outputs the synaptic weight in the form of an analog voltage in response to the input pulses. The output voltage (weight) is decreased during the input pulse and recovers toward a resting value in between the pulses. This circuit was further analyzed theoretically by Boegerhausen and colleagues in 2003 [15]. This STD circuit is implemented in our neuromorphic chips, and this forms the foundation of the design of our STP circuits. However, the circuit has few limitations, and we will discuss them later in this chapter. We proposed novel STP circuits to overcome the limitations. This chapter aims to present the design and the analysis of the proposed STP circuits. We will start with an overview of device level implementations of the STP. Later, we will move on to the circuit level STP implementations.

The inherent characteristics of the device (device physics) are exploited in device level STP implementations. Regarding emerging technologies, the use of memristors has raised interest in the neuromorphic community since the pioneering work of Leon Chua [24]. A memristor is a resistor with the memory. The resistance of the memristor can be modified based on the biases applied to its terminals. A memristor can be treated as a non-volatile analog memory device. It is suitable for low-power nano-scale integration. Memristors are the ideal candidates for building synapses, as they offer permanent storage of weights and plasticity of synapses in the neuromorphic systems. Several memristor based STP circuits have been recently proposed [20, 111, 98, 63, 4]. We will discuss a few of them in the following.

Memristors proposed in [20] are based on the movement of oxygen vacancies in a thin film of Tungsten oxides (WO_x). The oxygen vacancies form the conducting channels. By redistributing these channels, the conductance of the device can be changed. The conductance of a memristor is increased by repeatedly providing input pulses to the memristor. The conductance decays spontaneously in between the pulses. The change in the conductance can be comparable to the difference in the release probability of neurotransmitters that occurs during the STP. The conductance quickly reaches its maximum in response to the high-frequency stimuli. Therefore, the memristor responds to input pulse frequencies with high-pass filter characteristics similar to the STF (as discussed in Subsection 2.3.1). Another property of the memristor is that a repeated stimulation increases the retention time of the conductance change. Hence, short-term memories can be easily converted to long-term memories.

Memristors presented in [111] are based on the diffusion of oxygen ions from oxygen-deficient regions to oxygen-rich regions in an amorphous Indium Gallium Zinc oxide ($InGaZnO$) electrode. The top and bottom electrodes of the memristor are considered as pre-synaptic and post-synaptic neurons. The conductivity of the device is taken as the synaptic weight. Pulses with large amplitude and long duration cause a significant rise in the conductivity of the device. The conductance non-linearly decays in between the pulses. Learning mechanisms such as the Spike-Timing Dependent Plasticity (STDP) and the STP are demonstrated using this device.

Nano-ionic memristive devices called electrochemical capacitors are proposed in [63]. The top electrode of these devices is created by depositing reactive metals on top of the titanium-di-oxide TiO_2 layer. The bottom electrode is made up of an inert platinum (Pt). The device exhibit STF characteristics in response to a repeated stimulation of triangular pulses at the top electrode (in negative polarity). The device shows the STD dynamics when the polarity of the stimulus is reversed. Despite the STP characteristics, the physical phenomenon generating this behavior in these devices is unclear in the literature.

More recently, switched-capacitor based STP circuits were proposed by Noack and colleagues [82, 83]. As the name suggests, the switched-capacitors consists of capacitors and switches. They are commonly used in the signal processing domain. They are also used in the neuromorphic research [109, 35, 36] by exploiting their capability of implementing state-driven models in a mixed-signal realization. For instance, the value of the model variables is stored as a charge on a capacitor. The charge can be maintained by decoupling the capacitor from the actual circuit (open switch), thereby minimizing the leakage currents. The charging and the discharging of the capacitor (closed switch) are used to update the value of the model variables. This approach has also been used in the implementation of the STP models. In contrast to the standard neuromorphic implementations (e.g., [91]) these switched-capacitor based circuits consist mostly of digital building blocks. They support the use of standard digital synthesis and rapid prototyping. Large time-constants are achievable using these circuits, given that the leakage currents are negligible compared to the operating currents. Furthermore, the switched-capacitors impose the need for additional design techniques to minimize the leakage effects, especially when deep-sub-micron technologies are used to fabricate them.

In this research, we aim to empower the neuromorphic systems, as described in [23], with the short-term filtering properties. Given a specific system choice, the analog sub-threshold approach used by Rasche and Hahnloser [91] is best suited in our case. Like digital systems, switched-capacitors based circuits follow the building-block approach using standard elements such as amplifiers, switches, etc. These circuits can, therefore, be easily ported to small-scale technologies as soon as they are available, as opposed to full-custom analog VLSI circuits which require extensive re-design. Unfortunately, switched-capacitors are not a good choice for the integration with sub-threshold circuits due to the high impact of leakage currents in this operating regime of transistors. On the other hand, memristors show promising behavior given their size (few nanometers). However, there is no memristor compatible CMOS fabrication process commercially available yet, thereby making their immediate use unfeasible.

The STD circuit proposed in [91] is compatible with the design of sub-threshold neuromorphic circuits. But, the circuit lacks a tunable recovery rate of its output voltage (synaptic weight). As explained in the subsequent section, the rate of recovery of the output voltage is dynamic, and it depends on the instantaneous synaptic weight. This limitation imposes strong constraints on the attainable temporal dynamics and hinders a more general use of the circuit as a module for implementing a plethora of STP features. For example, strong synaptic depression followed by fast recovery cannot be achieved in this context. This temporal dynamic is essential to implement the burst-detection property of the STD. That is, when a burst of pulses is used to stimulate a STD synapse, the initial weight of the synapse causes a high post-synaptic current to mark the onset of the bursts. During the burst, the synaptic weight is reduced due to the

STD. As a result, the post-synaptic current becomes small, and no spike is generated at the post-synaptic neuron. The response to the onset of the next burst is mediated by a fast recovery of the synaptic weight during the inter-burst interval. In this way, the STD synapse can identify the onset of the bursts of spikes (digital pulses).

The novel circuits proposed in this research extend the state-of-the-art of STP neuromorphic circuits. We provided an external control over the recovery rate of the output voltage (synaptic weight) in the absence of input pulses. The cost of the improved functionality is low, given that the circuit designed in [91] and the STD circuit proposed here share the same number of transistors. Furthermore, the pMOS transistor of the circuit in [91] occupies a large silicon area because the bulk is not connected to the V_{dd}. A separate 'n'-well is needed to provide the bulk voltage. However, we minimized the design area of our STP circuit, by connecting the bulk to the V_{dd}.

4.2 ANALOG SUB-THRESHOLD NEUROMORPHIC STD CIRCUITS

Now we move on to the circuit level implementations, in which circuit blocks are designed to implement the STP dynamics. To faithfully highlight the novelty of our circuits, we must first describe the prior comparable works. Given a choice reported in the above section, we restrict our literature review to sub-threshold analog circuits that implement the synaptic STP. The sub-threshold STP circuits are having a high impact on this research, are discussed in the following.

A set of STD and STF circuits were presented in [60]. These circuits receive input currents from various synapses and output the scaled currents. We will discuss the operation of the STD circuit only, as the function of the STF circuit is analogous to its counterpart. The schematic of the STD circuit is shown in the Fig. 24. The input to the circuit is the sum of synaptic current I_i and the output of the circuit is the scaled version of the current I_o. An external digital pulse (inverted) turns on the circuit. During the pulse, the capacitor is charged at a rate set by the difference between the currents through the transistors with bias voltages V_u and V_d. In between the pulses, the capacitor is discharged at a rate set by the current through the transistor with the bias voltage V_d. The voltage across the capacitor is called the depression voltage. The depression voltage is supplied to the gate of the pMOS transistor which is connected in series with the current-mirror circuit (see Sec. 3.4). The depression voltage increases the source voltage of the output transistor of the current-mirror, thereby resulting in a reduced output current. The working principle of the STF circuit is similar to that of the STD circuit, except for the different scaling factor. The STF circuit has an addi-

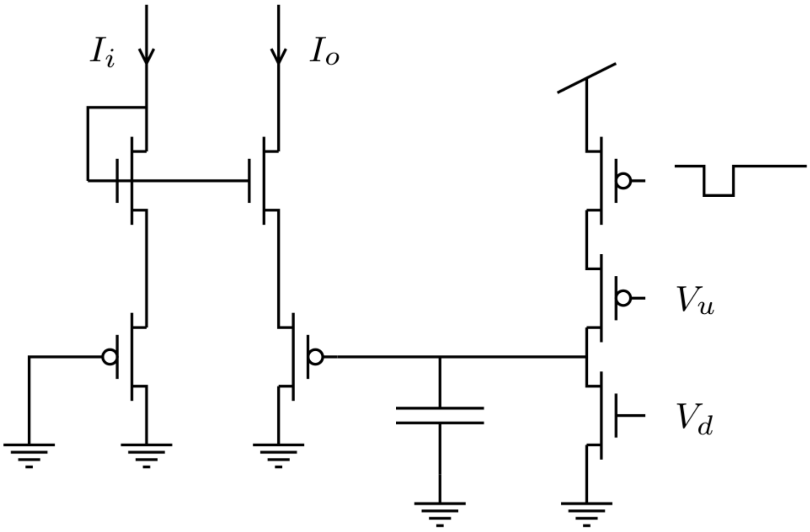


Figure 24: Schematic of the sub-threshold neuromorphic STD circuit proposed in [60] consisting of seven transistors and a capacitor. I_i is the sum of synaptic currents provided as input and I_o is the scaled output current. Digital input pulses (inverted) are applied to turn on the depression branch. During the pulse, the capacitor is charged by the difference between the currents set by the transistors with the bias voltages V_u and V_d. In between the pulses, the capacitor is discharged by the current set by V_d. The voltage across the capacitor called the depression voltage is supplied to a pMOS transistor, which is connected in series with the current-mirror. The depression voltage increases the source voltage of the current-mirror output, resulting in a reduced output current.

tional nMOS transistor placed in between the transistors with V_u and V_d voltages. This transistor limits the maximum scaling of the output current. The circuit is capable of reaching time-constants in the order of seconds. It can also be integrated with our DPI synapse with minor modifications in the design, for instance, by replacing the synapse's weight branch. However, the layout size of these STP circuits is big, considering the total number of transistors. It is true especially in case of the pMOS which occupy more space than the nMOS transistors. Better results regarding compactness and power consumption, can be achieved by designing a STP circuit, which is small in size and compatible with the DPI synapse.

A floating-gate based stochastic synaptic circuit was proposed in [112]. Floating-gates are widely used to implement digital memories (as for commercially available flash memories). Floating-gates are CMOS transistors with an electrically isolated (floating) gate. The common gate of the CMOS is referred to as 'control gate' in these devices. A positive charge at the floating-gate of an n-type transistor results in the channel formation between the source and the drain, whereas a negative voltage suppresses the channel formation. Two mechanisms can modify the charge of the floating-gate: tunneling and injection. Tunneling is a quantum-mechanical process during which an electric field is applied to ensure the electric charge dig through the potential barrier (between the floating-gate and the control gate). Hot-electron injection is the phenomenon in which the electric charge gains sufficient kinetic energy to inject over the surface of the potential barrier. Capacitors are used to access the floating node, and there is no discharge path for the current, thereby providing an extended charge-retention time. The floating-gates can also be used as analog memories, and several research groups have been proposing their integration with neuromorphic circuits [10, 16].

Peng and his colleagues proposed a STP circuit by using the floating-gate. For every input pulse, the capacitor (which is not a part of the floating-gate) is discharged by a sub-threshold current set by the weight control transistor. Therefore, the voltage across the capacitor drops. In between the pulses, the voltage recovers with an exponential recovery dynamics implemented by a resistor and a floating-gate circuitry. The charge at the capacitor is not fully restored when the input pulses arrive faster than the charge rate of the capacitor. As a result, the voltage across the capacitor is recovered to a smaller value than the resulting value from the previous spike. In this way, the circuit emulates the STD dynamics through the capacitor voltage. The circuit is compact regarding transistor count. However, the resistor is implemented by a sub-threshold pMOS transistor operated in the ohmic region. The operating range of the sub-threshold transistor is minimal in the ohmic region. Therefore, the biases should be precisely tuned to obtain the required recovery dynamics. Furthermore, high voltages are used for tunneling and injection at the floating-gates, which requires a dedicated power supply. The floating-gates are bulky in size compared to the conventional CMOS transistors due to the added floating-gates in their design which are used to store the charge for an extended period. All these constraints hinder the use of these circuits in large synaptic arrays of neuromorphic systems.

A depressing synapse circuit was proposed in [29]. The circuit is designed by utilizing the parasitic capacitance. Parasitic capacitance is usually an unwanted capacitance that exists within and between the components due to their proximity to the design. In this circuit, a parasitic capacitor is used to store the charges, whose amount varies during the operation of the circuit. The circuit consists of three nMOS transistors connected in series and a diode-connected pMOS transistor (see Sec. 3.4). The input pulses are provided to the first transistor. During the input pulse, the charge stored be-

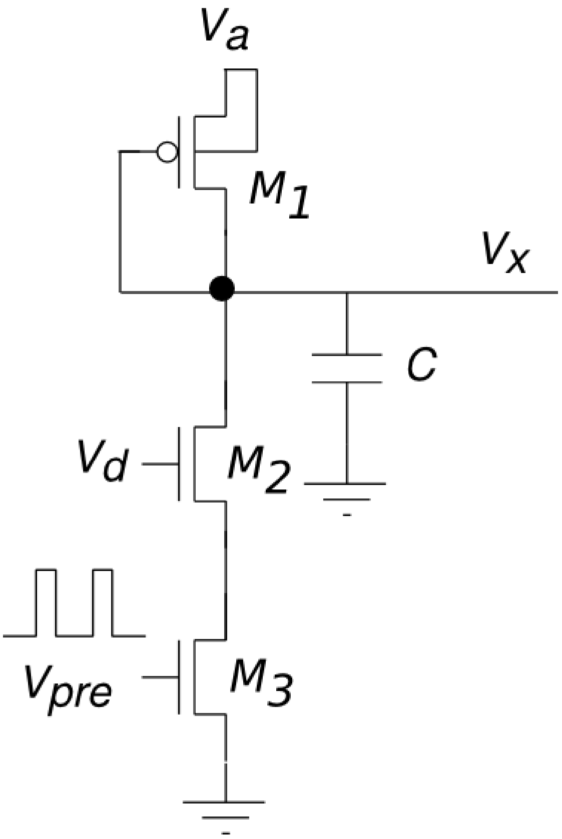


Figure 25: Schematic of the neuromorphic STD circuit proposed in [91] is shown. The circuit has two nMOS transistors, a diode-connected pMOS transistor and a capacitor C. Input pulses are applied at the gate of the transistor M_3. The bias voltage V_d sets the current flowing through the transistor M_2 in presence of an input pulse, therefore setting the discharge rate of the capacitor and consequent drop of the output voltage V_x. During the inter-pulse intervals, V_x recovers toward its resting value at a rate set by V_x itself due to the current flowing through the transistor M_1.

neath the gate-oxide of the second transistor (connected in series with the first transistor) is transferred to the parasitic capacitor (existing at the node of the diode-connected transistor and the first transistor). In between the pulses, the charge beneath the gate of this transistor is restored at a rate set by the third transistor (connected in series with the second transistor). For fast input pulses, the charge at the parasitic capacitor is incompletely restored resulting in a reduced voltage. Whereas for slow input pulses, the charge replenishment is complete. This mechanism is analogous to the neurotransmitter release in depressing synapses. The advantage of this circuit is the absence of a physical capacitor leading to a reduced silicon area. The parasitic capacitance is minimal and only fast time-constants in the range of 2-4 ms can be achievable with this circuit. However, the time-constants observed in biology are in the range of 50-200 ms [89]. This constraint limits the use of this circuit in biologically plausible implementations.

Rasche and Hahnloser [91] proposed a sub-threshold neuromorphic implementation of the STD model described in [2]. As already discussed in Sec. 2.2, this model captures the biologically realistic cortical gain-control properties of STP, and it has been widely used in computational neuroscience [1]. The schematic of the STD circuit presented in [15] is shown in Fig. 25. The circuit consists of two $nMOS$ transistors, one $pMOS$ transistor, and a capacitor. Digital pulses are provided as the input to the circuit. The output is the analog STD voltage measured across the capacitor. The output voltage of the circuit is depressed for the duration of digital pulses and attempts to recover in between the input pulses. The neuromorphic STD circuit proposed in [91] is compatible with the DPI synapse circuit (described in detail in Sec. 3.5) and it is implemented in the neural arrays of our neuromorphic chips (see Sec. 3.8). We discuss this circuit in detail in this section, as it forms the foundation of our STP circuit design.

The digital input signal V_{pre} is applied to the gate of transistor $M3$ which acts as a switch. When the digital pulse reaches its maximum amplitude, the switch $M3$ closes. The difference between the sub-threshold currents through the transistors M_1 and M_2 discharges the capacitor C. The consequent drop in the output voltage V_x depends on the bias voltage V_d, the duration of the pulse, the capacitance C and on itself (V_x). Therefore, for a fixed pulse duration and a capacitance, the bias voltage V_d controls the strength of the depression. In between the input pulses (i.e., the digital pulse reaches its minimum amplitude), the switch $M3$ is opened. The output voltage V_x tries to recover towards its resting voltage. The transistor M_1 is diode-connected (gate shorted to its drain terminal, see Sec. 3.4 for more details). The diode-connected transistor always operates in saturation ($V_{sd} \geqslant 4U_T$) as long as sufficient current flows through it. Therefore, the drain voltage V_x of M_1 is always smaller than the source voltage V_a due to its saturation region of operation. In other words, the output voltage V_x stays lower than the voltage V_a during the resting state. The rate to reach the resting voltage state is determined by the current flowing through the transistor M_1, which is an exponential function of the voltage V_x (the gate voltage of M_1). Therefore, this rate varies depending on the V_x value.

Next, let us discuss the circuit response to the pulse-train input. In [15], the transient response of this STD circuit was characterized for three values of the bias voltage V_d, as shown in blue, magenta and black curves in Fig. 26. The strength of depression marked by the size of the voltage drop is small for a small V_d (blue) and high for a large V_d (black). It can be seen for all three simulations, the output voltage V_x decreases with each input pulse and tries to recover towards its resting value in between the pulses. For $Time \leqslant 0.06s$, the recovery is slow when the output voltage V_x is close to V_a due

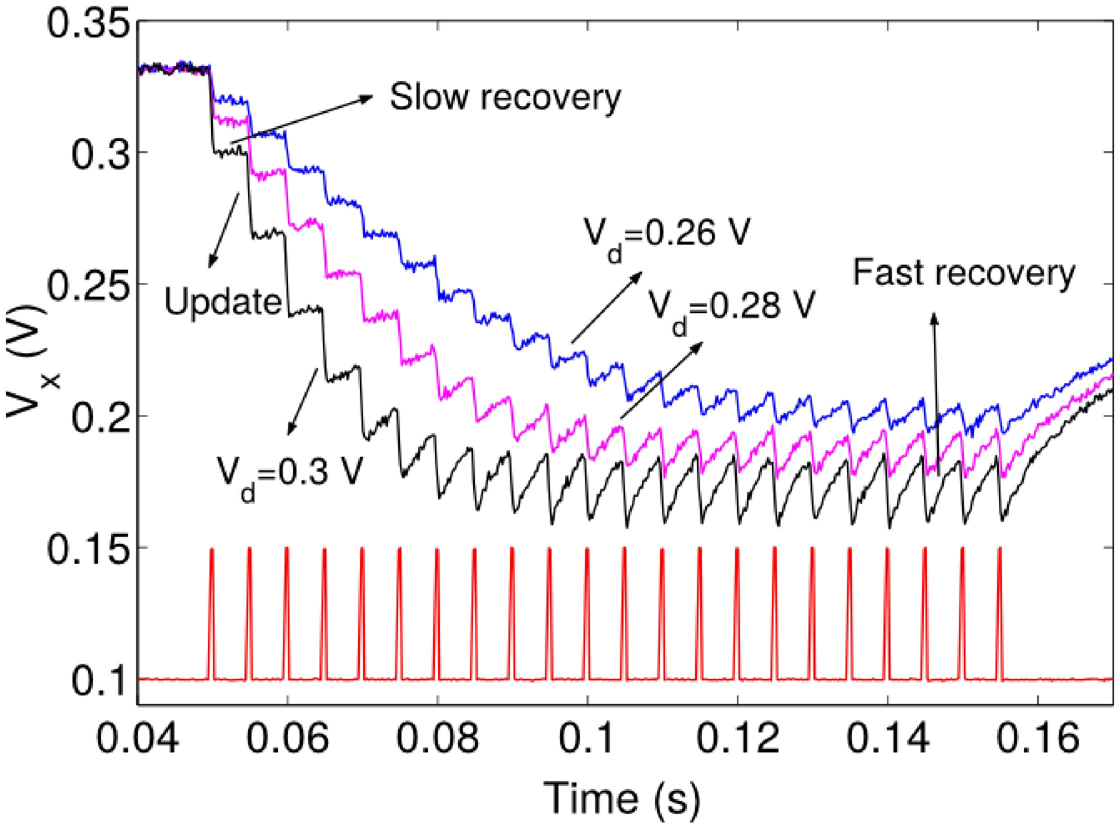


Figure 26: Output voltage responses of the STD circuit (blue, magenta and black traces) presented in [15] for different values of the depression (strength) voltage V_d. The red trace shows the input pulse train stimulus provided as V_{pre} to the STD circuit. The voltage V_d determines the amplitude of the output voltage V_x depressed during the pulse. In all blue, magenta and black traces, the recovery rate of the output voltage V_x towards its resting value is slow in between the first few input pulses for $\mathrm{Time} \leqslant 0.06$s. Over repeated stimulation for $\mathrm{Time} \geqslant 0.1$s, the V_x recovers faster with non-linear temporal dynamics.

to the small V_{gs} on transistor M_1. For $\mathrm{Time} \geqslant 0.1$s, the recovery becomes faster when V_x moves away from V_a (bigger V_{gs} on transistor M_1). The non-linearity arises due to variable regions of operation of the transistor M_1. For a large V_{ax}, the transistor M_1 shifts from a non-saturated region to a saturated region of the sub-threshold operation. In all three simulations, the output voltage reaches the steady-state for $0.14 \leqslant \mathrm{Time} \leqslant 0.16$s. It is important to note that the steady-state voltage is different from the resting-state voltage. The output voltage remains resting (close to V_a) when there is no input pulse. However, when there is a continuous stream of input pulses, the output voltage drops continuously and reaches the saturation value which is the steady-state value. For $\mathrm{Time} \geqslant 0.16$s, it is shown that the output voltage V_x attempts to recover towards to its resting value in the absence of input pulses.

The advantage of this circuit is its compact design with just three transistors and only one capacitor, which is one of the space-consuming components during the fabrication of the circuit. Biologically realistic time-constants in the order of tens of milliseconds to hundred milliseconds can be achievable with this circuit. The steady-state responses of the STD circuit are characterized for various input frequencies [91]. The stimulus consists of an input pulse train. The pulse frequency is varied from 1 to 200 Hz and presented to the circuit. The steady-state amplitudes of the output voltage are analyzed. The results suggest that the STD circuit exhibit low-pass filter characteristics analogous to the results presented in [107] (see Sec. 1). The same STD circuit responses are quantitatively compared to the Abbot's STP model as well in [15].

The drawback of the above mentioned neuromorphic STD circuit (schematics is shown in Fig. 25) from [91] is the lack of control over the rate of recovery of depression voltage, in between the input pulses. It is the design constraint because the transistor M_1 responsible for controlling the recovery rate is diode-connected. This limitation prevents the supply of an external bias voltage to the gate terminal of this transistor. In this case, the output voltage V_x controls the gate. Therefore, the output voltage V_x itself determines the transistor's current value, which controls its rate of recovery.

For example, it is evident from the Fig. 26, that the output voltage V_x recovers at a faster rate if the difference between the voltages V_x and V_a is high, in comparison to the recovery rate when V_x is closer to V_a. The circuit misses the precise control over its recovery, which might be useful in certain applications (See Sec. 5.3.2) that demand a strong depression followed by a fast recovery (for example, the STD model presented in [96]). On the other hand, these temporal dynamics are achievable using the phenomenological models of the STP [2, 71]. Therefore, the need for a new design of the neuromorphic STD circuit is irresistible.

So far, we only focused on the Short-Term Depression (STD) circuit. We will continue by discussing the implementation of the Short-Term Facilitation (STF) (another type of Short-Term Plasticity (STP)) in neuromorphic circuits.

A complementary variant of the STD circuit is proposed in [92] in which the STF circuitry is integrated together with the CMI synapse (see Sec. 3.5 for more details about CMI synapse), in their design. However, in the latest versions of the neuromorphic chips, the CMI synapses are replaced by the DPI synapses. This synapse can be tuned to operate in a specific regime to implement the STF as discussed in the last chapter (see Sec. 3.5). When the DPI synapse is biased with a small weight voltage and a slow time-constant voltage for a fixed threshold voltage, the output current becomes smaller than the threshold current. In this condition, the output current (EPSCs) builds-up slowly in response to the input spikes, resulting in the STF dynamics. Though the STF dynamics can be implemented by the DPI synapse, a dedicated STF circuit allows the DPI synapse to operate with a time-constant independent from the time-constant of the STF circuit. It allows more flexibility in tuning to implement both STD and STF with different time-constants at the same synapse.

To tackle the issues mentioned above, we designed a set of STD and STF circuits based on the design of [91], and we published in [90]. In our circuits, we implemented an independent control over the recovery rate for the output voltage in between the input pulses. The output voltage of the newly designed STF did not always reach the steady-state value, in response to the pulse train stimulus. Therefore, another set of STP circuits are designed with a negative feedback loop. All these circuits will be discussed in the next sections.

4.3 STP CIRCUITS WITH RECOVERY CONTROL

In this section, we will discuss our design of the neuromorphic STP circuits with a tunable recovery rate of the depression/facilitation output voltage.

4.3.1 *STD circuit*

Let us start by discussing the design of the STD circuit. Fig. 27 shows the schematic of the STD circuit (published in [90]). The STD circuit has a similar structure of the STD circuit proposed in [91], except for the diode-connected transistor. The circuit consists of three transistors and a capacitor. The digital input pulses V_{pre} are provided to the gate of the transistor M_1. These digital pulses represent the digital output spikes of the pre-synaptic neuron. The digital pulse turns ON the transistor M_1 for the duration of the pulse, which creates a path through the transistor M_2, for the capacitor C_w to charge. During the pulse, the capacitor C_w is charged at a rate set by the difference between the currents through M_2 and M_3 transistors, resulting in a decrease of the output voltage V_{out}. The gate voltages V_{wei} and V_{tau} determine the currents through M_2 and M_3. The capacitor is discharged in between the pulses at a rate set by the current through M_3, which is controlled by the gate voltage V_{tau}. Therefore, the output voltage V_{out} recovers with the tunable recovery rate, towards its initial value set by the voltage V_{up}.

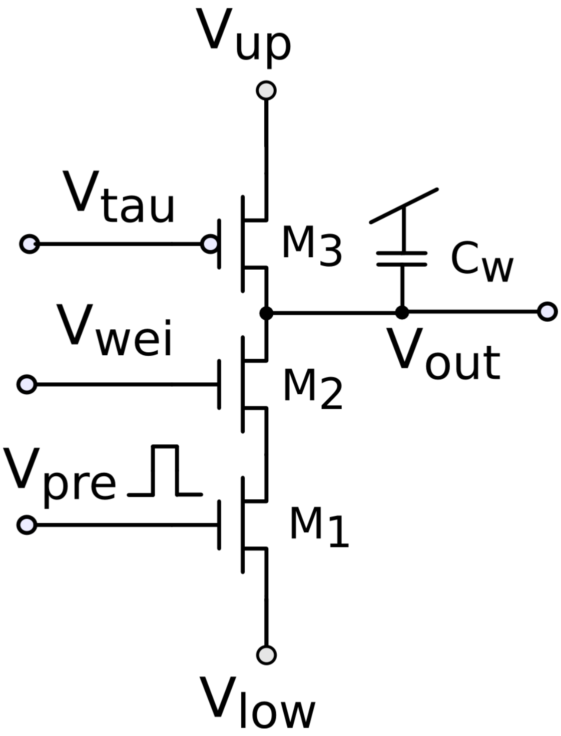


Figure 27: Schematic of a simple STD circuit (published in [90]) is shown. The circuit consists of two nMOS transistors, one pMOS transistor, and a capacitor. The digital input pulses V_{pre} are provided to the gate of the transistor M_1. The digital pulse turns ON the transistor M_1, which creates a path for the capacitor C_w to charge. The capacitor C_w is charged at a rate set by the difference between the currents through M_2 and M_3, thereby, decreasing the output voltage V_{out}. The capacitor C_w is discharged in between the pulses at a rate set by the current through M_3, which is controlled by the gate voltage V_{tau}. The output voltage V_{out} recovers with the tuned rate, towards its initial value set by the voltage V_{up}. The voltage V_{up} sets an upper boundary to the depression voltage to remain within the sub-threshold region, to bias a sub-threshold voltage to the weight transistor of the synapse, that follows the STP circuit in the design (not shown). The voltage V_{low} sets the lower boundary to the depression voltage to prevent the synapse from turning OFF.

The operation of this circuit is analogous to the working of the circuit proposed in [91]. The difference is that the diode-connected transistor of the STD circuit in [91], is replaced with a regular pMOS transistor M_3. This design allows the user to tune the external bias voltage V_{tau} (applied to the transistor gate terminal), which controls the rate of recovery. The voltages V_{up} and V_{low} set the upper and the lower limits of the output voltage. The upper voltage limit is necessary to bias a sub-threshold voltage to the weight transistor of the DPI synapse, which follows the STP circuit in the design (not shown). The lower voltage limit sets the lower-limit of the steady-state output voltage.

We derive the circuit response in the following, for a better understanding of the circuit operation. During the stimulus onset, it is safe to assume that the $nMOS$ transistor M_2 operates in a sub-threshold saturated region for the given V_{up} and V_{low} voltages.

At the point of arrival of the first input pulse, we can assume that $I_{M_3} << I_{M_2}$. Then the output voltage becomes:

$$C_w \frac{V_{out}(t)}{dt} = -I_{n0} e^{\frac{\kappa n V_{wei} - V_{low}}{U_T}} \tag{22}$$

After the pulse, only the positive current is active at the output voltage node. Therefore, the output voltage is driven towards its resting state:

$$C_w \frac{V_{out}(t)}{dt} = I_{p0} e^{\frac{-\kappa p V_{tau} + V_{up}}{U_T}} \left(1 - e^{\frac{V_{out}(t) - V_{up}}{U_T}}\right) \tag{23}$$

The proposed STD circuit is designed using the standard CMOS Austria Micro Systems (AMS) 350 nm technology. The transient responses of the circuit are characterized using the Spectre® simulator. To analyze the temporal dynamics of the depression voltage of the circuit, we performed two experiments with the STD circuit. During the first experiment, the depression voltage is plotted for various sub-threshold values of the voltage V_{wei}, for a given input pulse train and a constant V_{tau}. In the second experiment, the depression voltage is plotted for various sub-threshold values of the voltage V_{tau} for the same stimulus and a constant V_{wei}. The stimulus consists of a train of input pulses (not shown) of 100 μ seconds in pulse-width and 100 Hz in frequency. The stimulus is provided for a 0.5-second duration. The total simulation run-time is 1 second. The pause of 0.5 seconds after the stimulus duration allows the output voltage to recover completely within the simulation time. The parameters of the STD circuit are chosen such that the weight transistor of the following DPI synapse (not shown) operates in the sub-threshold region. The transistor source voltages (the limiting voltages) to the circuits, V_{up} and V_{low}, are set to 0.8 V and 0.4 V respectively.

The output voltage responses of the STD circuit for both the experiments are shown in Fig. 28. The top plot shows the output voltage responses to the V_{wei} voltage sweep. The bottom plot shows the output voltage responses to the V_{tau} voltage sweep. The V_{wei} changes the amplitude of the update of the output voltage V_{out} during the input spike

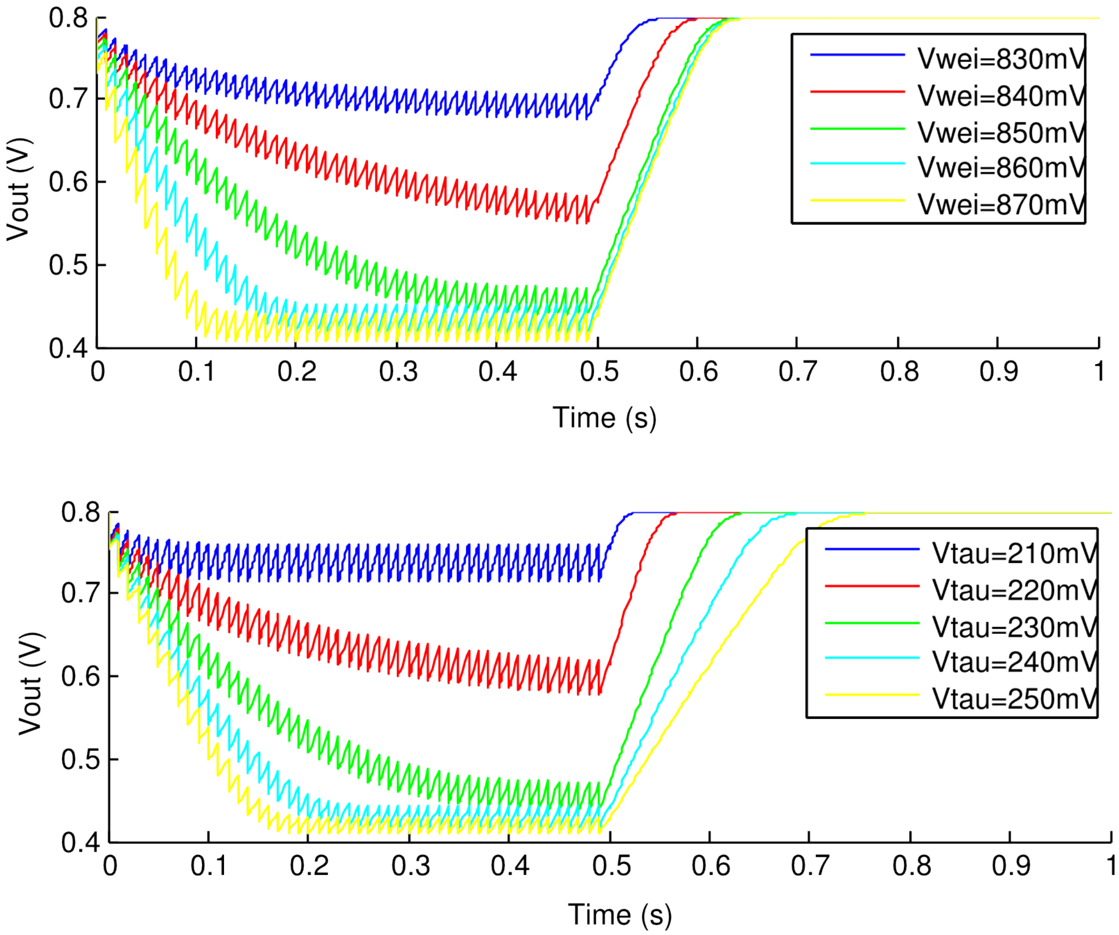


Figure 28: The output voltages of the proposed STD circuit of Sec. 4.3 (published in [90]) in response to the input pulse train stimulus (not shown). The stimulus presented is similar to the one shown in Fig. 26. The top plot shows the output voltages in response to the sweep of the V_{wei} voltage. The bottom plot shows the output voltages in response to the sweep of the V_{tau} voltage. The parameter values are chosen from the sub-threshold region. In the top plot, the size of the update of the depression output voltage V_{out} increases (from blue to yellow curve), in the order of increasing V_{wei} voltage. In the bottom plot, the speed of the recovery of the output voltage V_{out} drops (from blue to yellow), by increasing the V_{tau} voltage. In both the plots, the output voltages reach the steady-state values for the given parameter sets within the stimulus duration, except for the ones shown in red and green.

(not shown), in the order of the increasing magnitude of V_{wei} voltage (blue to yellow), as seen in the Fig. 28 (top). The V_{tau} voltage speeds up the recovery of the output voltage V_{out} in between the pulses (not shown), in the order of the decreasing V_{tau} voltage (blue to yellow), as seen in the Fig. 28 (bottom). Please note that V_{tau} is the gate voltage of the $pMOS$ transistor. When no stimulus is presented ($Time \geqslant 0.5s$), the output voltage V_{out} recovers toward to its initial value V_{up} completely (blue to yellow curves, in both the experiments). All output voltages except the 'red' in both the tests, reach the steady-state values (see $\mathrm{Time} \geqslant 0.4s$) for the given choice of the parameters. Reaching the steady-state condition is crucial for the circuit because it sets the maximum amount of synaptic depression for the synapse.

The steady-state condition is reached when the charge accumulated during a spike is equal to the charge removed during the inter-spike interval. It occurs due to the following reasons:

I. The voltage V_{ds} of the transistor M_2 decreases, reducing the charging current.

II. The voltage V_{ds} of the transistor M_3 increases, increasing the discharging current.

In [91], a non-linear p-type diode-connected transistor (the bulk of which is connected to its source) is used for controlling the temporal dynamics of the STD. In our design, we replaced the diode-connected transistor by a conventional transistor with a tunable gate voltage. This simple design modification offers complete control over the temporal dynamics and allows the circuit to perform specific computations which were not possible using the previous design. For example, the recovery of the STP can be set at a rate independent from the strength of the STP. Our design is compact compared to the previous design, which required a separate n-well for the bulk. Though not all output voltages (red curve in Fig. 28:top and bottom) of the STD circuit reached the value within the stimulus duration, the circuit can still be used as the STD circuit alongside the DPI synapse, considering the compactness of the design. We solved this minor limitation by adding a negative-feedback control to this circuit. We will discuss the details of this modified design in the next section.

4.3.2 *STF circuit*

Next, we discuss the design of the STF circuit. The STF circuit shown in Fig. 29 is the complementary version of the STD circuit (published in [90]). The $nMOS$ transistors of the STD circuit are replaced by the $pMOS$ transistors and vice-versa. The circuit consists of two $pMOS$ transistors, one $nMOS$ transistor, and a capacitor. The circuit has an additional circuit (inverter) to invert the digital input pulses. The presence of multiple $pMOS$ transistors makes the STF circuit bulky concerning the silicon area occupied compared to the STD circuit, which has only one $pMOS$ transistor.

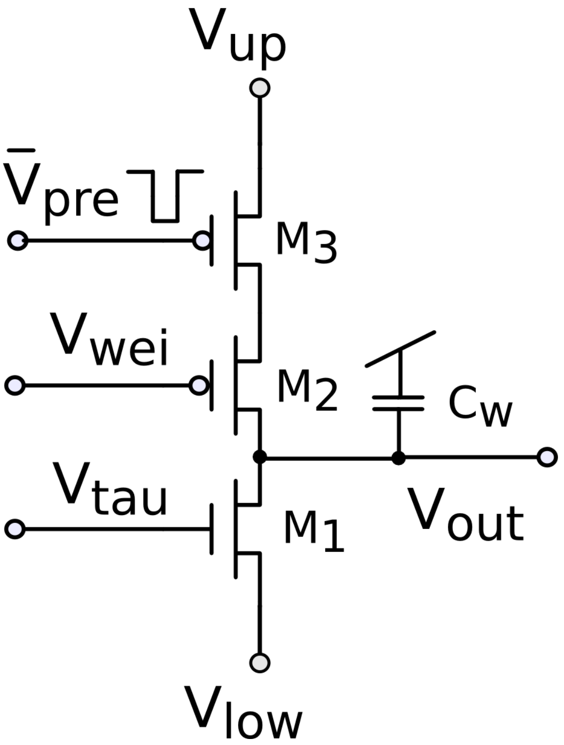


Figure 29: Schematic of the proposed STF circuit (published in [90]). Inverted pulses are sent as input to the STF circuit. During the inverted pulse, the capacitor C_w is discharged at a rate set by the difference between the currents through the transistors M_2 and M_1, thereby increasing the output voltage. In between the inverted pulses, the capacitor C_w is charged at a rate set by the current through the transistor M_1. Therefore, the output voltage recovers towards its resting voltage value V_{low}. A high value of the voltage V_{up} is supplied to operate the pMOS transistors in sub-threshold saturation region. A small value of the voltage V_{low} is supplied to keeps the following DPI synapse (not shown) always ON.

The operation of the STF circuit will be discussed now. The inverted input pulses provided to the gate terminal of the transistor M_3, turns ON the path for the capacitor C_w to discharge. The capacitor is discharged through the transistors M_2 and M_3, during the onset of the inverted pulses. The discharge rate is set by the difference between the currents through M_2 and M_1. During the onset of the inverted pulse, we can assume that the transistor M_2 operates in a sub-threshold saturated region by adequately setting the voltage V_{up} to a high value. The voltage V_{up} of the STF circuit is much larger than the voltage V_{up} of the STD circuit to operate the pMOS transistors in the subthreshold region. The capacitor is charged in between the inverted pulses at a rate set by the current through M_1, with the gate voltage V_{tau}. This time, the facilitation output voltage recovers towards its initial resting voltage value (close to V_{low}). The small value of V_{low} prevents the DPI synapse from being completely turned OFF.

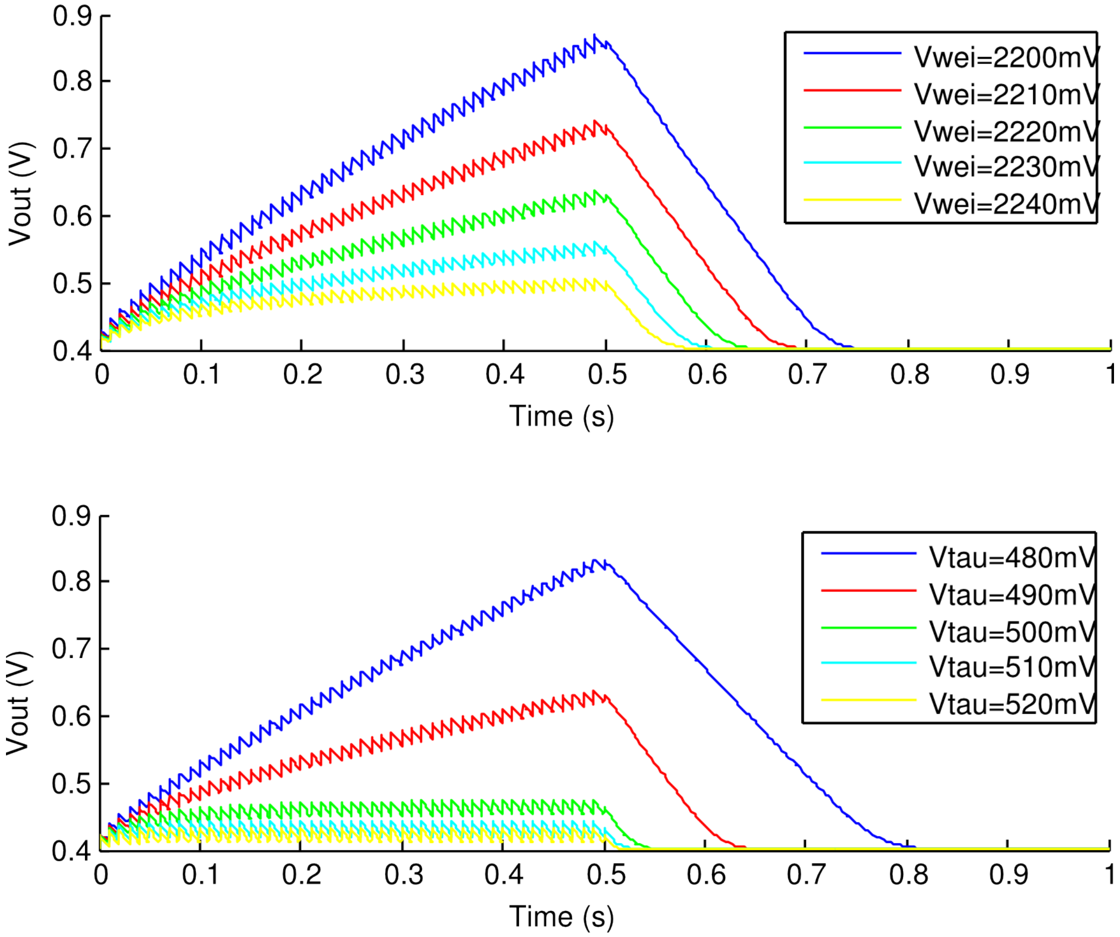


Figure 30: The output voltages of the proposed STF circuit of Sec. 4.3 (published in [90]) in response to the input pulse train stimulus (not shown). The top plot shows the output voltages in response to the sweep of the voltage V_{wei}. The bottom plot shows the output voltages in response to the sweep of the voltage V_{tau}. The voltage values are chosen from the sub-threshold region for both the p-type and the n-type transistors. In the top plot, the size of the update of the output voltage V_{out} decreases (from blue to yellow curve), in the order of increasing V_{wei} voltage (as the absolute voltage decreases). In the bottom plot, the speed of the recovery of the output voltage V_{out} increases (from blue to yellow), by increasing the V_{tau} voltage. In the top plot, no output voltage reaches the steady-state value within the stimulus duration. In the bottom plot, except for the voltages in red and blue, all the other output voltages reach the steady-state values.

The proposed STF circuit is designed using the standard CMOS 350 nm Austria Micro Systems (AMS) technology. The transient responses of the circuit are characterized using the Spectre® simulator. The same experiments as discussed in the previous subsection are carried out in the STF circuit as well. The circuit's output voltage responses are shown in Fig. 30. The top subplot shows the output voltage responses to the V_{wei}

voltage sweep. The bottom subplot shows the output voltage responses to the V_{tau} voltage sweep.

The amplitude of the update of the output voltage V_{out} increases (blue to yellow) during the pulse, in the increasing order of the V_{wei} voltage (blue to yellow), as shown in Fig. 30 (top). The speed of the recovery of the output voltage V_{out} increases (blue to yellow) in the increasing order of the V_{tau} voltage (blue to yellow), as shown in Fig. 30 (bottom). Unlike the STD, only a few of the output voltage responses reaches the steady-state values (yellow in top; yellow, cyan, and green in the bottom). The steady-state values are not achieved due to the linearity in the recovery of the output voltage resulting from the saturated transistors, which is explained in the following.

The output of the STF circuit has to be in the sub-threshold range to provide a sub-threshold bias to the input $nMOS$ weight transistor of the DPI synapse [9]. However, the upper limit voltage V_{up} is high (for $pMOS$) as mentioned earlier, so that the transistor M_2 is always saturated. Due to this condition, the output voltage of the STF circuit does not reach the steady-state voltage within the sub-threshold range for many of its parameters. If the transistor M_3 gets saturated after an inverted input pulse, and, if the charge per inter-pulse interval provided through the transistor M_2 is smaller than the charge per pulse removed through the transistor M_1, then the output voltage V_{out} cannot reach the steady-state value. To tackle this issue, we designed another set of the STP circuits by adding a negative feedback loop to the current STF design. This feedback loop allows the output voltage of the STF circuit to reach steady-state values. These feedback STP circuits are discussed in the next section.

Nevertheless, the STF circuit can still be used with different bias settings. For instance, the sub-threshold condition of the $pMOS$ weight transistor can be satisfied by shifting the boundary voltages V_{up} and V_{low} towards V_{dd}. With this setting, the circuit can be used to control the 'threshold' transistor of the DPI synapse which requires the high bias voltage to operate. In this case, the STF circuit is used as the STD circuit, that controls the 'threshold' transistor of the synapse. At the same time, a long-term learning circuit can be used to control the 'weight' transistor of the synapse. Thus, implementing two forms of synaptic plasticity at the same synapse is feasible with our STP circuits design.

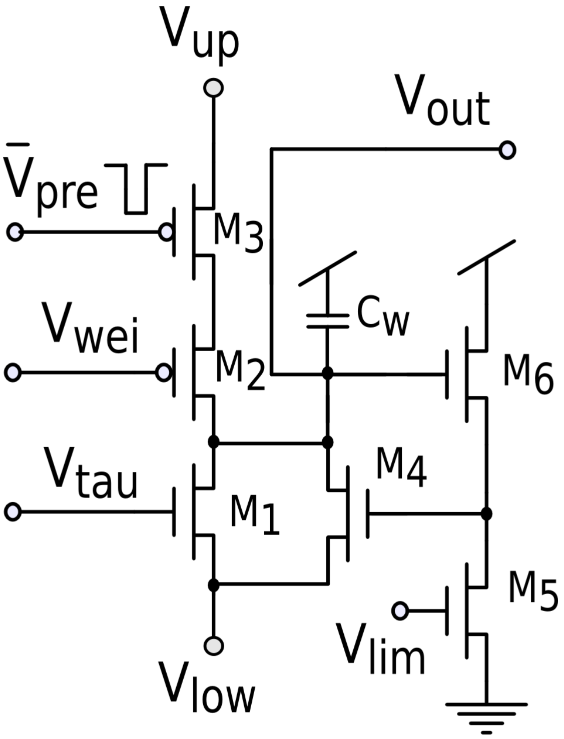


Figure 31: Schematic of the STF circuit with the feedback control (published in [90]). The design is an extension of the proposed STF circuit by adding a source-follower circuit. The output voltage V_{out} is supplied to the gate of the source-follower's input transistor M_6. The voltage V_{lim} is provided as the bias voltage to the source-follower's transistor M_5. The output voltage of the source-follower is connected to the gate of the (additional recovery) transistor M_4, which is connected in parallel to the transistor M_1. The voltage V_{lim} affects the source-follower output voltage (feedback-control voltage) depending on the facilitation output voltage V_{out}. The feedback-control voltage provides an additional control over the recovery rate of the voltage V_{out} set by the current through the transistor M_4, in between the inverted input pulses.

4.4 STP CIRCUITS WITH FEEDBACK RECOVERY CONTROL

The problems for the circuits presented in Sec. 4.3 are more pronounced for the case of facilitation. Therefore we will start with the effect of the feedback on the STF circuit. Then, we will briefly discuss the STD circuit with feedback, to avoid redundancy in the text.

4.4.1 *STF circuit*

Fig. 31 shows the schematic of the STF circuit with a feedback control (published in [90]). Transistors of the previously discussed STF circuit did not always operate in the saturation region. Therefore the steady-state was reached only for a few parameter sets. A negative-feedback loop is used to tackle this issue. A source-follower circuit (refer Sec. 3.4 for more details on the source follower circuit) is added to the previous STF circuit design, implementing the feedback loop. The working of the STF circuit with a feedback control is similar to the working of the STF circuit without the feedback, except for the additional feedback loop. We covered the operation of the STF circuit in the last section. Therefore, we restrict our explanations to the influence of the additional negative-feedback loop on the STF circuit.

The bias voltages V_{up}, V_{low}, V_{wei} and V_{tau} are tuned to be the same values as discussed in the last section. The inverted-input pulses are provided to the gate of the transistor $M3$. The output voltage of the STF circuit is supplied to the gate of the source-follower's input transistor $M6$. The feedback-control voltage V_{lim} sets the voltage of the source-follower's bias transistor $M5$. The output of the source-follower is connected to the gate of the transistor $M4$. The transistor $M4$ offers an additional path for the capacitor charging current during the recovery of the facilitation output voltage in between the inverted input pulses. The voltage V_{lim} influences the recovery rate of the voltage V_{out} through the transistor $M4$, by discharging the capacitor at a rate depending on the output voltage V_{out}. This negative-feedback loop enables the output voltage to reach a steady-state value. A theoretical understanding of this circuit's working principle is discussed in the following.

The source follower of the feedback STF circuit has two $nMOS$ transistors which operate in the sub-threshold saturation region. The transfer function of the source follower is:

$$V_{out} = \kappa_n(V_{Input} - V_{bias}) \tag{24}$$

According to our circuit design, this function can be rewritten as:

$$V_g = \kappa_n(V_{out} - V_{lim}) \tag{25}$$

where V_g is the output voltage provided to the gate of the transistor $M4$.

The current through the transistor $M4$ rises exponentially with the charge of the capacitor, limited by an offset, set by the voltage V_{lim}, thus forming a negative-feedback loop. This new path enables the voltage V_{out} to reach the steady-state value even with fully saturated sub-threshold transistors.

If the transistors M_1, M_2, and M_4 are operated in saturation, the equation describing the voltage V_{out}, following one pulse of duration t_{pw} and an arbitrary recovery time t, is:

$$CV_{out}(t) = CV(0) + I_{M_2}t_{pw} - I_{M_1}t - \int_0^t I_{M_4}(V_{out}(t'))dt' \tag{26}$$

Steady-state value is reached when $V_{out}(\frac{1}{f_{in}}) = V(0)$.

The integral is not analytically solvable without further assumptions. However, it provides an intuitive understanding of the qualitative dynamics of the circuit. Assuming a constant positive inter spike charge difference:

$$(I_{M_2}t_{pw} - \frac{I_{M_1}}{f_{in}}) > 0 \tag{27}$$

The output voltage rises linearly per spike until:

$$\int_0^{\frac{1}{f_{in}}} I_{M_4}(V_{out}(t'))dt' = \int_0^{\frac{1}{f_{in}}} I_{n0}e^{\frac{\kappa n(V_{out}(t')-V_{lim})-V_y}{U_T}}dt' \tag{28}$$

This value is high enough to settle the difference. The saturation condition is not necessarily fulfilled, but it should be in the regular case as the feedback is provided by the transistor M_4.

If the input charge per spike $I_{M_2}t_{pw}$ is significantly greater than the constant $\frac{I_{M_1}}{f_{in}}$, the steady-state value is determined by the voltage V_{lim}. We can use the voltage V_{lim} to limit the highest steady-state value. In this configuration, the voltage V_{up} sets the recovery for small frequencies. It also provides a smoother feedback than the sudden cut-off given by the non-saturation transistors in both Sec. 4.3 and [91], resulting in a wider bandwidth of intermediate states between the lowest and the highest steady-state values.

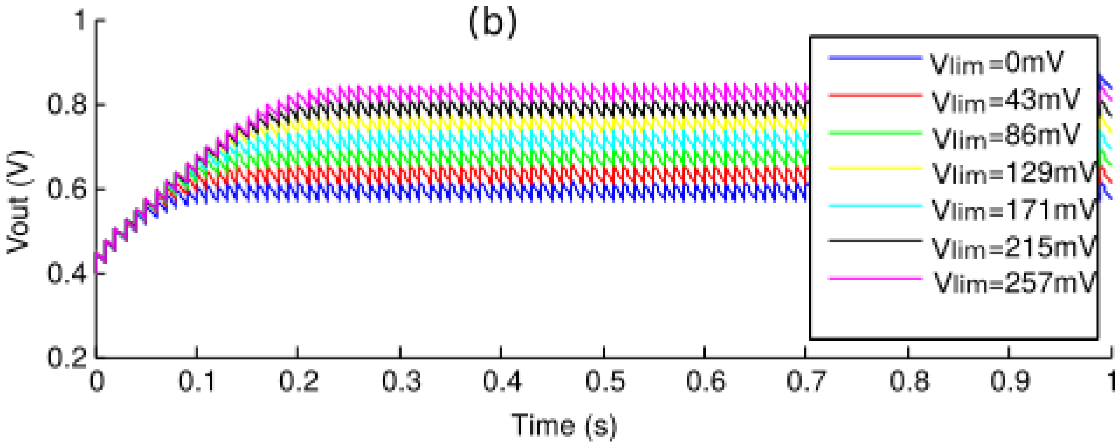


Figure 32: The output voltages of the STF circuit with the feedback recovery control of Sec. 4.4 (published in [90]) in response to the input pulse train stimulus (not shown). The feedback-control voltage V_{lim} is swept and each curve in the plots shows the corresponding change in the output voltage. The voltage V_{lim} determines the offset of the steady-state values. All the output voltages (curves) in both the plots reach the steady-state values. The non-linearity in the recovery rate responsible for the steady-state behavior is clearly visible in the magenta curve of the top plot.

The proposed STF circuit with a feedback recovery control is designed using the standard CMOS Austria Micro Systems (AMS) 350 nm technology. The transient responses of the circuit are characterized using the Spectre® simulator. An experiment is conducted to analyze the influence of the V_{lim} over the steady-state values of the output voltage of the updated STP circuit. The responses are not characterized for V_{wei} and V_{tau} to avoid redundancy in the results. The input pulse train is provided for 1 second (same as the simulation time). The input frequency is 100 Hz, and the pulse width is 100 μ seconds. V_{wei} and V_{tau} voltages are fixed within the sub-threshold range. Only the feedback-control voltage V_{lim} is varied in this experiment. The output voltage of the STF feedback circuit is plotted in Fig. 32. The output voltage of this new STF circuit reach the steady-state values (blue to magenta for $Time \geqslant 0.3s$) for all values of V_{lim} in the experiment. The steady-state condition is achieved by the non-saturated operation of the sub-threshold transistors and a non-linear recovery rate. The speed of the non-linear change in the recovery rate is controlled by the voltage V_{lim}. Therefore, the bias V_{lim} determines the steady-state values. The transient responses of the output voltage ($Time \leqslant 0.1s$) remain the same due to the fixed voltages of V_{wei} and V_{tau}.

The output voltage of the STF circuit with a feedback recovery control can reach the steady-state values. This modified STF circuit is still compact regarding the silicon area because we used a simple n-type source-follower and a $nMOS$ transistor to imple-

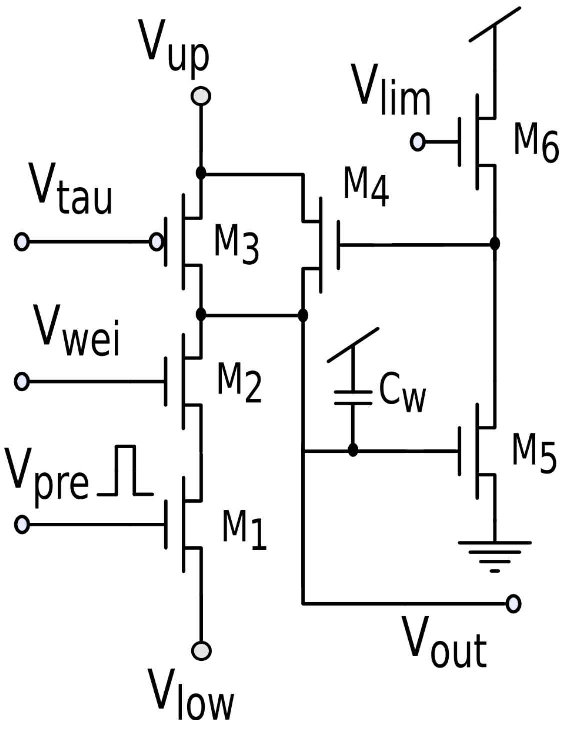


Figure 33: Schematic of the STD circuit with the feedback recovery control (published in [90]). The negative-feedback loop is implemented using the source-follower. The feedback-control voltage V_{lim} is provided to the source-follower's input transistor M_6. The output voltage V_{out} of the STD circuit is supplied as the bias voltage to the source-follower's bias transistor M_5. The output voltage of the source-follower is used to bias the (additional recovery) transistor M_4. The voltage V_{lim} controls the additional recovery path of the depression output voltage V_{out} set by the current through the transistor M_4.

ment the feedback loop. Therefore this STF circuit can be directly integrated with the vast arrays of the DPI synapse. A dedicated STF circuit also offers an independent tuning of the facilitation time-constant in contrast to the implementation of STF using the DPI synapse itself.

4.4.2 *STD circuit*

The operation of the STD circuit with a feedback recovery control is analogous to the STF circuit with a feedback recovery control as discussed above. This variant of the STD circuit is briefly discussed in this section to avoid redundancy in the text. Fig. 33 shows the schematic of this STD circuit (published in [90]). Again, the source-follower is used to provide the negative-feedback loop. The major difference is that the feedback-control voltage V_{lim} is provided to the gate of the input transistor of the source-follower (instead of the bias transistor in the STF circuit with a feedback recovery control). The output voltage V_{out} of the feedback STD circuit is supplied to

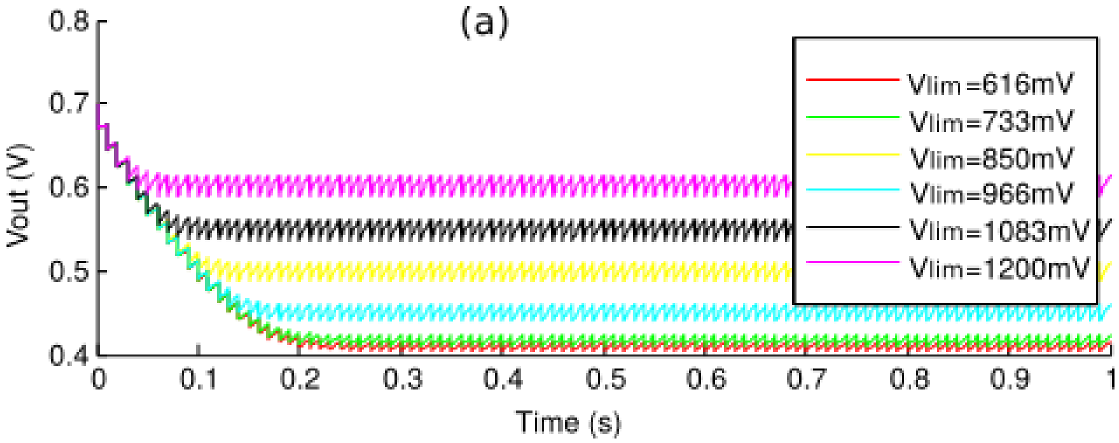


Figure 34: The steady-state values of the output voltage of the STD circuit with a feedback recovery control (published in [90]) in response to the input pulse train stimulus (not shown). The feedback-control voltage V_{lim} is swept, and each curve in the plots shows the corresponding change in the output voltage. The voltage V_{lim} determines the offset of the steady-state voltage. All the output voltages (curves) in both the plots reach the steady-state values. The non-linearity in the recovery rate responsible for the steady-state behavior is visible in the magenta curve of the top plot.

the bias transistor of the source-follower. The output voltage of the source-follower is connected to the gate of the (additional recovery) transistor $M4$. The voltage V_{lim} controls the recovery rate of the output voltage V_{out} through the transistor $M4$, by discharging the capacitor at a rate set by the output voltage V_{out} itself. This way, the output voltage of the STD circuit reaches the steady-state value.

The proposed STD circuit with a feedback recovery control is designed using the standard CMOS Austria Micro Systems (AMS) 350 nm technology. The transient responses of the circuit are characterized using the Spectre® simulator. The same experiment conducted to analyze the steady-state values of the output voltage of the STF circuit with a feedback recovery control for different values of V_{lim} is performed for the STD circuit with a feedback recovery control as well. Sub-threshold bias values are chosen for all the parameters. The results of the experiment are plotted in Fig. 34. From the plots, we can conclude that the V_{lim} voltage determines the steady-state values of the output voltage (red to magenta: for $Time \geqslant 0.3s$). The transient responses of the output voltage are the same for all the curves, due to the constant V_{wei} and V_{tau} voltages.

Though the STD circuit without a feedback control performs right, there are few parameters, for which the steady-state value is not reached (red curve in Fig. 28:top and bottom). The STD circuit with a feedback recovery control solves this limitation by allowing all the output voltages to reach the steady-state values (red to magenta) using V_{lim}. However, the modified STD circuit occupies a more extensive design area with three additional $nMOS$ transistors, compared to the STD circuit without feedback. Therefore, depending on the design space available, the specific type of the circuit can be used to implement the Short-Term Depression (STD).

4.5 SIMULATIONS OF STP CIRCUITS

As mentioned earlier in Sec. 2.3.1, the STP exhibits filtering properties. To validate this concept in silicon, we characterized the filtering properties of the proposed STP circuits, by plotting the steady-state values of their output voltages. All four STP circuits are designed using the standard CMOS Austria Micro Systems (AMS) 350 nm technology. The steady-state responses of the circuits are characterized by the Spectre® simulator. This experiment is comparable to the one performed in biology, mentioned earlier in Sec. 2.3.

The stimulus to the circuits was kept the same as described in the previous experiments. The frequency of the input pulses are varied, and the steady-state voltage responses of the circuits are plotted in Fig. 35. The biases (V_{wei}, V_{tau}, V_{lim}, V_{up} and V_{low}) are tuned in such a way, that the output voltage of these circuits reaches the steady-state value within the stimulation time. We discussed already that the output voltage of the STF circuit without a feedback recovery control could not achieve the steady-state values for most of the input stimulus frequencies (see Fig. 30). Therefore, the value of the output voltage at the end of the stimulus train is taken as the steady-state value just to compare the circuit's filtering characteristics with the other circuits' characteristics.

The output voltage oscillates between two amplitudes during the steady-state (see the circuits' responses in Fig. 28, Fig. 30, Fig. 32, Fig. 34). V_{up} (or V_{low}) determines one end of the steady-state amplitude. The speed of recovery determines the value at the other end of the amplitude. Therefore, mean and SD of these two steady-state amplitudes is obtained. The output voltage remains in the steady-state for a certain period. Therefore, a small time-window is chosen towards the end of the simulation time. The mean and the SDs of the steady-state amplitudes are averaged within this time-window and plotted in Fig. 35. The filter characteristics of the responses of the STF circuit without the feedback recovery control are shown in top-left; the STF circuit with the feedback recovery control in bottom-left; the STD circuit without the feedback recovery control in top-right; the STD circuit with the feedback recovery control in bottom-right.

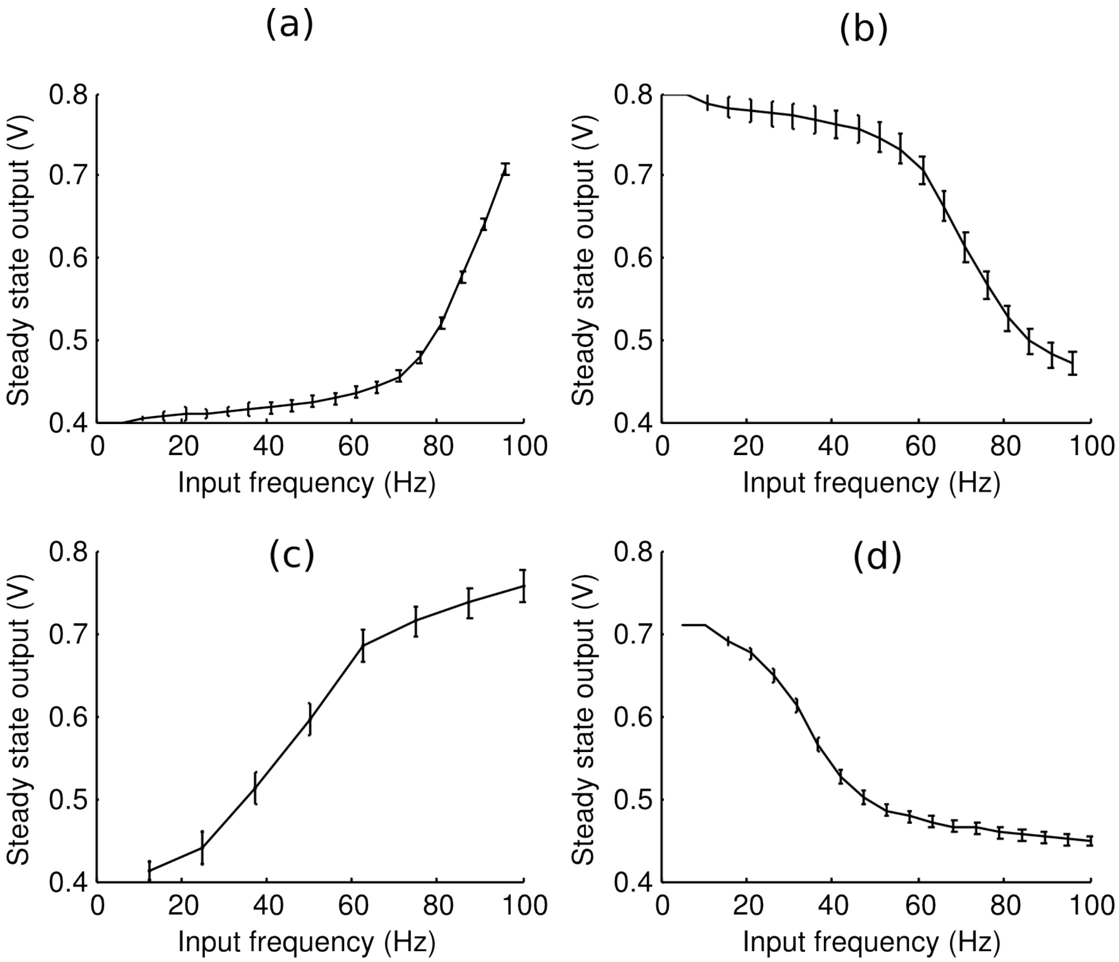


Figure 35: The filter characteristics of the STP are demonstrated using the STP circuits (published in [90]): the STF without the feedback (a); the STD without the feedback (b); the STF with the feedback (c); the STD with the feedback (d). A digital pulse train of fixed pulse-width is used as an input stimulus. The frequency of the input stimulus is varied, and the output steady-state voltage responses are plotted. Each point in the curve represents the mean of the steady-state output voltages averaged over a small time-window of the steady-state voltages. The error-bars denote the averaged SDs of the steady-state voltages within the time-window. All circuits are tuned to achieve the steady-state responses by choosing the appropriate parameters, except for the STF circuit without the feedback, which did not show the steady-state responses within the stimulus duration, for all input frequencies. In this case, the value of the output voltage at the end of the stimulus (0.5 s) is arbitrarily taken for visual comparison to the other plots. Analogous to the observations from the models of the STP (see Sec. 2.3), the STD circuits responses (right: top and bottom) show low-pass filter characteristics, while the STF circuits responses (left: top and bottom) show high-pass filter characteristics to varying input frequencies.

Each dot in the plot is the averaged mean of the steady-state amplitudes, over a small time-window. The error bars represent the averaged SDs of the steady-state amplitudes within the same time-window. As expected from the models of the STP (see Sec. 2.3), the averaged mean steady-state responses of the STD circuits show low-pass filter characteristics (Fig. 35: right) and the responses of the STF circuits exhibit high-pass filter characteristics (Fig. 35: left), in response to the input frequencies. The STP filter curves (averaged mean) can be shifted in either direction by choosing the appropriate parameters of the circuits. In all these plots, the averaged SDs remain almost constant for all input frequencies. The size of the SD of the steady-state amplitudes depends on the choice of the parameters. Please refer table 3 in Chapter 7 for details of the parameters used to obtain the temporal filters in Fig. 35.

Next, we present a simulation result in Fig. 36 to verify that our STP circuits can be used to detect bursts. In this experiment, the STD circuit without the feedback control (see Fig. 27), is stimulated with the bursts of pulses of 200 Hz burst frequency and 25 ms interval between the bursts for 500 ms duration. The stimulus (input voltage) V_{pre} is shown as a green trace in the figure. The parameters V_{wei} and V_{tau} are chosen to be high values. The parameters V_{up} and V_{low} are fixed as 800 mV and 300 mV respectively, which is the optimum range to operate the DPI synapse that follows this circuit. The output voltage $V_{out-std}$ is shown as a red trace in the figure. The output voltage of the STD circuit decreases fast in response to the first pulse of the burst. The pulses within the bursts are too soon for the output voltage to be wholly recovered towards its initial value. However, during the inter-burst interval, the output voltage recovers entirely towards its initial value. These dynamics of strong depression followed by a fast recovery of the STD output voltage is only possible by adding the independent control over the recovery-rate to the STD circuit we designed in this research. Therefore, our STD circuit can be used as a burst onset detector which is a useful property in modeling bio-inspired temporal filters.

The filtering properties of the STP as observed in the biology can be achieved using the STP circuits proposed in this research. These filtering features are used to model the neuron (synapse) as a filter, as in the case of implementing calling song recognition network of crickets in the neuromorphic hardware. It is the other half of this research, which we will discuss in detail in the following chapter.

4.6 FABRICATION AND TESTING OF STP CIRCUITS

To faithfully realize the computations of Short-Term Plasticity (STP) in silicon, the proposed STP circuits were fabricated in a neuromorphic test chip. We call it a testchip-1 and the block diagram of the STP blocks in this chip is shown in Fig. 37. The testchip

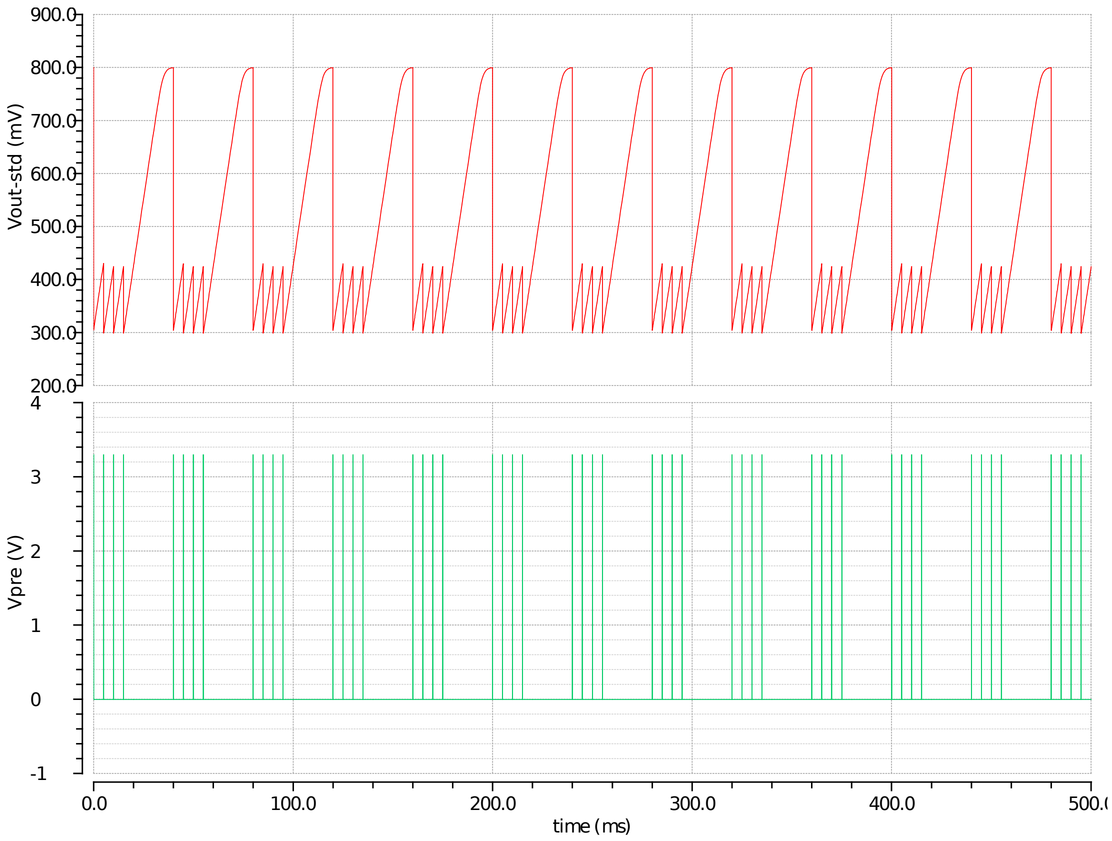


Figure 36: The output voltage $V_{out-std}$ of the STD circuit without a feedback recovery control (shown in red) in response to the input bursts of pulses V_{pre} (shown in green). The output voltage is decreased fast to V_{low} during the onset of the burst due to the strong depression set by V_{wei}. During the burst, the output voltage does not entirely recover towards its initial value due to the high frequency of the pulses within the burst. After the burst, the output voltage completely recovers towards its initial value V_{up} due to the fast recovery rate set by V_{tau}. These dynamics of strong depression followed by quick recovery can be used to detect bursts in a stimulus.

consists of four STP circuits (STD and STF - with and without feedback), a band-pass filter circuit, a calcium-based plasticity circuit and a DPI synapse circuit (refer Sec. 3.5 for more details). The band-pass filter circuit is an additional filter circuit which is explained in detail in Chapter 7. The calcium-based plasticity circuit is not a part of this research. A multiplexer is used to select the STD and STF circuits without feedback and an another multiplexer for the STD and the STF circuits with feedback. The control signal S2 is shared between these two multiplexers. The output of these two multiplexers are fed into the third multiplexer which selects between the STP circuit with and without feedback. The outputs of the band-pass filter circuit and the calcium circuit are multiplexed and its control signal S1 is shared with the third multiplexer which is mentioned earlier. The output of these two multiplexers are fed into the final multiplexer which selects between the STP circuits and the rest. The output voltage of the

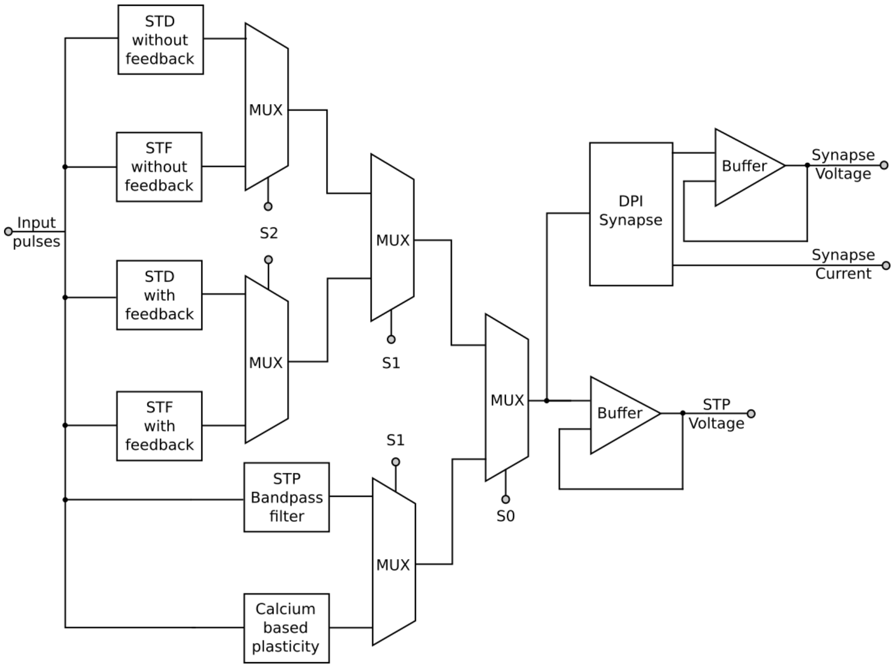


Figure 37: Block diagram of the STP circuits block designed in testchip-1 is shown. The design consists of four STP circuits (STD and STF - with and without feedback), a band-pass filter circuit, a calcium-based plasticity circuit and a DPI synapse circuit. The input pulses are sent to all STP circuits. Multiplexers are used to select the corresponding circuit block through the control signals S2, S1 and S0. The output voltage of the selected STP circuit is buffered and routed to the output pad, and also sent to the gate of the weight transistor of the DPI synapse. The output voltage of the DPI synapse is also buffered and routed to the output pad, and the output current shows the effect of STP on its amplitude.

multiplexer is buffered through an opamp buffer to decouple the output voltage from the output pad. The output voltage of the multiplexer is also supplied to the gate of the weight transistor of the DPI synapse. The input pulses are provided to stimulate all the STP circuits. However, due to the multiplexing only one of the STP circuits is active at a particular time. Therefore, the DPI synapse receives only one of the STP output voltages. The output voltage of the DPI synapse is also buffered, and the output current of the synapse is also routed to the pad. For a given input pulses, the effect of STP on the synaptic weight in terms of can be estimated through the change in amplitude of the synaptic currents, in our design. Given the limited design area of the testchip, neurons are not included in the design. Therefore, the AER communication protocol is not included in the design of our part of the testchip.

The layout design of the STP block consisting of four STP circuits is shown in Fig. 38. The circuits were designed and laid-out using a standard CMOS AMS 180 nm technology. Conventional transistors of width: 1 µm and length: 0.36 µm are used in the

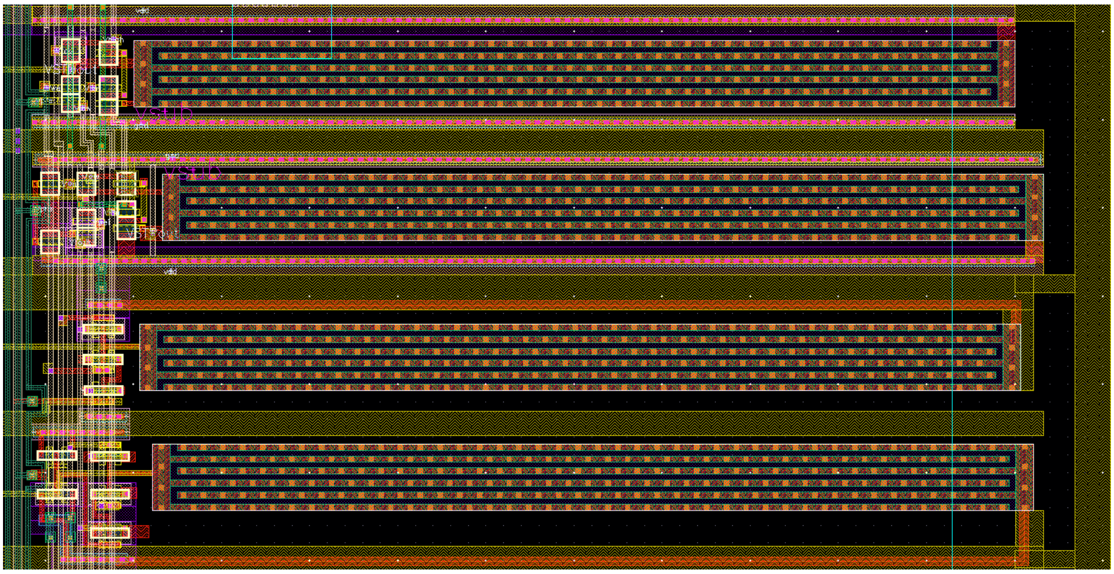


Figure 38: The layout of the proposed STP circuits designed using the standard CMOS AMS 180 nm technology. The circuits (with the design area) organized from the top are as follows: STD with the feedback (62.74*7.0 μm^2); STF with the feedback (62.74*7.0 μm^2); STD without the feedback (62.74*7.59 μm^2); STF without the feedback (62.74*8.44 μm^2). The transistors are located at the left of the layout and occupy a small area. The capacitors on the right half of the layour occupy most of the silicon area. Further design aspects are provided in the table 1

.

design. The capacitor occupies the area of 50*3.76 μm^2 and it is designed using the four 'Metal' metal-insulator-metal (MIM) for a total capacitance of 100 fF. In total, six layers of 'Metal' are used in the design. The biases are shared across the STP circuits, due to the limitation of the total count of the I/O pins that need to be bonded from the chip. The layout of the circuits (with the design area) organized (Fig. 38) from the top are as follows: the STD circuit with a feedback recovery control (62.74*7.0 μm^2), the STF circuit with a feedback recovery control (62.74*7.0 μm^2), the STD circuit without the feedback (62.74*7.59 μm^2) and the STF circuit without the feedback (62.74*8.44 μm^2). Please refer the table 1 for more details about the dimensions used in the layout design.

The STD circuit proposed by C. Rasche and R. Hahnloser in [91] (see Fig. 25 of Sec. 4.2 for more details) was fabricated using 1.2 μm technology. The transistors were 5 μm in width and 5 μm in length, and the capacitor was 0.2 pF. Compared to this circuit, the transistors of our STP circuits fabricated using 180 nm technology occupy 70 times the smaller area and the capacitor is 2 times smaller in its capacitance value. The transistors located on the left of the design occupy a small area compared to the capacitors on the right, which consume most of the silicon area of the circuits. It is true

Block	Type	Length	Width	Value
STD without feedback	M1-M4 Cap	50μ	3.76μ	96.2fF
	Transistor	0.36μ	1μ	-
	Silicon occ.	62.74μ	7.59μ	-
STF without feedback	M1-M4 Cap	50μ	3.76μ	96.2fF
	Transistor	0.36μ	1μ	-
	Silicon occ.	62.74μ	8.44μ	-
STD with feedback	M1-M4 Cap	50μ	3.76μ	96.2fF
	Transistor	0.36μ	1μ	-
	Silicon occ.	62.74μ	7μ	-
STF with feedback	M1-M4 Cap	50μ	3.76μ	96.2fF
	Transistor	0.36μ	1μ	-
	Silicon occ.	62.74μ	7μ	-

Table 1: Dimensions of the STP circuit blocks designed in testchip-1.

for all four STP circuits. The layout of the STP circuits with a feedback recovery control is slightly larger than that of the STP circuits without the feedback. It is due to the presence of three additional transistors of the negative feedback circuit. The output voltage node of these STP circuits is connected to the operational-amplifier buffer (not shown in the layout), which decouples the output voltage of these circuits from the pads. The pads are used to safely measure (or supply) the voltages (or currents) from (or to) the chip.

We tested the STP circuits fabricated in the neuromorphic test chip. During testing, we analyzed the output voltage of all four STP circuits by presenting the pulse train stimulus. The waveform generator provides the input pulses of 100 μ seconds in pulse-width and 150 Hz in frequency for a duration of 250 milliseconds. The biases V_{lim}, V_{tau}, and V_{wei} are tuned in such a way, that the output voltages of the STP circuits reach their steady-state values within the stimulus duration. The voltages V_{up} and V_{low} are set to 0.8 V and 0.3 V respectively, such that the weight transistor of the DPI synapse stays in the sub-threshold region. The output voltages of the STP circuits are recorded through the oscilloscope. The recorded data are plotted in Fig. 39. The top plots show the output voltages measured from the STP circuits without the feedback. The bottom plots show the responses from the STP circuits with the feedback. The left half of the figure shows the STD circuit responses, and the right plots show the STF

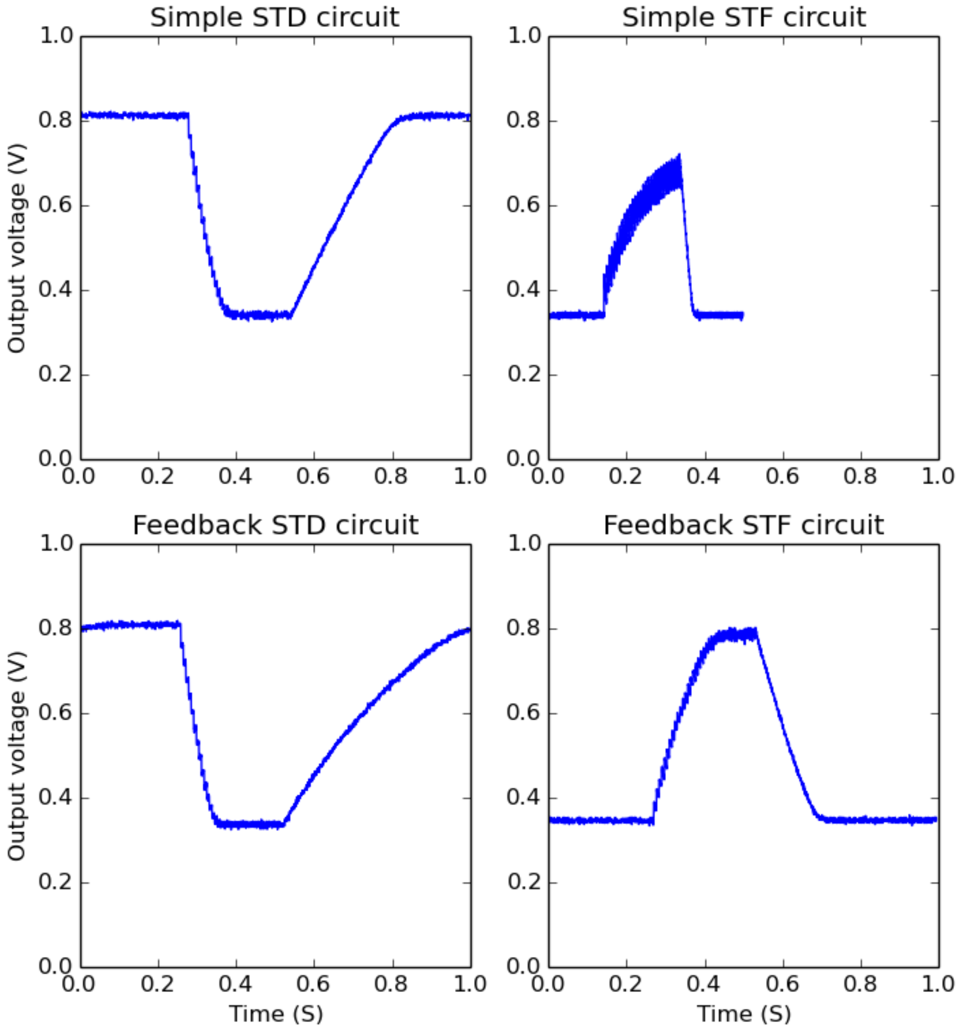


Figure 39: The output voltage responses of the fabricated STP circuits recorded through the oscilloscope are shown. Top-left: STD circuit without the feedback; Top-right: STF circuit without the feedback; Bottom-left: STD circuit with the feedback; Bottom-right: STF circuit with the feedback. A train of input pulses of 150 Hz frequency and 250 milliseconds in duration is provided as a stimulus to these circuits through the waveform generator. The voltage biases are tuned, such that the output voltages of these circuits reach the steady-state values within the stimulus duration. The traces show that all of the STP circuit responses reach the steady-state values within the stimulus duration, except for the STF circuit without the feedback. This behavior is expected from the simulation results (as discussed in the previous section).

circuit responses. As expected from the simulations, all except the output voltage of the STF circuit without the feedback reach the steady-state values within the duration of the stimulus.

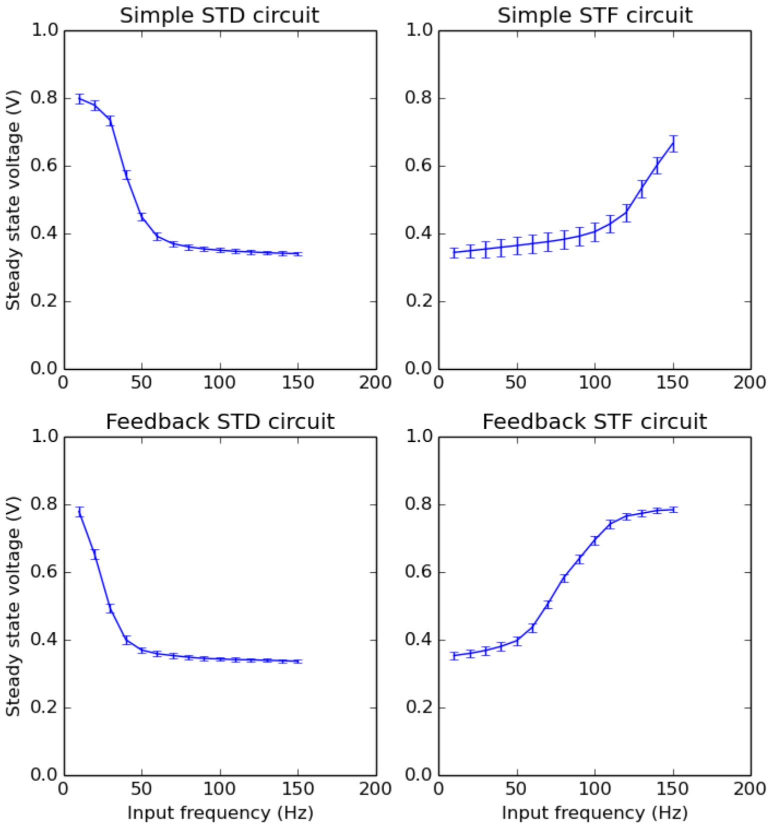


Figure 40: The filtering properties of the fabricated STP circuits are shown. Top-left: STD circuit without the feedback; Top-right: STF circuit without the feedback; Bottom-left: STD circuit with the feedback; Bottom-right: STF circuit with the feedback. The input is a train of pulses with a fixed pulse-interval of 100 μ seconds sent from the waveform generator. The input pulse frequency is varied in this experiment. The stimulus duration is varied based on the input frequencies, such that the output voltage reaches the steady-state values within the stimulus duration. This variation in the stimulus duration is consistent for all the circuits. Each point in the curve is the mean of the steady-state amplitudes, and the error bars represent the SDs of those amplitudes. As expected, the STF circuit without the feedback did not achieve the steady-state responses within the stimulus duration. Therefore, the mean of the amplitudes at the end of the stimulus is plotted as the steady-state values. As expected from the simulation results, the STD circuits responses show low-pass filter characteristics and the STF circuits responses show high-pass filter characteristics. The absolute peak-to-peak amplitudes of the steady-state output voltages decrease for high-frequency inputs, due to the short ISIs, resulting in small SD values.

We also tested the filtering characteristics of the fabricated STP circuits responses. In this experiment, a train of input pulses is sent through the waveform generator to the chip. The input pulses have a pulse-width of 100 μ seconds, and the input frequency is varied from 10 Hz to 150 Hz in steps of 10 Hz. The duration of the stimulus is changed during the experiment depending on the frequency of the input pulse. The length of the stimulus is varied, such that the output voltage reaches the steady-state value within the stimulus duration. The biases are kept the same as the previous experiment. The output voltages from the chip are recorded through the oscilloscope.

The mean and the SDs of the steady-state amplitudes are computed for each circuit in response to each stimulus frequency. As expected, the output voltage of the STF circuit did not reach the steady-state value for most of its input frequencies. Therefore, the mean of the output voltage amplitudes at the end of the stimulus is taken as the steady-state value to compare one-to-one with the circuits' responses.

The steady-state responses of the four STP circuits to the input frequencies are shown in Fig. 40. Every point in the plot is the mean of the steady-state amplitudes. The error bars represent the SDs of the steady-state amplitudes.

As expected from the simulations, the responses of both the STD circuits (top: left and right) show low-pass filter characteristics and the responses of both the STF circuits (top: left and right) exhibit high-pass filter characteristics. The top-right plot shows a sharp rise in the high-pass filter profile, due to the non-steady-state responses of the STF circuit without the feedback control. The shape of the filter curves can be altered by choosing another set of the parameters. For instance, the parameters can be tuned further for a better low-pass filter response of the feedback STD circuit. The time-constant of recovery of the output STP voltage is fixed to all inputs. During high-frequency stimulation, the absolute peak-to-peak amplitude of the steady-state output voltage of the STD circuit without the feedback, recovers to a small value compared to the low-frequency inputs, due to the short ISIs and this applies to all four STP circuits. Therefore, the SDs of their steady-state output voltages decrease during high-frequencies. However, the output voltage of the STF circuit without the feedback, recover entirely towards the resting voltage for specific frequencies. Therefore, the corresponding SDs of the steady-state amplitudes are significant for low-frequencies (see $50 \leqslant \mathsf{Freq} \leqslant 100\mathsf{Hz}$). Please refer table 4 in Chapter 7 for details of the parameters used to obtain the temporal filters in Fig. 40.

The results presented here are analogous to the circuit simulations and are comparable to the STP filters characterized in [1]. The filters of the STP can be used to model the post-synaptic neuron responses to select specific pre-synaptic input frequencies. This idea is implemented in the model of the calling song recognition network of crickets. More details of this implementation can be found in the next chapter.

The individual STP circuits are functioning as expected, in the fabricated test chip. These STP circuits are laid-out again in another test chip (using the same technology) and are used as computational STP filter blocks in a small neural array. The purpose of this neural array is to implement the calling song recognition network of crickets in a dedicated neuromorphic chip. The characterization of this new test chip is out of the scope of this research. Please refer Chapter 7 for more information on this neural array implementation.

4.7 CONCLUSION

The STD circuit proposed here is identical to the linear charge-and-discharge synapse circuit proposed in [5]. The main difference is the usage of this circuit. In [5], the circuit is used as a synapse, whereas in our implementation, the output of the circuit is voltage-limited and used to bias the 'weight' transistor of the DPI synapse. The advantage of these new STD circuits is the independent control over the recovery dynamics which was missing in the previous design of [15]. With the explicit recovery control, the circuit can detect the bursts, a property of the Short-Term Plasticity (STP) as discussed in Sec. 2.3.2. If the bursts of spikes are presented to the synapse, the circuit identifies the onset of bursts by the strong STD voltage, which recovers fast before the arrival of the next burst. These temporal dynamics are possible only with the tunable recovery of our STD circuit. The STP circuits without the feedback are compact in design, but the steady-state non-convergence in the STF circuits leads to another design by adding a negative feedback loop. The STF circuits, provide more flexibility to the DPI synapse, especially the STF circuit with the feedback recovery control. The STF circuit allows the DPI synapse to operate with large currents during the STF implementation, which is not the case if DPI synapse is used to implement the STF. The STF circuit also provides independent time-constant of the STF without limiting the operating range of the DPI synapse. The usefulness of the neuromorphic STP is demonstrated in the next chapter. We also combined the STD and STF circuits in the design, to obtain band-pass filter characteristics. This variant of the STP circuit might be useful to filter out specific frequencies. The circuit is discussed in detail in the Chapter 7.

5 NEUROMORPHIC MODEL FOR CRICKET CALLING SONG RECOGNITION NETWORK

5.1 INTRODUCTION

Timing is a crucial factor in sensory processing. For instance, spatiotemporal cues are the essential components of object recognition in human vision [68]. Temporal structures of audio signals are also crucial in sound recognition. The auditory neurons of the brain encode these temporal patterns of the sound. Understanding this neural coding scheme is a key to the speech-recognition research. A simple neural sound recognition system is found on grasshoppers [74], crickets [101], drosophila [25]. These systems can be used as an object of study considering the simple architecture of their neural networks. Insects have relatively small nervous systems with a small number of neurons compared to vertebrates. For example, an adult Drosophila brain has approximately 100 thousand neurons [21], whereas the mouse brain is estimated to have around 75 million neurons [78]. This quality ensures insects model organisms to study. In our research, we study 'crickets', a class of insects that have been extensively studied for their acoustic communication abilities [100].

Crickets display a variety of acoustic-oriented behavior such as singing (or responding to) rivalry songs, courtship songs and calling songs. Males produce calling songs by rubbing their wings against each other to attract females. Females respond to these species-specific calling songs by positioning (rotating) and moving in the direction of their preferred source of the sound. This behavior of female crickets is called phonotaxis [105]. It is crucial for the survival of their species because the females evaluate the fitness of the male through the quality of the calling song and mate the fittest male in the field.

In this research, we study the neural network that is responsible for recognizing calling songs during cricket phonotaxis. The network offers the understanding of the principles of auditory processing of the individual neurons. We modeled a calling-song recognition network in the neuromorphic hardware using the STP synapses. The network is modeled based on the neurophysiological evidence of crickets' auditory neurons. This model provides an understanding of the network architecture constructed in the neuromorphic hardware. The model also demonstrates the temporal filtering properties of the STP (as discussed in Sec. 2.3.1) during calling song recognition. The details of the model will be explained in the 'methods' section of this chapter. The biological evidence of phonotaxis will be discussed first in this chapter, followed by the implementation of the model, the results obtained from the neuromorphic hardware and finally, the discussion about the results will be presented.

5.2 NEUROBIOLOGY OF CRICKET PHONOTAXIS

Several behavioral experiments on phonotaxis have been carried out across various species of crickets [105], [43], [57]. In all these experiments, an artificial stimulus, reproducing the main features of original songs were presented to the crickets, and their response was tested. A typical artificial stimulus (see Fig. 43 black waveform) lasts for a duration of 300-500 milliseconds (Chirp Duration (CD)) and separated by intervals (Chirp Intervals (CIs)) of approximately 500 milliseconds in duration. These values were chosen following the characteristics of the natural calling songs. For instance, the CI of the male calling song ranges from 300 ms to 700 ms. Each CD is further divided into pulses and pauses of short duration defined as Pulse Duration (PD) and Pulse Interval (PI). The combinations of PD and PI are referred to as Pulse Periods (PPs). The pulses (within the PD) are made of an amplitude modulated sinusoidal carrier wave of 5 kHz frequency and 80 dB intensity (the preferable range for female crickets).

In [105], the behavioral response of crickets was measured in response to the artificial calling songs. During the experiment, the female cricket was clamped on a trackball, called the walking compensator. The trackball moves in the opposite direction of the insect's movement, such that the actual position of the subject always stays constant. The trackball was placed in a chamber with two loud-speakers on each side, placed at different angles. Artificial calling songs ranging from 10 ms to 98 ms of PPs were presented through the loud-speakers. The songs are shown in Fig. 41 (top). In response to these songs, the female crickets start walking towards the source of sound (corresponding speaker). The values of the angle and the velocity of walking were extracted from the trackball. Based on the collected data, phonotaxis scores were computed for few seconds of tracking per minute of a trial. The score is defined by the speed of the alignment of the animal and the speed of the movement on the trackball, towards the right sound source. The resulting scores of the behavior were plotted in Fig. 41. The scores were plotted for various PPs. The overall profile of the scores resembles a band-pass filter with the highest response, around 30-50 milliseconds PPs. Therefore, female crickets prefer these values of the PPs, possibly revealing a good fitness of the singing male.

Neuroscientists have been studying the cricket brain with the goal of elucidating the neural substrate triggering this behavior. In the seminal work of Schildberger [101], six types of auditory neurons were identified in *Gryllus bimaculatus*. The neural responses were again characterized as a function of the PPs and interpreted regarding temporal filter properties (e.g., high/low/bandpass). The auditory system of crickets begins with the hearing organs (ears) in their front legs. Approximately 70 sensory receptor fibers (or afferent neurons) are identified in each ear of the cricket. The auditory afferents have small axons that terminate in the pro-thoracic ganglion. The auditory neurons originating from the pro-thoracic ganglion were grouped into three classes based on their anatomy [101]. They are Ascending Neurons (ANs), Central Brain Neurons 1s (BNC1s) and Central Brain Neurons 2s (BNC2s). Ascending neurons (Ascending Neuron 1 (AN1) and Ascending Neuron 2 (AN2)) ascend from the pro-thoracic ganglion to the brain. BNC1 are located in the brain, and their dendrites are physically co-located with the axons of ascending neurons. Anatomical evidence supports the hypothesis of ascending neurons being the primary input to BNC1 [101], but the synaptic connections have not been reported. BNC2 are also located in the brain, but they do not arborize in the projection field of the ascending neurons. Based on the frequency tuning, the Brain Neurons are further classified into four types namely a, b, c and d. The a and the b neuron types respond to the low-frequency stimuli, while the c and the d neuron types respond to the high-frequency stimuli.

Next, let us analyze the responses of these identified neurons as they form the basis of our model. AN1 neurons are highly sensitive to low-frequencies (approximately 5 kHz), in the range of the carrier frequency of the calling songs of crickets. On the other hand, AN2 neurons respond only to high frequencies (in the range of 10-20 kHz, which are the lowest frequencies of bats' sonar [30]), thus providing a detection signal to avoid predators. Since AN2 neurons respond only to high input frequencies, they do not play a role in calling songs recognition which is low-frequency signals. During calling song recognition, the latency of the onset of the spiking response increases from AN1 neurons to BNC1 and BNC2. The latency provides a clue about the feed-forward topology of the network.

The role of these neurons in cricket phonotaxis was characterized in [101], by studying their responses to chirps with various temporal structures (similar to the previously mentioned behavioral experiments). The neural responses were normalized, and the percentages of the measures were computed. These percentages are called the relative magnitude of the responses. For the given CDs and the duty-cycle, the PP of the chirp was increased from 10 ms to 98 ms as shown in Fig. 41 (top). The relative magnitude of responses was averaged over twelve stimulus presentations per neuron for three sensory brain neurons. These responses are shown in Fig. 41 (bottom). The figure shows the responses of BNC1d (filled squares), BNC2b (filled triangles) and BNC2a (filled circles). The neural responses measured from other crickets are also shown (empty shapes) to demonstrate the degree of variability across different crickets. The shaded region represents the phonotaxis scores (as mentioned earlier) re-plotted from [105].

Few neurons of the BNC1 group did not spike for a 10 ms PP stimulus, but they responded to the stimulus with long PPs (70-100 ms). This response resembles the characteristics of a low-pass filter [101]. Similarly, high-pass and band-pass filter characteristics were found in the responses of BNC2 neuron group. It is postulated in [101], that the low-pass response might be due to the summing up of graded potentials, especially during the long PP (70-100 ms) stimuli. In other words, a small change in the EPSP occurs for short PPs. However, the EPSP builds up to the threshold to elicit spikes for long PPs. Another hypothesis was proposed in [101], which stated the band-pass response might be a result of the logical AND operation of the low-pass and the high-pass responses. The neurons with low-pass and high-pass responses converge onto the neuron with a band-pass response, which functions as a coincidence detector of the incoming responses. That is, the neuron responds only if both the input responses are high. This neurophysiological evidence presented in [101] matches the cricket's behavior measured in [105]. This neural data also confirm that the female crickets prefer a specific type of the calling songs.

Many researchers continued the work of Schildberger, often using similar methodologies. Specifically, in [58] more auditory local inter-neurons were identified. Their responses were characterized for various PPs and duty-cycles of the calling song. The data presented in [58] gives a good insight into the recognition mechanisms of calling songs. A set of local neurons, namely B-LI2, B-LC3, B-LI3, and B-LI4, were identified alongside the previously discussed ascending neuron AN1 which is called TH1-AC1. The naming scheme is changed in their work, and it is explained as follows: B refers to the brain, TH1 refers to the 1st thoracic ganglion, A to ascending, L to local, I to ipsilateral and C to the contra-lateral position of the axons. The auditory neurons were stimulated for various PDs, PIs and PPs and their responses were measured. The experiments conducted by Schildberger in [101] and Kostarakos and colleagues in [58] are different. For instance, the duty-cycle of the stimuli was fixed to 50% in Schildberger's

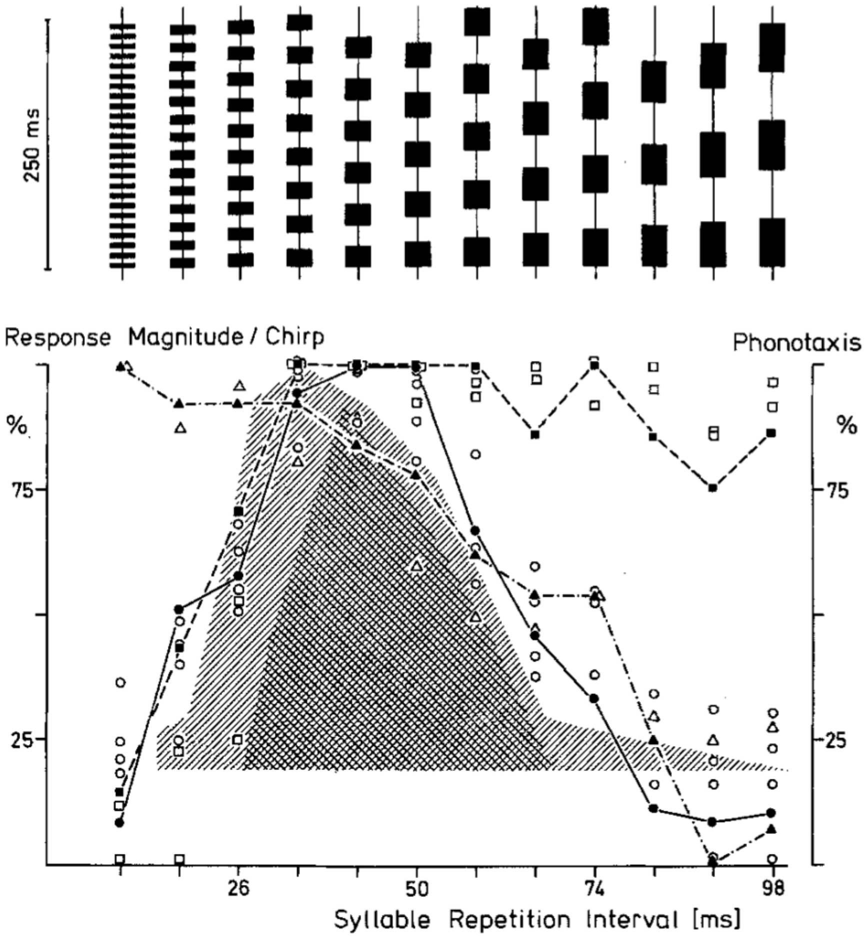


Figure 41: Responses of the auditory brain neurons of crickets to various PPs of the stimulus presented in [101]. Different types of artificial songs consisting of a chirp of 250 ms CD (top) with increasing PPs (top: left to right) are shown. The relative magnitude of the responses of three auditory neurons averaged across twelve trials are shown (bottom). Refer text for details of this measure. The filled shapes denote the responses from one cricket. The empty shapes denote the responses of the corresponding neurons in other crickets. The squares denote the BNC1d responses that resemble a low-pass filter profile. The triangles denote the BNC2b responses resembling a high-pass filter. The circles denote the band-pass profile of the BNC2a neuron responses. The shaded region in the middle shows the phonotaxis scores of two species of cricket presented in [105]: *Gryllus campestris* (single stripes) and *Gryllus bimaculatus* (cross stripes). Refer text for details about the calculation of these scores.

work [101], whereas in Kostarakos and colleagues' work [58], the duty-cycle was also varied along with the PPs of the chirps. The change in the duty-cycle is essential, as the duty-cycle of the natural calling songs are not fixed.

The neural responses presented in [58] is shown in the Fig. 42. The relative response of the neurons was computed by summing up the total number of spikes within a chirp and averaging across different animals. These averaged values were normalized, and the percentages were computed. The x-axis represents the PDs, and the y-axis represents the PIs. The secondary diagonal represents the responses to increasing PPs with a constant duty-cycle, similar to Fig. 41. The peak responses are marked with asterisks. The color-code of the responses is shown below the subplots. The first subplot of Fig. 42 shows the behavioral data of cricket phonotaxis obtained from the track-ball experiment, similar to the one mentioned earlier [105]. The phonotaxis scores are high, in response to the stimulus with 34 $\mathcal{ms}$ PP and nearby region. The TH1-AC1 and the B-LI2 neurons showed unspecific response patterns compared to other neurons. The B-LC3 and the B-LI3 neurons showed band-pass filter like responses, whose peaks are centered around 34 $\mathcal{ms}$ PP stimulus. These two local neurons were responsive to the PPs other than 34 $\mathcal{ms}$ as well, meaning that their responses are not strictly band-pass. However, the B-LI4 neuron showed a fine-tuned band-pass filter like responses, which are centered around 34 $\mathcal{ms}$ PP. In [58], the authors showed that the B-LI4 neuron's responses (bottom-right subplot) were highly correlated with the phonotaxis behavior (top-left subplot). Therefore, the authors conclude that the B-LI4 neuron might tune the behavior to respond selectively to calling songs. They suggested that the B-LI4 is a 'feature detector' neuron, which detects the attractive stimulus when the temporal pattern of the stimulus coincides with an internal template.

Various theories on pattern recognition of the calling songs have been proposed. Matching or cross-correlation between the temporal pattern of the stimulus and the internal neural template were suggested as possible mechanisms for temporal selectivity in cricket brain in [44, 41]. In [17], the authors proposed that the temporal selectivity might be the result of intrinsic membrane-potential oscillations in resonance with the frequency of the stimulus. As aforementioned, Schildberger proposed in [101], that the band-pass temporal selectivity might be a result of the logical $\mathcal{AND}$ operation of low-pass and high-pass filter responses. However, Kostarakos and colleagues suggested that coincidence detection might play a role in calling song recognition [58]. On considering these available theories, we can conclude that the auditory neurons of crickets exhibit band-pass filter like selectivity. At the same time, the neural mechanisms responsible for the selection process are still unverified.

Studying the auditory neurons of crickets has the potential to uncover the computational principles of temporal processing in the brain. The encoding mechanisms of the auditory neurons associated with the temporal structure of the stimulus are not fully

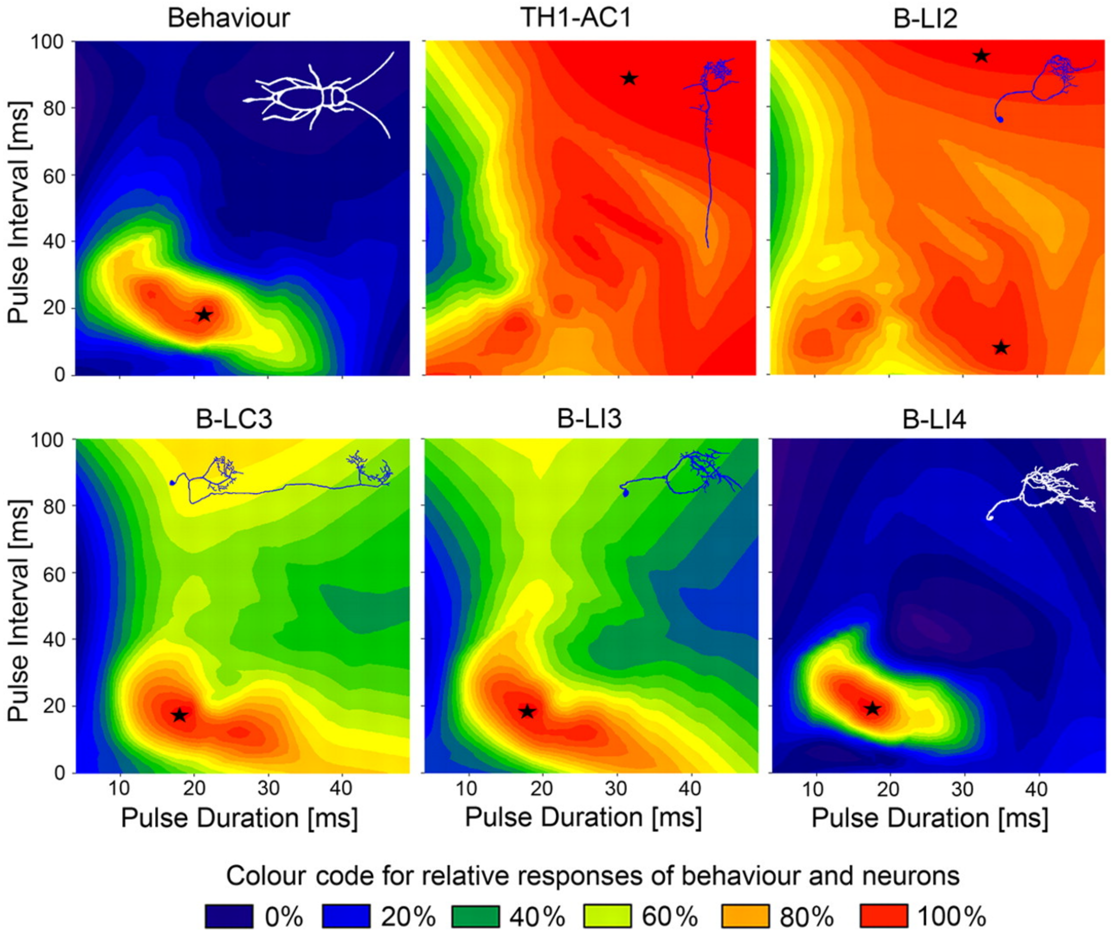


Figure 42: Relative responses of cricket brain neurons along with the relative phonotaxis scores (top left) presented in [58], corresponding to a chirp stimulus varying in the PP and the duty-cycles. Refer text for details of the relative response measure. The asterisks denote the activity peaks.The color-code is shown below the plots. TH1-AC1 and B-LI2 respond high to almost all PPs. B-LC3 and B-LI3 respond to all PPs and showed high responses to 34 ms PP region. B-LI4 respond high to 34 ms PP region and respond minimum to other PPs. The behavioral responses showed a selective response to the 34 ms PP region.

understood in phonotaxis literature. The computational phenomenon underlying the temporal selectivity of sensory neurons can be uncovered by following how neurons recognize the specific pattern of calling songs when putting together in a network. Since we propose that the recognition occurs at the network level, it is essential to understand the structure of the network. Two kinds of networks have been introduced in neuroscience literature stemming from the neurophysiological evidence. In [113], a calling song recognition network was proposed, based on the anatomy of auditory neurons and the latency of their spikes (the flow of information). Another calling song recognition network was introduced in [103], based on the latencies of the spikes of the auditory neurons. Despite the previous works, computational principles of individual neurons within the calling song recognition network is not yet achieved. That is, the algorithm of the network to recognize the calling songs that respond with a maximum number of spikes has not been found. Therefore, we built a spiking model of calling song recognition network in the neuromorphic hardware that recognizes the artificial calling songs. As we discussed earlier in Chapter 3, the neuromorphic hardware offers a silicon model of the neurons and the synapses operating in real-time with biologically realistic time-constants. These features enable us to model the temporal computation of each auditory neuron individually and allow us to construct a functional network. The neurons of the network are tuned, such that the output responses of the network qualitatively match the behavioral data. Besides the computational principles of individual neurons are analyzed in this research, the computational architecture of the whole system is also revealed. We will cover the implementation of this model in the next section.

5.3 EXPERIMENTAL METHODS

In this section, we will discuss the experimental methods used in this research, which include the computational principles of the model, the structure of our network, the stimulus presented to our network, and the details of the neuromorphic hardware implementation. In this section, we characterize the computational paradigms used in the network. Later, we test the tuned network with different input spike patterns. Please refer table 5 in Chapter 7 for the details of the parameters used to emulate the networks. We will start with the stimulus used in our experiment.

5.3.1 *Stimulus*

In biological experiments [43, 73], artificial chirps resembling the male cricket calling songs are presented to the female crickets (see Fig. 43). Meckenhäuser and colleagues in [73] analyzed the structural features of the calling songs that affect the phonotaxis behavior. The calling songs were presented by varying the temporal features of the song, and the behavioral responses were recorded. An artificial neural network was trained by using this data for eight combinations of six temporal features, to predict the phonotaxis scores for the new data. From this model of [73], Meckenhäuser and colleagues proposed that four temporal features were sufficient to represent the artificial calling song: two from a short time-scale (5-50 milliseconds) and two from a long time-scale (300-800 milliseconds). The short-time features are the Pulse Duration (PD), and the Pulse Interval (PI), while the long-time features are the Chirp Duration (CD) and

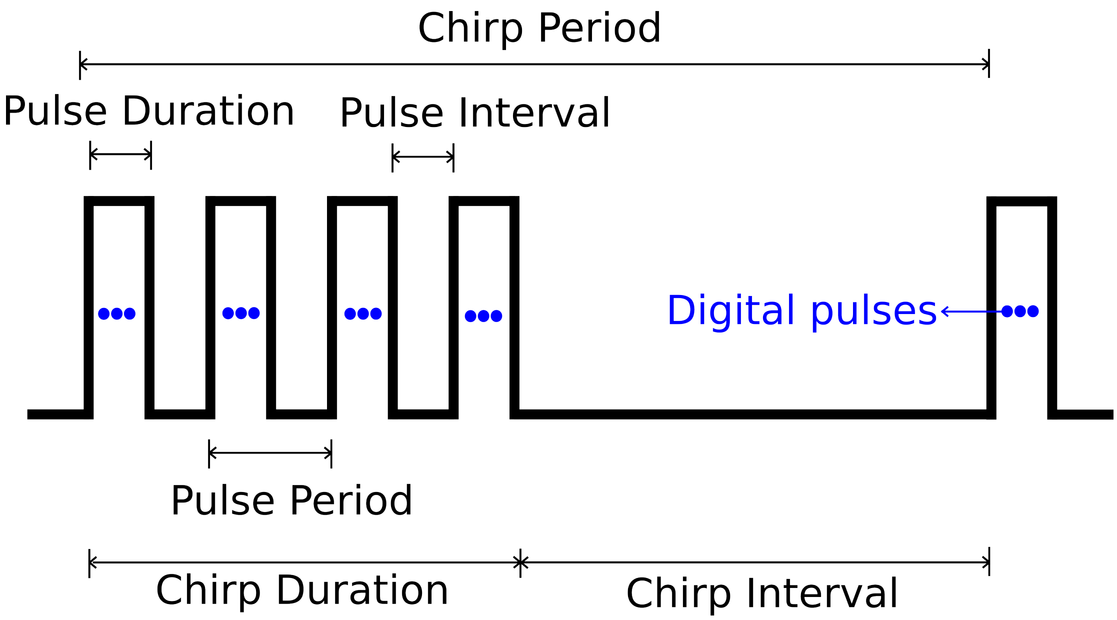


Figure 43: Artificial stimulus used in our experiments that represents a chirp of the cricket calling song. The chirp is provided for a particular duration (Chirp Duration (CD)) with the intervals in-between (Chirp Interval (CI)). The total duration of the Chirp Duration (CD) and the Chirp Interval (CI) is called the Chirp Period (CP). Each Chirp Duration (CD) consists of short pulses of specific Pulse Durations (PDs). The pauses within the Chirp Duration (CD) are called the Pulse Intervals (PIs). The sum of the PD and the Pulse Interval (PI) is called the Pulse Period (PP). In our experiments, we used digital pulses (shown as blue dots) of 800 Hz frequency to represent the sound signal, and the Pulse Duration (PD) are made of these digital pulses.

the Chirp Period (CP). These features are typically used in the biological experiments of cricket phonotaxis.

To faithfully model the neural responses analogous to the biology, we used the stimulus similar to the one used in the neurophysiological and the behavioral experiments. A chirp with various combinations of the PD and the PI was used in [58]. The PDs and the PIs used in one of the experiments in [58] were adopted in our stimulus. The behavioral and the neuronal responses in [58] suggest that the band-pass selectivity of phonotaxis occurs within the PD of 50 ms and the PI of 50 ms. Therefore, we adopted this scale in our experiments. However, in [58], the maximum of the PI (100 ms) was twice the duration of the PD (50 ms). Our artificial stimulus consists of a chirp that lasts for 260 ms. The temporal selectivity of neurons reported in [101, 58], occurs within one chirp. Therefore, only single chirps (and no CIs) are considered in our experiments. The chirps are further divided into the PDs and the PIs, as shown in Fig. 43. The PDs and the PIs used in the experiments are: (5, 9, 13, 17, 21, 25, 29, 34, 38, 42, 46 and 50 ms). All combinations of these PDs and PIs are employed, to obtain all PPs. Unlike in

[101], we did not fix the duty-cycle. From the Fig. 42, we can conclude that the duty-cycle variations can significantly capture the band-pass filter like responses. The PD is made of bursts of digital pulses (shown as blue dots in Fig. 43) of 800 Hz frequency. These digital pulses represent the output spikes of the sensory neurons that project into the ascending neurons. The structure of these digital pulses resembles the shape of the calling song. The frequency of these pulses is regular, meaning that the pulses (spikes) are spaced at uniform intervals, to capture the inherent noise of the neuromorphic hardware effectively. As mentioned earlier in Sec. 3.7, the responses of the neural circuits vary from one another, due to the effects of 'device-mismatch' resulting from the fabrication process. Therefore, (non-uniform) responses to the uniform inputs allow us to characterize the mismatch in the hardware. The inherent noise of the analog hardware is used to show the deviations in the responses across the networks. The networks are fine-tuned in four stages to respond to a regular spiking input. The neurons of the network use a sparse number of spikes. The model also relies on specific spike-times within the PDs during a chirp. The number of the output neurons showing desired responses is already limited due to the influence of device-mismatch. Using a time-variant noise such as poisson spikes would drastically affect the performance of the system. To avoid this problem, we instead use a uniform spike-train input to demonstrate the STP in our network.

5.3.2 *Network model*

We have modeled the calling song recognition twice in this research. We will start by discussing our first model followed by the recent implementation. Our first model is an extension of Rost's Master thesis [96]. In his thesis, Rost has modeled four auditory neurons using STP synapses in a Brian Simulator, based on the physiological evidence by Schildberger in [101]. Together with Rost, we modeled a calling song recognition network in our neuromorphic hardware and published in [97]. The network consisted of an ascending neuron and neurons with low-pass, high-pass and band-pass filter characteristics. The ascending neuron projects to low-pass and high-pass filter neurons, through the STF and the STD synapses. These two neurons converge to the band-pass filter neuron through the STD synapses. The structure of the first network is not shown to avoid ambiguity. The details of the STP implementation can be found in [97]. The network was simulated using a Brian Simulator [40] and emulated using a 'NeuroP' hardware (see Sec. 3.8 for more details of this hardware). Chirps of various PPs were presented with the fixed duty-cycles of 50% (similar to the one presented in the Fig. 41). The neurons were tuned to respond to the attractive PPs of the chirp. This implementation confirmed that the STP could be used to model the temporal filtering

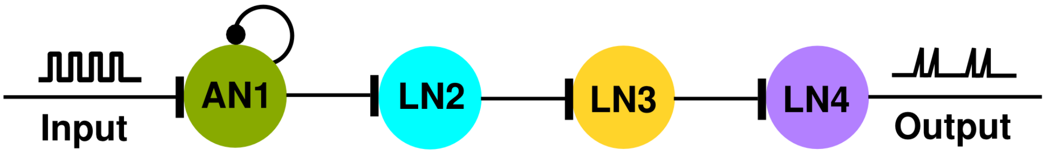


Figure 44: The spiking neural network modeled in the neuromorphic hardware to recognize the calling songs of crickets. The network consists of four neurons: the AN1, the LN2, the LN3 and the LN4. The bars at the end of the synapse between the neurons represent the excitatory synapse type. The circle at the end of the synapse represents the inhibitory synapse. The ascending neuron AN1 exhibits the SFA which is modeled using a feedback inhibitory synapse. Other local neurons: the LN2, and the LN4, exhibit the STF in their synapses which is modeled using the specific dynamics of the (DPI) synapse itself. The input pulses used to stimulate the network are generated off-chip. The output spikes are measured directly from the chip.

properties of the silicon neurons, similar to the observations reported in [101]. However, during the implementation of the high-pass filter response of a neuron, we found out that certain dynamics of STP cannot be achieved by the neuromorphic circuit. This finding led to the design of a new series of the STP circuits, proposed in [90]. The details of the problem and these circuits can be found in the previous chapter.

Next, we discuss the model of our latest network. A more recent work by Kostarakos and colleagues in [58], have shown the band-pass selectivity of the auditory neurons exists in response to duty-cycles variations of the PPs. Evidence of the neuron with high-pass filter properties to the duty-cycle variations of the chirp is unavailable in the neuroscience literature. Therefore, our previous network is redesigned based on the recent neurophysiological evidence [58]. We modeled our network using four types of neurons: the Ascending Neuron 1 (AN1), the Local Neuron 2 (LN2), the Local Neuron 3 (LN3), and the Local Neuron 4 (LN4). We used a simple naming scheme because the naming schemes of cricket auditory neurons are inconsistent in the literature [101, 58, 103].

We constructed a feed-forward spiking neural network shown in Fig. 44, based on the latencies and the firing rates of the cricket auditory neurons. These two features represent the flow of information in the network. Based on these features two networks were proposed in the literature [113, 103], as discussed in the previous section. The neurons of our network shown in Fig. 44 are color-coded: AN1 (green), LN2 (cyan), LN3 (yellow) and LN4 (purple). The excitatory synapses are represented as lines ending with bars, and the inhibitory synapse is shown as a line with a filled circle at the end. In our network, the filtering properties of the auditory neurons of crickets are modeled as computations in their synapses. We chose the auditory neurons of crickets reported in [58], to model in our network because their responses were characterized by various combinations of the PDs and the PIs with several duty-cycles, in [58]. The AN1 neurons exhibit the Spike Frequency Adaptation (SFA) (explained later in this section), and this is modeled by a feedback inhibitory synapse. The LN2, the LN3, and the LN4 are inspired by the local neurons B-LI2, B-LI3, and B-LI4 of crickets. The filtering properties of the LN2 and the LN4 neurons are modeled using the Short-Term Facilitation (STF) in their synapses. The filtering property of the LN3 is modeled using a conventional excitatory synapse without any plasticity.

The analog neuromorphic chips have inherent variability among the circuits called the 'device-mismatch' resulting from the fabrication process (refer Sec. 3.7 for more details). Special design techniques can be employed to reduce the device-mismatch. On the contrary, the device-mismatch can be utilized as well. Schmuker and his colleagues [102] analyzed the effect of the device mismatch and the temporal noise (due to the temperature effects on analog circuits) in the accelerated neuromorphic hardware during the implementation of a spiking neural network. They caliberated the hardware to reduce the mismatch and quantified the temporal noise by measuring the variability in the output spike count. They exploited the resulting variations as stochasticity in their network. Nevertheless, these mismatch effects differ across the neuromorphic chips. To demonstrate this variability across the neurons, we modeled twenty individual networks in our sub-threshold mixed-signal neuromorphic hardware (see Sec. 3.8 for further details of the hardware) as shown in Fig. 45. The networks are built using the same four types of auditory neurons as discussed before. Therefore, eighty neurons are used to form these networks, with twenty neurons of each type. The connection scheme (see Fig. 44) is preserved for each kind of neuron. Therefore, twenty independent networks were created.

The resulting feed-forward networks were constructed using the neuromorphic multichip setup (see Sec. 3.8 for more details). To be more precise, we used two 2DIFWTA chips in our implementation. The synapses were implemented using the DPI synapses [9] (refer Sec. 3.5 for more details) and the neurons were implemented using the low-power integrate-and-fire neurons [49] (refer Sec. 3.6 for further details) of the neuromorphic chip. The parameters are global and are shared across the arrays of neurons and synapse types in the chip. Therefore, the synapses and neurons cannot be tuned

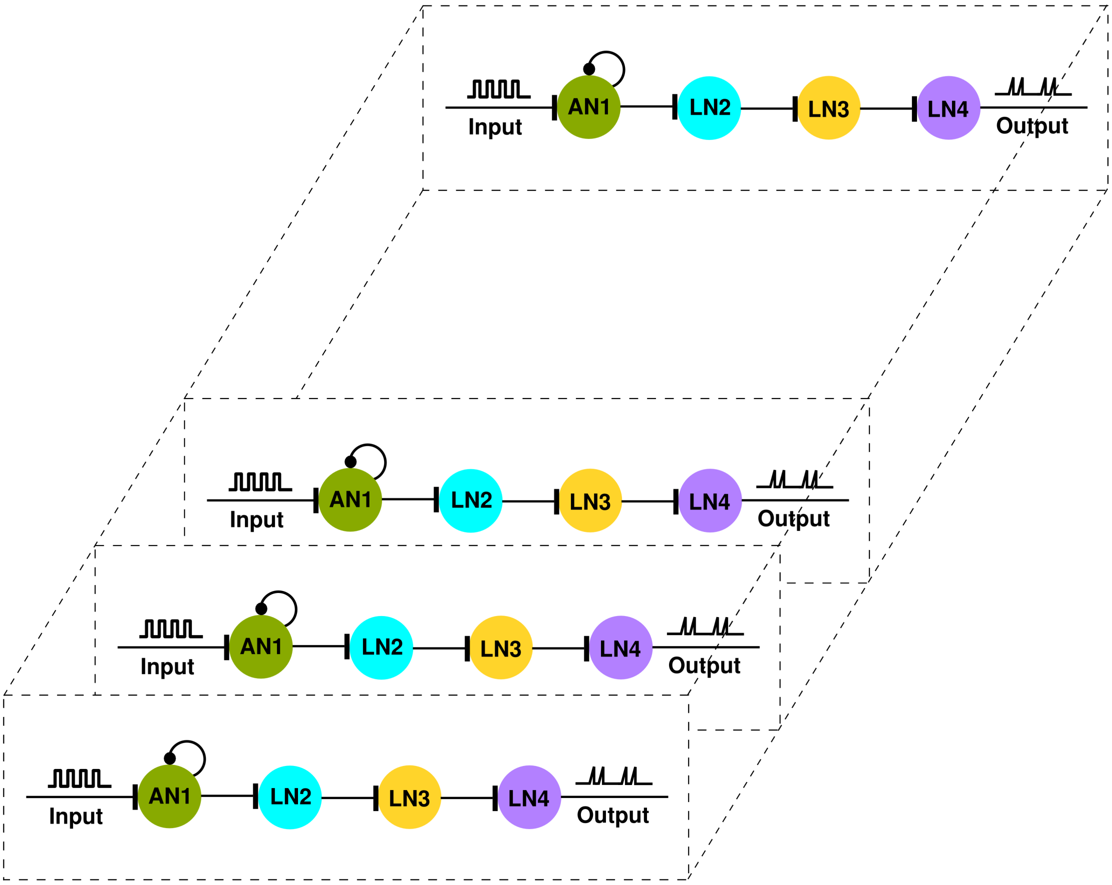


Figure 45: The architecture of the calling song recognition networks. Twenty independent networks are implemented in the neuromorphic hardware. All these networks are operated parallel in real-time. This arrangement exploits the mismatch effects of the hardware, making every network distinct.

individually. Four neuron types with one distinct synapse each are required to build our network. Each '2D' chip has two distinct excitatory synapse arrays that do not share their parameters. Therefore, we used two '2D' chips to implement four types of non-identical synapses.

Though the parameters are shared within each chip, the responses of the circuits vary within each chip due to the effects of 'device-mismatch', as we discussed earlier in this section. In this context, we consider the degree of variability in silicon circuits is analogous to the variability among testing different animals in biological experiments. The neurons from both the chips were calibrated beforehand to have the minimal mismatch in their responses. A chirp of 300 ms PD and 300 ms CD (not shown) is used to calibrate the neurons. 100 neurons (arbitrarily chosen) were picked from each chip. The parameters of the synapses were fixed. The neuron parameters (from both the chips) were tuned to match their mean firing-rates to a maximal extent. Despite this calibration, it is significant to mention here that the spike-times and the spike frequencies of the neurons always vary within the chip and across the chips. However, the calibration can be useful for exporting the network to other neuromorphic '2D' chips (provided that the chips are fabricated with the same technology).

In this research, the neuromorphic hardware is used to model the temporal dynamics of the synapses and to model the neurons. Biologically realistic time-constants in the order of 50 - 200 milliseconds [89] are necessary to model the temporal dynamics precisely. These time-constants are achievable in our neuromorphic chips, thanks to the sub-threshold operation of the transistors (see Sec. 3.3 for more details). Our neuromorphic chips operate in real-time and consume low-power compared to conventional processors. For instance, a typical silicon neuron with a 100 Hz firing rate consumes few µW of power [49]. The real-time operation of our neuromorphic hardware is useful, especially when the system is integrated with any real-time sensor (e.g., silicon retina [62] or silicon cochlea [65]) or used in a robot. On the other hand, it is difficult to incorporate any real-time sensor with the accelerated neuromorphic hardware due to the scaling of operating times. Nevertheless, all these neuromorphic platforms offer brain-like parallel computations in silicon and faithfully capture the temporal dynamics of the computational elements such as neurons and synapses. The silicon implementation of neurons and synapses provides the advantage of full parallelism. We take advantage of this property by simulating the twenty networks in parallel. Building a neuromorphic system that emulates the phonotaxis behavior of crickets can serve as a dedicated computational module for sound guided tasks, especially in bio-inspired robots.

5.3.3 *Computational primitives*

One of the goals of this research is to demonstrate the use of Short-Term Plasticity (STP) and Spike Frequency Adaptation (SFA) as computational primitives leading to the temporal filtering properties observed in biology. To this end, we designed a neural network by modeling the temporal dynamics found in the vital biological neurons reported in the literature. We provide an educated guess about the neural substrate of the cricket behavior in response to original songs and elucidate the role of the chosen computational primitives in producing the biological dynamics. The computational primitives are basic signal-processing mechanisms involved in recognition of the calling songs. In our network, this identification can be achieved by adequately tuning the Spike Frequency Adaptation (SFA) of a neuron and the STF of the synapses, the details of which will be discussed in the following subsections.

5.3.3.1 *Spike frequency adaptation*

Many neurons adapt their spiking responses to a sustained stimulation. Initially, their firing-rates are high which are subsequently reduced, and finally, the steady-state values are reached. This phenomenon is called SFA which may result from any of the following biophysical mechanisms: inactivating sodium channels; activating voltage dependent potassium currents during depolarization; enabling calcium dependent potassium currents following the hyperpolarization. In a network, the SFA can result from a negative feedback to the neuron, through the inhibitory synapse. Adaptation plays a significant role in bursting behavior of neurons. The adaptation currents can determine the neurons to act as resonators or integrators [45]. Combined with the Short-Term Depression (STD), SFA ensures the neurons sensitive to the temporal derivative of the stimulus [87]. SFA can also improve the reliability of information encoding [33, 34].

Among the auditory neurons of crickets, the AN1 neuron exhibits SFA in response to a continuous stimulation of pulses within the chirp [101, 113, 58]. The slow stimulus components depending on the intensity variance, are suppressed by the SFA. Therefore, the SFA of the AN1 creates an intensity invariance in the auditory pathway [11]. The SFA was implemented in our network through an inhibitory synapse, which is connected in a negative feedback loop with the silicon neuron. The number of the output spikes start decreasing after the onset of the SFA and stay constant after the SFA reaches the steady-state value. The SFA adds the high-pass filter characteristics to the transfer function of the AN1 neuron of our network. As a result, the AN1 neuron becomes less responsive to a sustained stimulus. For example, a chirp with a long PD equal to the CD and no PIs in between. The SFA property of the AN1 neuron is characterized as follows.

We characterized the Spike Frequency Adaptation (SFA) of the AN1 neuron during our first implementation of the network in [97]. We aimed to characterize the SFA using both the software and the hardware. However, the goal of the work is not to match the hardware results and the simulation results. The neuron is modeled in the software (using the BRIAN simulator) as well as in the hardware (using the IFSLWTA neuromorphic chip). A leaky IF neuron with SFA is used in the simulations. The parameters of the AN1 neuron are tuned to obtain the output firing-rate and the time-constant of the Spike Frequency Adaptation (SFA) based on the neurophysiological evidence. A step input current is used to stimulate the AN1 neuron. A small variability is introduced to the neuron by adding a filtered noise (modeled as an Ornstein-Uhlenbeck process) as an additional input current, and the simulations are repeated over 100 trials. Before

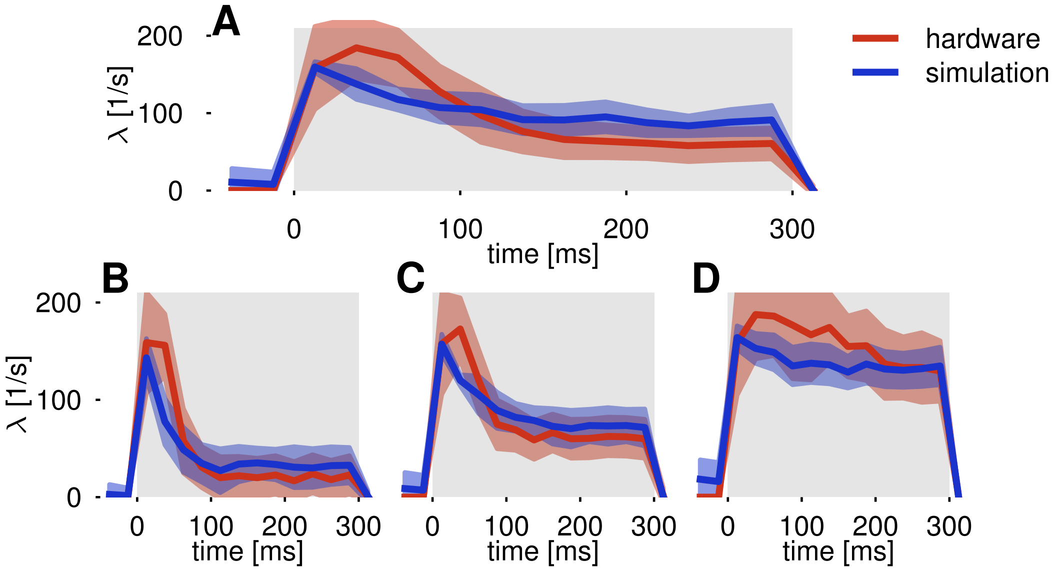


Figure 46: The SFA of the AN1 neuron characterized using BRIAN simulations (blue) and neuromorphic hardware emulations (red), published in [97]. The neurons are stimulated with a chirp of 300 ms duration (CD), consisting of only one PD without a PI (shown as the gray region). The mean firing rates (λ) (dark lines) are computed for 100 neurons in the hardware and 100 repeated trials during the simulation. The shaded region denotes the SD. (A) An ideal set of parameters are chosen to implement the SFA of the AN1. (B - D) Different adaptation behavior obtained during the simulation and from the hardware by varying the parameters of the synaptic weight and the time-constant.

we discuss the outcome of this model, let us first explain the hardware implementation. The SFA is implemented in the neuromorphic hardware, through a feedback inhibitory synapse to the neuron. When stimulated with the pulses (digital), the inhibitory synapse produces a negative current that is proportional to the output firing rate of the neuron [9]. In our implementation, the inhibitory synapse receives the neuron's output spikes and injects its output current back into the neuron. To characterize the response variability of the neurons of our hardware, we stimulated an array of 100 low-power leaky IF neurons with a 560 Hz input spike train for 300 ms in duration. The parameters of the neuromorphic chip are chosen to match the simulation results, concerning the peak firing rate and the adaptation characteristics, such as time-constants and steady-state frequencies.

The results of the characterization of the SFA (published in [97]) is shown in Fig. 46. The SFA profile of the neurons is analyzed for four sets of weight and time-constant values of the synapse. The results obtained both from the simulations (blue), and the hardware (red) are shown in the figure. The thick lines represent the mean of the firing rates, and the shaded region denotes the SD of the firing rates. These statistical measures are computed across different trials in the simulations. However, they are computed across the neuron array in the hardware. The time-constant of the feedback inhibitory synapse is long in the hardware. Therefore, the SFA profile of the hardware is slower compared to the simulations (see $0 \leqslant Time \leqslant 100ms$ in all subplots). The SDs of the SFA profile from the software and the hardware are also different. The SDs of the SFA simulation is almost absent at the peak of the mean frequencies and high at the low frequencies of the steady-state values. Unlike simulation results, the SDs from the hardware are high at the peak of the mean frequencies and are low during the steady-state values. This effect is visible in Fig. 46(B). The strong adaptation led to a decrease in the Standard Deviation (SD) due to the reduction in the total number of spikes. This way, the SFA filters out the low-frequencies in our network.

We characterized the AN1 neuron in our first model published in [97]. The function of the AN1 neuron remains the same in our latest network. Therefore, this characterization is still valid, despite the differences in parameter settings of the AN1 between these implementations.

5.3.3.2 *Short-term facilitation*

The Short-Term Facilitation (STF) is a synaptic enhancement process, during which the probability of release of the neurotransmitters is increased for a short period. The STF occurs due to the influx of calcium ions into the axon terminal, after the generation of a spike [115]. The STP influences the interaction of the pre-synaptic neuron with the post-synaptic one, resulting in a selective communication between the neurons [54]. The STP tunes the synapse to be a low-pass or a high-pass filter by changing the release probability of the neurotransmitters [1]. For more computational properties of the STP, refer Sec. 2.3. The STP has been reported in the primary visual cortex of the mammalian brain [108], and the avian auditory system [69] (see Subsec. 2.3.5 for more details). The STP was also reported in insects, especially the mushroom bodies of fruit flies (*Drosophila*) [106] and honeybees [75]. However, the presence of STP in crickets' auditory system is not evident from the literature. Nevertheless, this research provides a bottom-up approach to verify the presence of STP in cricket brain by demonstrating the computations of the auditory system in the neuromorphic hardware. We implemented the STF in the synapses of the LN2 and the LN4 neurons. The

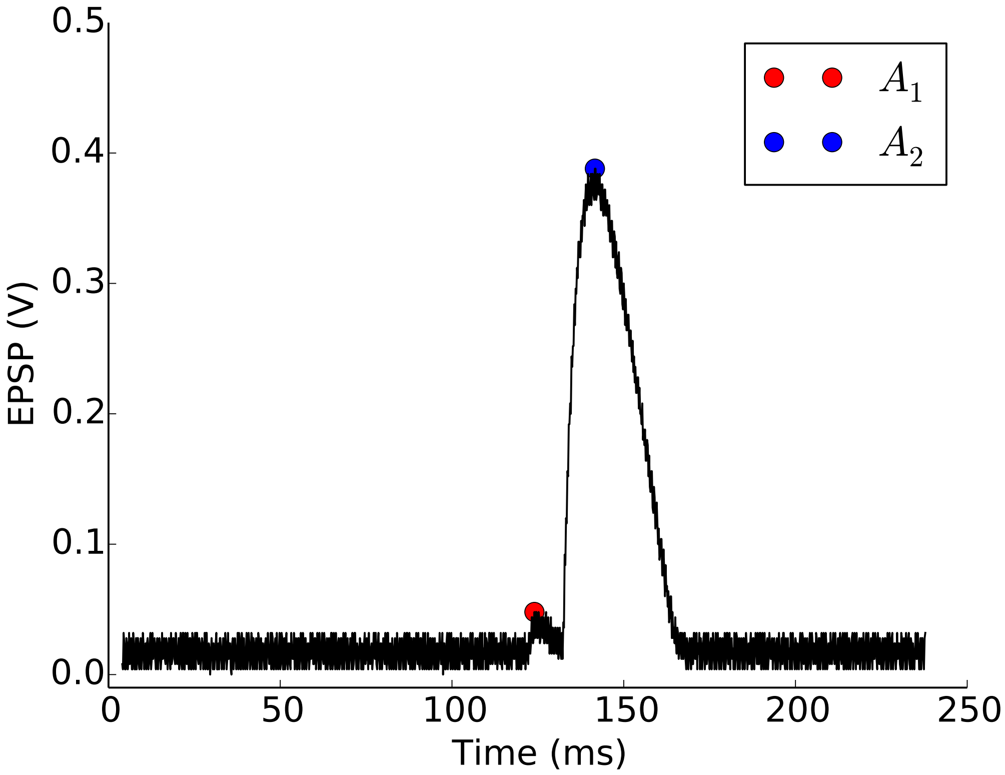


Figure 47: The response of the EPSP of a neuron with the STF synapse to two input pulses stimulated at 10 ms ISI and recorded from the oscilloscope. The synapse and the neuron are tuned in a way, that the EPSP stays below the spiking threshold. The red dot denotes the first peak (A_1) of the initial rise in the EPSP, responding to the first input pulse. The blue dot denotes the second peak (A_2) in the EPSP, responding to the second input pulse. The A_2 is larger than the A_1 due to the STF synapse. The response is time-shifted (moved to the right) to avoid a negative time in the trace, caused by the oscilloscope trigger.

STF is implemented using the neuromorphic Differential-Pair Integrator (DPI) synapse (refer Sec. 3.5 for more details of the synapse). The synapse implements STF by operating with the parameter setting of a small weight and a long time-constant. The time-constants of the STF synapses differ between the LN2 and the LN4. We used the STF to model the band-pass filter characteristics of the LN2 and the LN4 neurons.

We characterized the STF in the neuromorphic hardware. As mentioned earlier, STF is implemented by choosing a small-weight and a long time-constant of the DPI synapse. This parameter setting results in a build-up of EPSC (see Sec. 3.5 for more details of the EPSC response). An array of twenty low-power integrate-and-fire neurons are stimulated in the hardware, to characterize the response variability. The stimulus consisted of two digital input pulses. The ISIs are varied from 10 ms to 100 ms in steps of 10 ms.

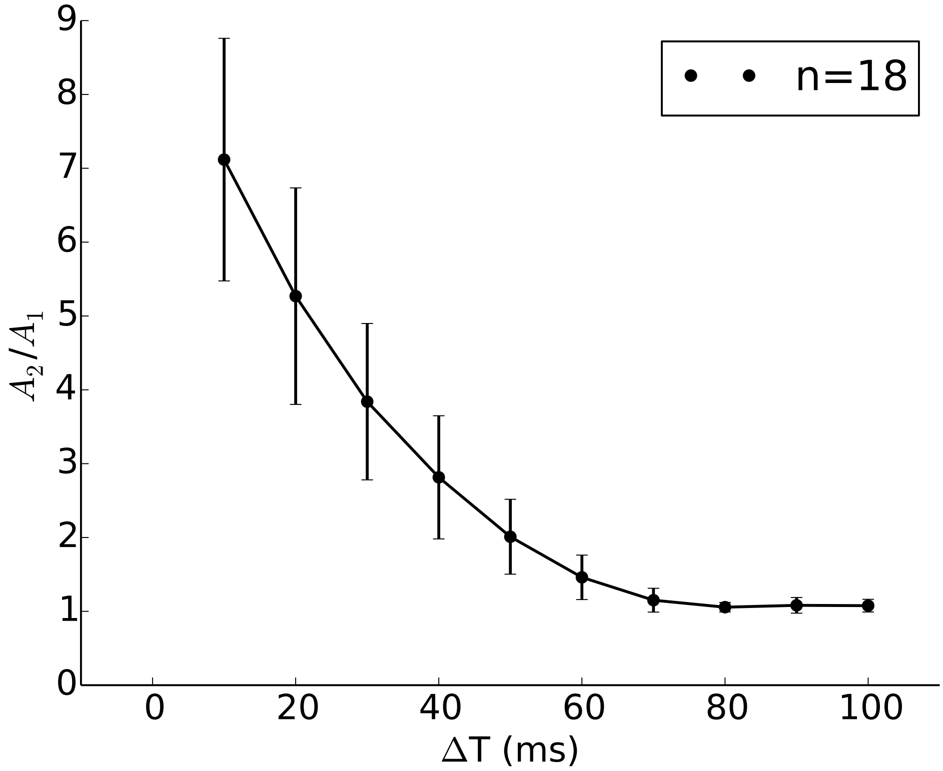


Figure 48: The ratio of A_2/A_1 of Fig. 47, called the PPR, is computed and plotted as a function of the ISI ΔT of the input pulses. Two input pulses of varying ISIs are used to stimulate twenty neurons. Out of twenty, two neurons are spiking, and the responses of those two neurons are omitted in the plot. Each point in the plot represents the mean PPR across the neuron array, and the error bar represents the Standard Deviation (SD). The PPR is high for short ISIs due to the build-up of the increasing EPSCs' amplitudes. The EPSCs almost recover back to their initial values during the long ISIs resulting in small PPR. The profile of the PPR shows high-pass characteristics of the STF to input frequencies.

The neurons are tuned such that their EPSPs rise in response to the incoming synaptic currents. At the same time the EPSPs stay below the spiking-threshold of the neurons. The EPSPs of the neurons are recorded directly from the chip using the oscilloscope. A sample trace of the EPSP is shown in Fig. 47. The EPSP trace is recorded from one of the neurons with the STF synapse when stimulated with two input pulses of 10 ms ISI. The onset delay due to the oscilloscope trigger is compensated by offsetting the response in time to avoid the negative time in the trace. The effect of the STF is evident from the amount of increase in the size of the EPSP peak of a neuron.

The PPRs are computed for each ISI of the stimulus. A PPR is the ratio between the second peak of the EPSP and the first peak. Out of twenty neurons of the array, two neurons are spiking during the presentation of the stimulus. Therefore, the PPRs are not computed for those two neurons. The PPRs of the remaining non-spiking neurons (18) are plotted as a function of the ISIs (ΔT) in Fig. 48. Each point in the plot represents the mean PPR, averaged across the neuron array. The error-bars denote the SDs. In response to a stimulus with the shortest ISI of 10 ms, the synapse is highly facilitated. The strong facilitation strength results from the slow build-up of increasing amplitudes of EPSC (with a long time-constant). The absolute amplitude of the second EPSP peak is much higher than the absolute amplitude of the first peak. Therefore, the mean PPR is substantial during this ISI. The amount of increase in the EPSP is not uniform throughout the neuronal array due to the device-mismatch. As a result, the SD of PPR for this ISI is large.

In response to a stimulus with the longest ISI of 100 ms, the amplitude of the EPSCs recovers completely from being high (due to the facilitation), before the onset of the second pulse. As a result, the amplitude of the second EPSP peak is almost equal to the amplitude of the first peak. Hence, the mean PPR almost reached the value of one. More interestingly, irrespective of the device mismatch effects, the EPSCs are recovered to their initial values almost completely. As a result, the SD for the 100 ms ISI decreases. From the characterization of the STF (see Fig. 48), we can conclude that the STF exhibits high-pass filter characteristics to the input frequencies, similar to the characterization of the STF using a software model presented in [107]. This property is useful in shaping the responses of the neuron in our network.

5.4 RESULTS

To analyze the response of our network, we presented a chirp consisting of 800 Hz digital pulses to stimulate all twenty networks (see Fig. 43),. The Chirp Duration (CD) was fixed to 260 ms. The Pulse Durations (PDs) and the Pulse Intervals (PIs) were varied from 5 to 49 ms, in steps of 4 ms. The neuron parameters were calibrated and fixed. The synapses were tuned for each neuron type, to obtain the desired filter responses of the neurons based on the neurophysiological evidence from [58]. The responses of each neuron type are presented in this section. We present the neural dynamics regarding membrane potentials recorded from one of the networks and the spiking responses from all the neurons of all the networks in the raster-plots. Finally, we show the filter-like responses concerning spike-count of the neurons from one of the networks and the variability in the filter responses by plotting the spike-count of one neuron type (LN4) from all the networks.

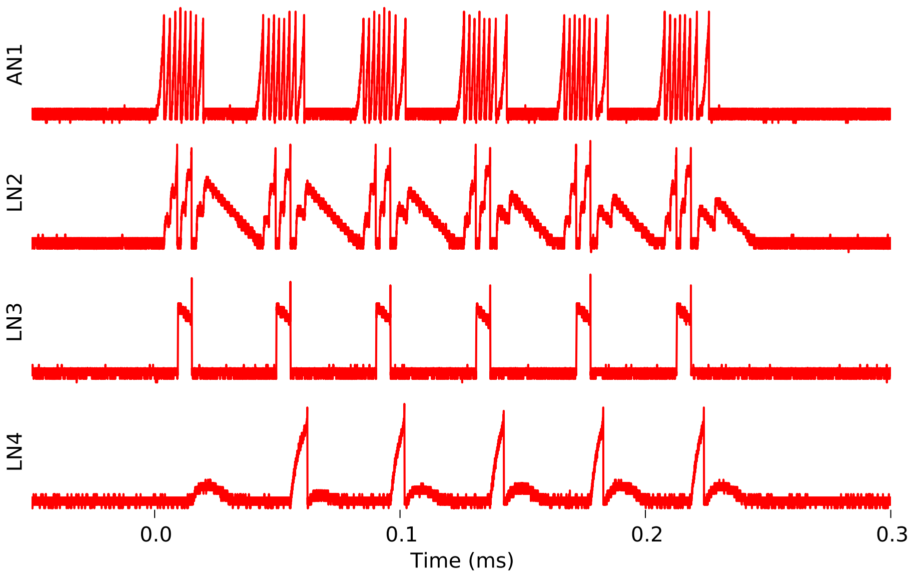


Figure 49: The membrane potentials recorded from the neurons of one of the networks. The stimulus is a chirp of 260 ms CD made of 800 Hz evenly distributed pulses of 20 ms PD and 20 ms PI (stimulus not shown in the figure). The AN1 neuron responded with 8 pulses to each PD. The SFA of the AN1 is visible at the last AN1 spike of every PD. The LN2 neuron responded to every third spike of the AN1 through the facilitating (STF) synapse. The amplitude of the second peak of LN2's EPSP is higher than its first due to the STF. The LN3 responded to every second spike of the LN2 through a regular excitatory synapse. The LN4 responded from the second input spike of the LN3 with the ISI of a PI, due to the slow STF. The slow rise in the EPSP after the spike is the result of integration of the EPSC by the neuron because the EPSC did not recover during the PIs due to the slow STF.

Let us start by discussing the neural dynamics of the network, shown in Fig. 49. We presented a 20 ms PP stimulus (not shown), and the membrane potentials were recorded for all neurons of one of the networks. This specific PP is chosen, as it evokes one of the high responses of the network. The AN1 synapses were tuned with low-weight values to output a 200 Hz spike frequency to the high-frequency input. The high-frequency input was used to tune the inhibitory synapse to invoke a strong feedback to implement the SFA of the AN1. In other words, it is a constraint posed by the inhibitory synapse circuit to operate in the desired regime. The time-constant of adaptation is slower than the given PD. Therefore, the AN1 neuron shows a minimal adaptation towards the end of every PD. The LN2 neuron integrates the EPSC through its STF synapse from the AN1. The result is a rise in the absolute amplitude of the LN2's EPSP to every input spike during the PDs. The EPSP crosses the threshold to spike (and spike) for every third input spike, given that specific ISI from the AN1. The ISI of the input spikes was long at the end of the third PD due to the SFA of the AN1. As a result, the STF strength is reduced (see Fig. 48) and the absolute amplitude of the EPSP peak becomes smaller than its value during the short ISI. Hence, the EPSP can reach the resting-state values during the PIs after the third PD. The LN3 neuron integrates the EPSC through the conventional excitatory synapse from the LN2. The LN3 spiked for every second input spike with that specific ISI from the LN2. The time-constant of the STF in the synapse of the LN4 neuron was tuned to be slower than that of the LN2's. As a result, the EPSP of the LN4 can cross the spiking threshold only after the first PD and the PI (or the first PP). Therefore, the LN4 neuron responded to every second spike of the LN3 neuron whose ISI is $\geqslant 20\text{ms}$ (also the PI). However, the facilitated EPSC of the LN4 did not recover during the PIs. The LN4 integrates this slowly recovering EPSC, and as a consequence, its EPSP rises after every output spike (in the absence of input spikes).

It is significant to note that the synaptic time-constant is tuned to be slow to implement the gradual time-constant of the STF. In this case, the STF is coupled to the synapse, meaning that we used the synapse to implement the STF, as we did not have any dedicated STF circuit during this implementation. However, this problem can be solved by a specific STF circuit which we designed after this implementation (see Sec. 4.4). Nevertheless, through this STF implementation, we were able to achieve the desired response from the network. That is, the total number of spikes within the CD decreases at each stage of the network and the latency of the onset of the spikes increases through the network. These two outcomes of our results coincide with the idea of the network models proposed in the literature [113, 103]. The above-presented results showed us the dynamics of the neurons from one of the networks in response to a chirp. Next, we are going to analyze the responses of the neurons of all the networks to three stimulus patterns.

We presented three variants of a chirp of 260 ms CD and 800 Hz frequency to the networks. We plotted the spike-times of the neurons of all the networks in response to these three chirps in the raster-plots shown in Fig. 50, Fig. 51, and Fig. 52. The raster plots are organized by the neuron type from top to bottom. Horizontal bars in grey separate the neuron types. The PD of the stimulus was varied between 5 ms (shown in blue stripes), the 25 ms (orange stripes) and the 49 ms (green stripes). The duty-cycle was fixed at 50%. Therefore the PIs were equal to the PDs. The PIs are shown as white spaces in between the colored stripes of PDs. Let us analyze the responses of the networks one-by-one for each stimulus.

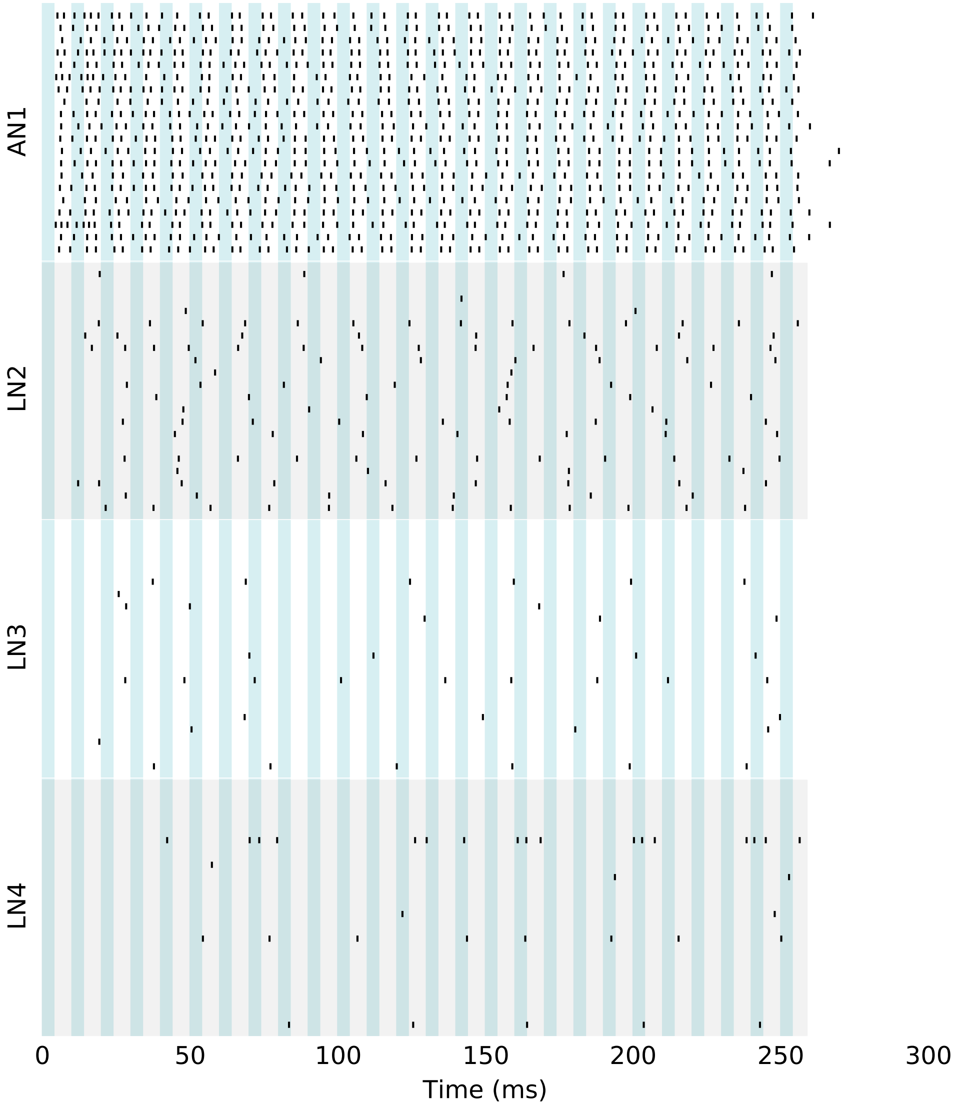


Figure 50: Rasterplot of the neurons of all the networks, in response to the 10 ms PP stimulus with the constant duty-cycle. 5 ms PDs are shown in the blue stripes, and 5 ms PIs are represented as the white vertical stripes in between. The neuron types are distinguished using the gray horizontal bars. The spikes are represented as small vertical bars. The n^{th} neuron of every neuron type in the plot belongs to the n^{th} network, thereby preserving the network structure.

The raster-plots of the networks to the 10 ms PP stimulus are shown in Fig. 50. Most of the AN1 neurons do not show the SFA because the PD is shorter than the time-constant of SFA. The AN1 neurons exhibit the low frequency of output spikes in response to the high input frequency, due to the low-weight parameter setting of the AN1 synapses, as discussed earlier. As a result, the latency of approximately 5 ms occurs during the onset of the AN1 spikes. It is evident from the figure that all 20 neurons of the AN1 elicit spikes to this shortest PP stimulus. The EPSC amplitudes of the LN2's STF synapses build-up shortly during the short PDs. The EPSCs recover almost wholly (less than or equal to the amount of the neuron's leak current) during the PIs. As a result, the LN2 neurons elicit a small number of spikes compared to the AN1s. The EPSPs of the LN3 also recover almost entirely during the long ISIs of the LN2 spikes. Therefore, there is only a small or no build-up of the EPSPs that cross the spiking threshold resulting in almost no spikes from the LN3. The long ISIs of LN3 spikes or no LN3 spikes reduce the overall activity among the LN4 neurons, due to the complete recovery of EPSPs between the two input spikes. Few of the LN4 neurons begin to spike after the second input spike, due to the STF synapses. In this case, the ISIs between the input spikes are shorter than the time-constants of the STF. A prolonged recovery of the EPSC after the first spike (or a low threshold) causes one of the LN4 neurons (topmost LN4) to respond with more output spikes than its input spike-count (starting from the second input spike). The discrepancy between the response profiles of the neurons within the same neuron type occurs due to the device mismatch effects. From the figure, we can infer that the networks (LN4 neurons) filter out these high-frequency inputs.

Next, we present the networks' responses to the 50 ms PP stimulus as shown in Fig. 51. Despite the AN1 neurons, all other neurons types spike more to this stimulus than to the 10 ms PP. The AN1 spike counts within the PDs are sufficient to trigger the negative feedback loop of the SFA. Therefore, the AN1 neurons begin to adapt at the end of the PDs. The AN1 neurons elicit enough number of spikes with close ISIs, to elicit more than one LN2 spike during the PDs. As a consequence, the LN3 neurons output at-least one spike for every two input spikes during the PDs. The LN4 neurons respond after two input spikes from the LN3 with the ISI $\geqslant 50\mathrm{ms}$, due to the slow STF as discussed earlier. The effects of the device mismatch modify the total spike count and the onset of the output spikes with each neuron type. Nevertheless, the networks (LN4) find this stimulus attractive, by responding with high spike-counts.

Finally, we present the networks' responses to the 98 ms PP stimulus as shown in Fig. 52. Contrary to the responses shown in Fig. 50 and Fig. 51, the SFAs are visible in the AN1 output spikes. The LN2 neurons respond similarly to their responses to 50 ms PP stimulus in Fig. 51, before the onset of the SFA of the AN1. After the onset of the SFA, the ISIs of the AN1 spikes increase. The EPSCs of the LN2s' STF synapses recover almost completely (less than or equal to amount of the neuron's leak current) during these long ISIs, thereby, resulting in a small number of LN2 output spikes after the

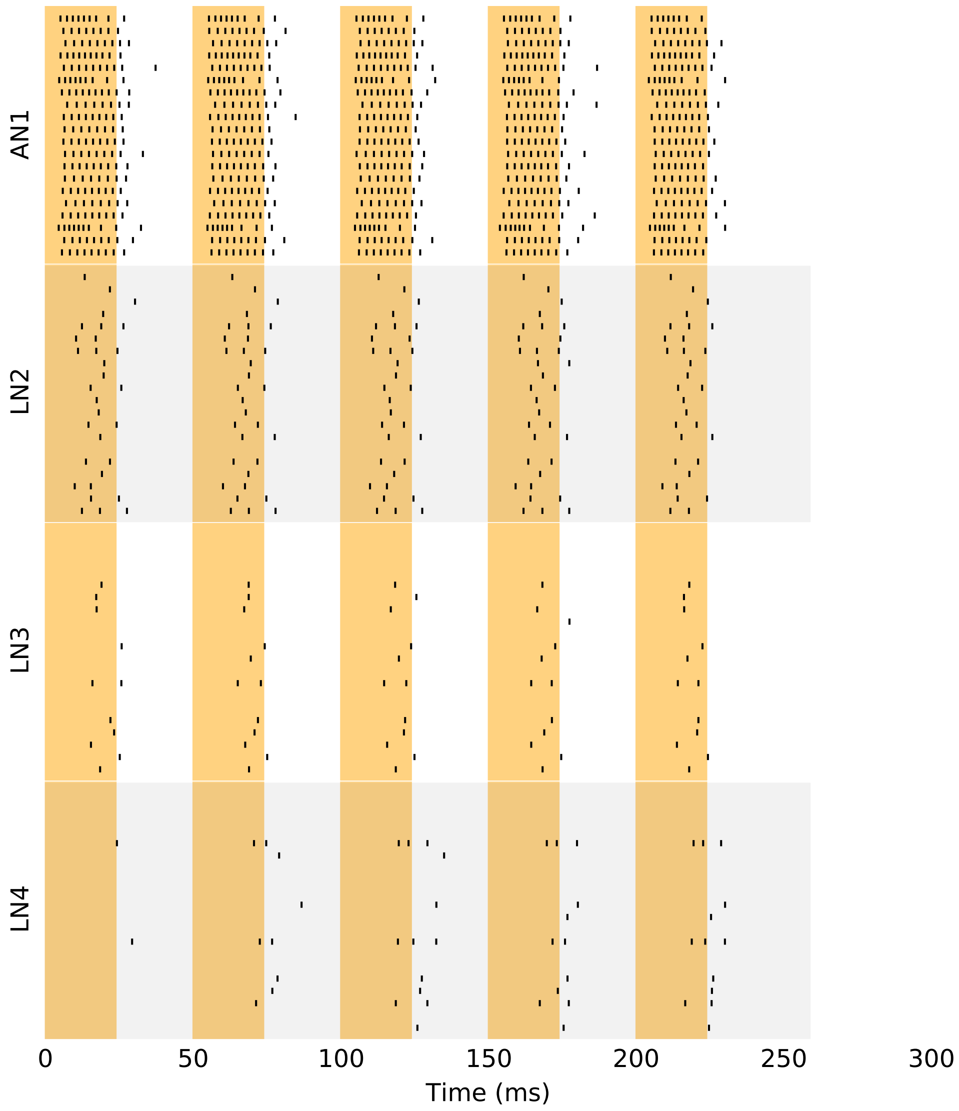


Figure 51: Rasterplot of the neurons of all the networks, in response to the 50 ms PP stimulus with the constant duty-cycle. 25 ms PDs are shown in the orange stripes, and 25 ms PIs are represented as the white vertical stripes in between. The neuron types are distinguished using the gray horizontal bars. The spikes are represented as small vertical bars. The n^{th} neuron of every neuron type in the plot belongs to the n^{th} network, thereby preserving the network structure.

AN1 adaptation. The LN3 neurons respond in the same way to the 50 ms PP stimulus. However, the total number of the PDs are smaller (three) during the 98 ms PP stimulus, compared to the (five) 50 ms PP stimulus of Fig. 51. The number of PPs are varied to

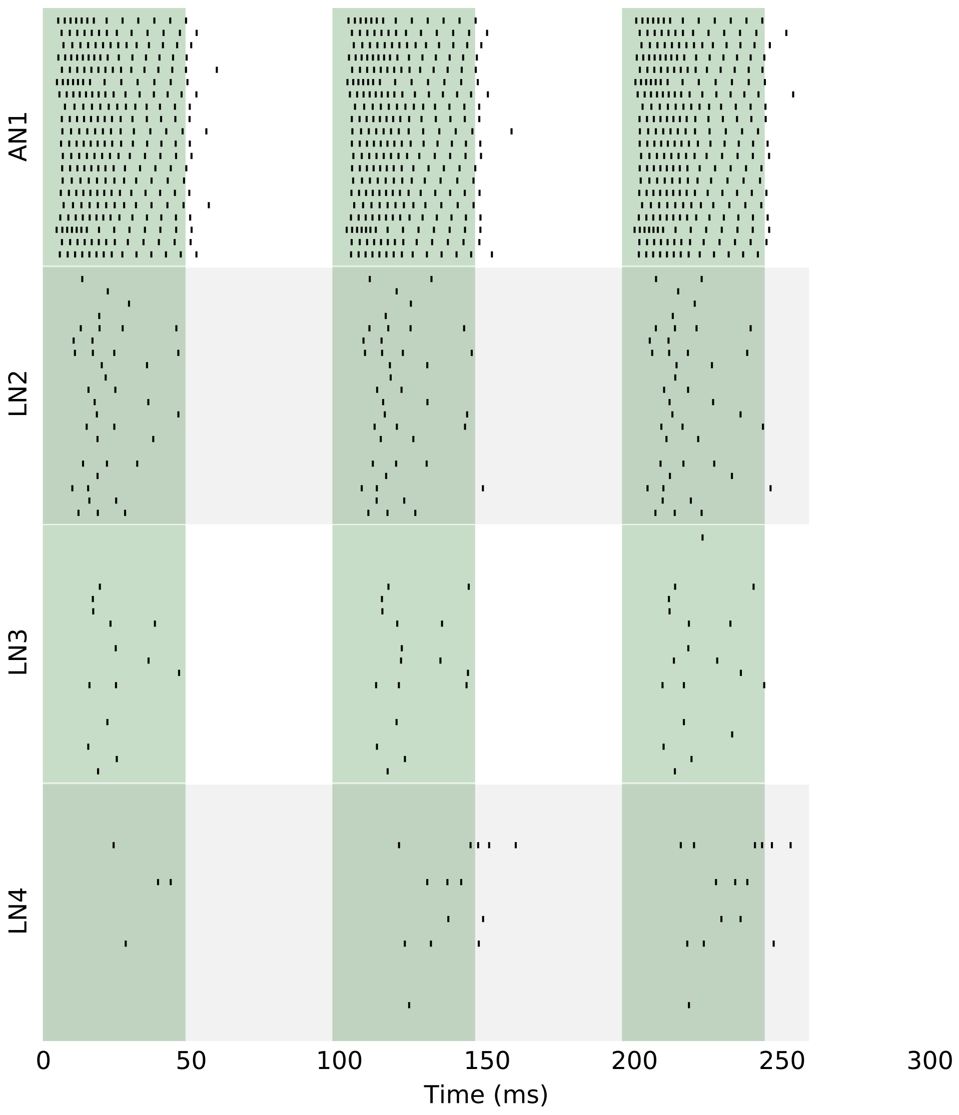


Figure 52: Rasterplot of the neurons of all the networks, in response to the 98 ms PP stimulus with the constant duty-cycle. 49 ms PDs are shown in the green stripes, and 49 ms PIs are represented as the white stripes in between. The neuron types are distinguished using the gray horizontal bars. The spikes are represented as small vertical bars. The n^{th} neuron of every neuron type in the plot belongs to the n^{th} network, thereby preserving the network structure.

maintain the constant CD in a chirp. As a result, the overall spike count of LN3 decrease during this longest PP stimulus.

Many LN4 neurons are silent during this stimulus because the PIs are long that the EPSCs recover almost wholly (less than or equal to the amount of the neuron's leak current) without allowing the facilitation. However, there are a few LN4 neurons that spike after two input spikes using slow STF due to the mismatch effects. Nevertheless, the networks (LN4 neurons) respond with a smaller spike count than that of the 50 ms PP stimulus (see Fig. 51). This way, the networks filter out these low-frequency inputs.

To this extent, we examined the responses of the networks to three variants PP of a chirp. Further, we present the network responses characterized by each neuron type, to more PPs of a chirp. We introduced a chirp whose PPs are varied from 10 ms to 98 ms in steps of 8 ms, by keeping the duty-cycle constant at 50%. We measured the total spike count of each neuron type of all the networks for each PP variant of the stimulus. We computed the mean and the SD of these measures across the neuron array of each neuron type and plotted these values in Fig. 53. Each point in the black curve represents the mean of the total number of spikes from each neuronal type in response to a particular PP variant of the stimulus, and the error bars represent the SDs among the neurons of the same kind.

As we discussed earlier, the SFA of the AN1 is evident during long PPs of the chirp. Therefore, the mean spike-count of the AN1 decreases from approximately 50 for the 10 ms PP to approx. 30 for the 98 ms PP. The resulting mean spike-counts display the slope with two peaks at 10 ms PP and 34 ms PP. The neurons respond high to the high-frequency stimulus due to the large number of the 10 ms PP within a chirp. The neurons begin to adapt after the 34 ms PP, resulting in a drop in the mean spike-count. The decline in the slope explains the high-pass filter like the response of the AN1 to the non-attractive stimuli.

The LN2 neurons flatten the slope of the AN1 responses, resulting from the strong SFA and the reduced PP count within the chirp. The maximum mean spike-count of the LN2 is smaller than the AN1's due to the combined effects of the facilitating synapse and the integrating neuron as discussed earlier. The number of the LN2 spikes within the PDs increases for long PPs. However, the spike count of LN2 decreases for long PPs after the onset of the SFA. Meanwhile, the number of the PDs declines for long PPs. As a result, the LN2 neurons display two peaks in their mean responses, the highest peak at 34 ms PP and the second highest at 70 ms PP. The mean response profile resembles a band-pass filter like the response. Therefore, the band-pass selectivity towards the attractive stimulus appears already at this stage of the network.

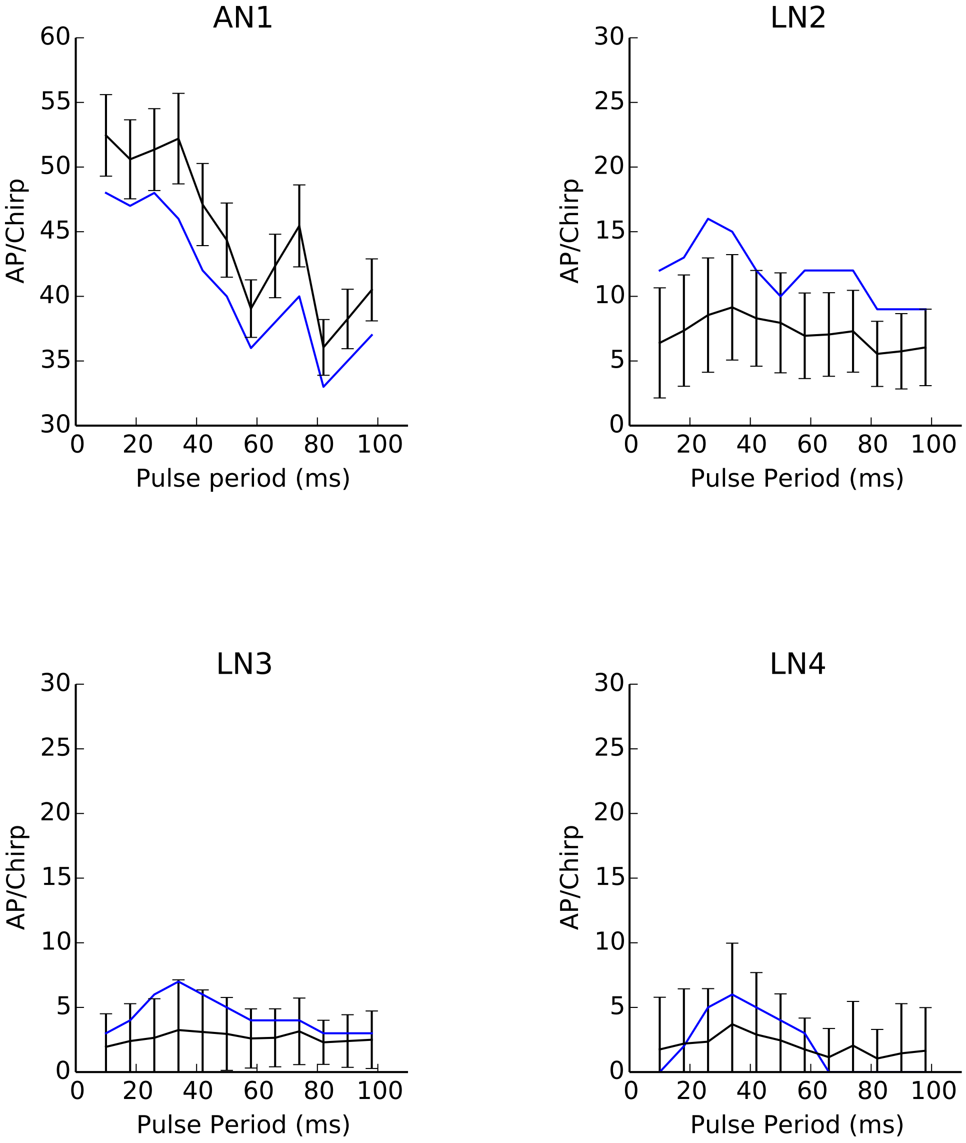


Figure 53: The response profiles of the networks characterized by each neuron type, in response to a chirp stimulus. Each point in the curve represents the mean (in black), and the error bars represent the SD of the total sum of action potentials of each neuronal array. The stimulus varies in the PPs with a fixed duty-cycle of 50%. One of the networks' response (shown as blue curves), whose LN4 neuron best represents the neurophysiological recordings of BNC2a presented in [101] and B-LI4 in [57]. The AN1's y-axis is offset by 30 to keep the scale of the y-axis constant for all the neurons.

Four of the LN3 neurons are silent, and their zero responses are included in the calculation of the mean. Therefore, the mean spike-count of the LN3 is almost constant to all PPs. The LN3 start eliciting a spike per PD if the PD $\geqslant 34$ms. Therefore, the SD of the

LN3 spike-count showed a peak value at 34 ms PP. For the long PPs, the total number of the PPs decreases within a chirp resulting in a decrease in the total spike-count of the LN3 neurons. Therefore, the SDs of the LN3 show a weak band-pass profile, for increasing PPs. The resulting band-pass filter has a wide bandwidth because the LN3 neurons respond minimally to the long PPs with at-least one spike per PD.

The input spike-counts are small during the short PPs. The EPSPs reach the resting-state values during the long PPs. As a result, eight out of twenty LN4 neurons is completely silent due to the mismatch effects. Therefore, the absolute value of the SD is higher than the mean at 34 ms PP. The mean spike-counts (and the SDs) of the spiking LN4 array show a stronger band-pass filter profile than the LN3 for increasing PPs. Therefore, the LN4 neurons prefer the 34 ms PPs over the other. The band-pass filter profile of the LN4 neuron differs from one another due to the device match. Therefore, averaging these responses results in a broad bandwidth of the band-pass filter profile. The individual response of the LN4 neurons will be discussed later in this section. Nevertheless, we present the results of one of the networks whose LN4 responses closely matched the neurophysiological evidence presented in [101] and [57], to understand the individual response to the PP variations with a constant duty-cycle. In this case, it is the fifth network's LN4 neuron (in the order) among the twenty networks.

The blue curve in each subplot of the Fig. 53 corresponds to this network's responses. The responses of the AN1 neuron follow the trajectory of the AN1 mean response and remain within the SD values of AN1. However, the responses of the LN2, LN3, and LN4 neurons of this network fall outside of their corresponding SD values. The LN2 response shows the second highest peak at 34 ms PP (first peak at 18 ms PP). The LN3 response exhibit a wide band-pass profile, with the maximum at 34 ms PP. The LN4 response displays a narrow band-pass profile, retaining the LN3's peak at 34 ms PP. Therefore, the network is tuned to select this PP of the stimulus. The 34 ms peak of the band-pass profile of this LN4 neuron is consistent with the values from the neurophysiological recordings of BNC2a presented in [101] and B-LI4 in [57].

We analyze the responses of this particular network we picked earlier (responses shown as the blue curves in Fig. 53), for several PPs with multiple duty-cycles. We presented a chirp, whose PD was varied from 5 ms to 49 ms in steps of 4 ms and the PI was varied from 5 ms to 49 ms in steps of 4 ms. We recorded the spikes from the neurons of all the networks in response to all 144 combinations of the PDs and the PIs. Considering the abundant number of the data, we present in the Fig. 54, only the interpolated data of the total spike count of neurons of the network whose LN4 responses closely matched

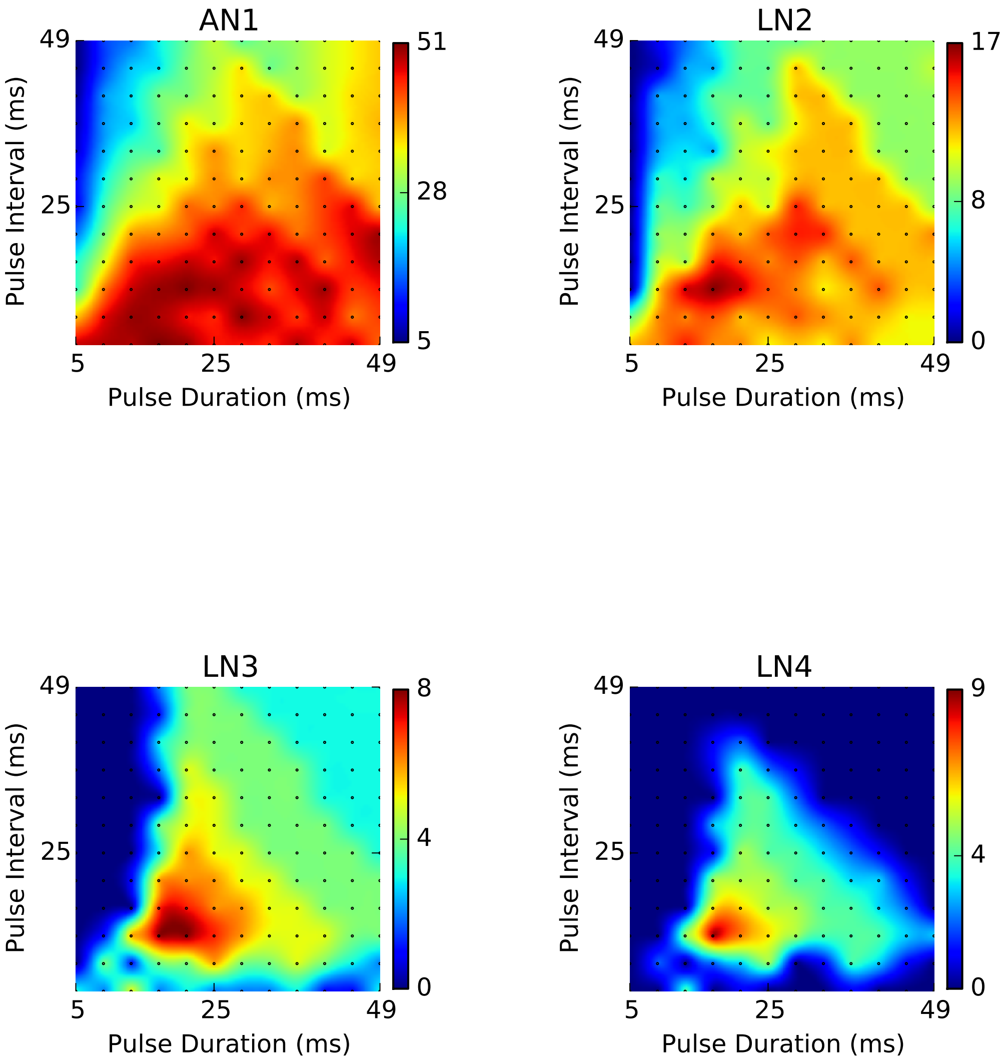


Figure 54: The heat-map plots showing the activity of neurons from one of the networks (fifth), the LN4 of which best represents the neurophysiological evidence of BNC2a presented in [101] and B-LI4 in [57]. The PD and the PI of the chirp are varied for different duty-cycles. Each point in the heat-map represents the total number of the action-potentials per chirp. The regions in between the points are interpolated. The color-code of the plots is drawn next to each of the plots. The maximum, the minimum and the median of the activity are marked in red, blue and green accordingly.

the neurophysiological recordings. These responses of this network are sufficient to explain the functionality of all the networks. The difference between the plots in Fig. 53 and the plots in Fig. 54 is that the duty-cycle of the PPs were kept constant at 50% in

Fig. 53, whereas the duty-cycle was varied in Fig. 54. The secondary diagonal of each subplot shows the spike-counts with 50% duty-cycle, which corresponds to the blue curve in Fig. 53. The heat map represents the maximum spike-count in red and minimum spike-count in blue, and their actual numbers are shown in the color-bars. The response of the network is already discussed for the stimulus with fixed duty-cycles (secondary diagonals). Therefore, the rest of the network activity will be discussed here.

For short PDs and the long PI, the number of PDs within the chirp is small. As a result, the AN1 neuron show low responses to the 5 ms PD and 49 ms PI. The activity in the top right corner of long PDs and long PIs is not high due to the small number of PDs within the chirp. For PDs longer than the PIs (except at 4 ms), more than one spike is elicited by AN1 during each PD (in most cases). Therefore, the AN1 show high responses in the regions of all the PDs when the PI was short (5 ms). In the lower middle region of the 30 ms PP, the PDs are shorter than the time-constant of the SFA, and the number of the PDs is high within the chirp. A large number of spikes with short ISIs are elicited here, marking the region of the maximum activity. The SFA creates a high-pass filter effect on the AN1 by shifting the maximum activity towards short PDs and lowering its activity during long PDs. This way, the preference for the PDs and PIs starts at the first stage (neuron) of the network.

For 5 ms PDs and 15-49 ms PIs, the incoming spikes from AN1 are small in number within each 5 ms PD and the PIs were longer than the time-constant of the STF, resulting in a small EPSC that does not elicit output spikes from the LN2 neuron. The LN2 neuron showed a small response during the short PDs due to the low-pass filter property of the STF. The LN2 neuron exhibited a minimal response to the AN1's adapted spikes with long ISIs during the long PDs as discussed earlier. Therefore, the LN2 activity in the region of the 5 ms PIs and all the PDs is less than the AN1's response. The AN1's adapted spikes have a minor impact on the LN2's response during long PDs and long PIs. Hence, the top right region of the plot showed an intermediate activity (in green), while the total spike-count of this region remained almost the same. The maximum spike-count of the LN2 neuron is smaller than the AN1 (as expected due to the STF). The overall shape of the LN2 neuron's heat-map profile resembles the AN1's. The major differences are: the region with the maximum activity of the LN2 is narrow compared to the wide-spread region of the AN1; the minimum of the spike count of LN2 is zero, whereas the AN1's minimum value is five. The LN2 neuron shows a weak band-pass profile in the primary diagonal but not in the secondary diagonal.

For 5-21 ms PDs and the 17-49 ms PIs stimulus, the input spikes from the LN2 are minimal in number, resulting in small EPSPs that stay below the threshold of the LN3. The LN3 neuron outputs one spike within each PD to every two incoming spikes from the LN2. As a result, together with the small number of spikes from the LN2, the LN3's responses in the regions of the 49 ms PD and all the PIs are significantly reduced. The high-pass filter effect is created across the primary diagonal by the combined property of the LN3's one spike per PD and the small number of the PDs during the long PDs. The structure of the stimulus played a significant role in shaping the LN3 neuron's activity profile. The maximum activity of the LN3 was more centered around the region of 30 ms PP, compared to the LN2. The LN3 neuron shows weak band-pass profiles across both the primary and the secondary diagonals.

The LN4 shows no responses in the regions of 5 ms PDs and all the PIs because the responses are already suppressed at the LN3. The responses of the LN4 in the regions of the 5 ms PIs with all the PDs are almost zero due to the following reasons: small responses from the LN3 neuron; slow build-up of the EPSC amplitudes due to the STF; the EPSCs recovered almost completely during the long PIs. The LN4 activity on the top right corner was canceled out by the combined effects of the STF and a small number of the PDs. The LN4 neuron exhibits active band-pass filter like responses across both the primary and the secondary diagonals, with the peak activity being centered at the region of the 30 ms PP. The band-pass profile of the LN4 neuron closely resemble the values from the neurophysiological recordings of BNC2a presented in [101] and B-LI4 in [57].

In comparison to the LN4's peak obtained when the duty-cycle is kept constant (blue curve in Fig. 53), which is at the 34 ms PP, the new peak obtained from the duty-cycle variations (Fig. 54) is centered at the 30 ms PP. This profile is more accurate than the fixed duty-cycle responses, as the responses are analyzed for more features of the chirps. The presented band-pass filter profile represents the response of one of the networks. However, due to the device mismatch effects, the filter response varies across the other networks. Let us analyze the responses of the LN4 neuron (band-pass filter neuron) from all the networks.

To demonstrate the variations in the responses across the networks, we present the band-pass filter responses of all the networks (LN4s) in Fig. 55. The total spike-counts of the LN4 neurons within chirp are plotted, in response to the variants of the PD and the PI (including the previously discussed LN4 from the fifth network). The LN4 neurons from the networks: 2,6,7,12,17,18,19,20 are silent, because of the slow STF as discussed

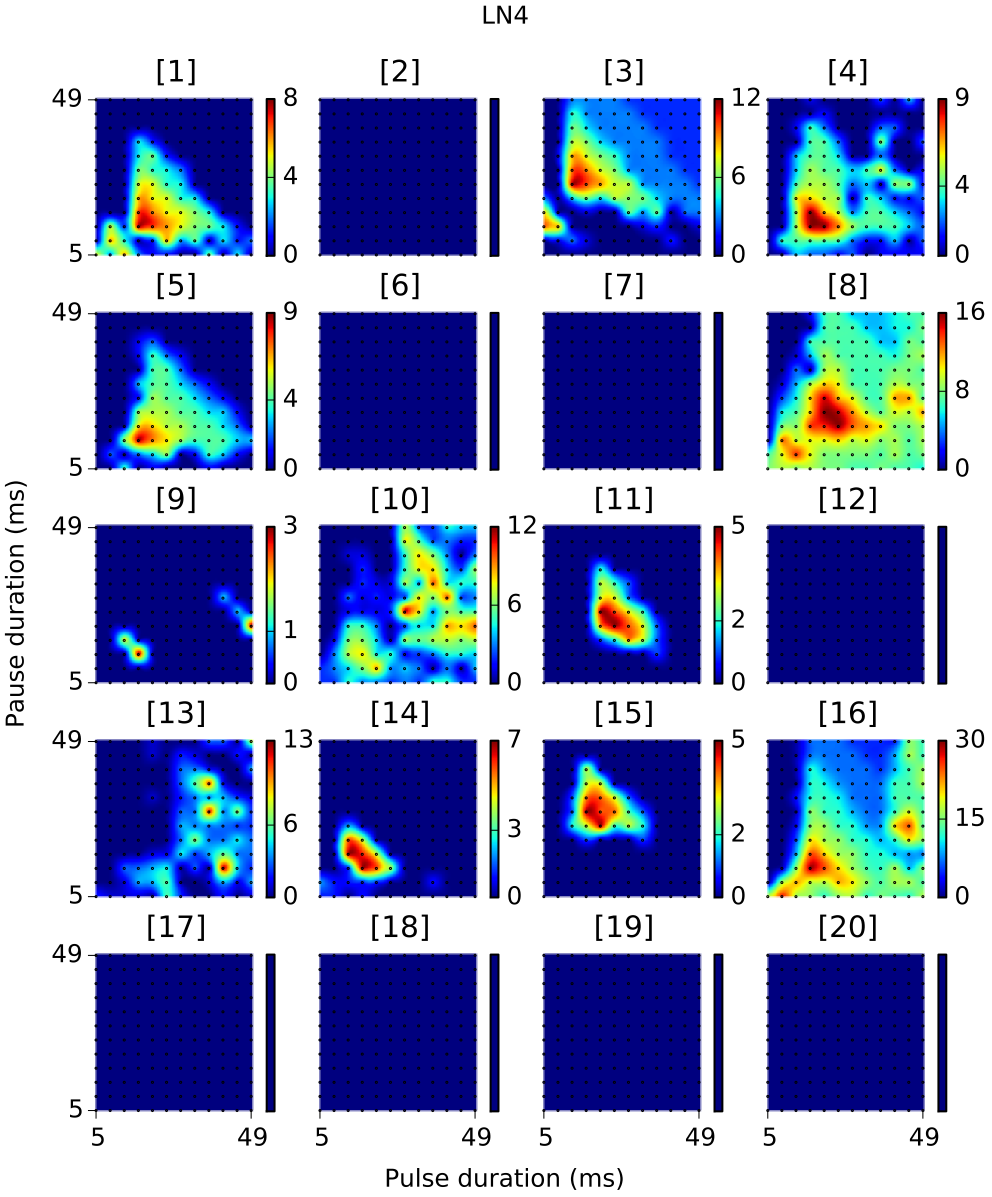


Figure 55: The heat-map plots showing the activity of the LN4 neurons of all the networks. The responses are the sum of the total number of action-potentials per chirp, in response to a chirp with varying PDs, PIs and duty-cycles. The variations in the neural responses due to the device mismatch are visible in the subplots. These deviations result in different shapes of the band-pass filter profiles.

earlier. The LN4 from the networks: 9,10,13 show non-specific response patterns. The LN4 from the networks: 1,3,4,5 show desired band-pass filter like response patterns. The LN4 from the networks: 11,14,15 show band-pass responses with highly narrow bandwidths. However, the LN4 from the networks: 8,16 show band-pass responses with wide bandwidths. Note that the maximum spike-counts are high in these two patterns, due to the strong STF (or low spike-threshold). Therefore, due to the device mismatch effects at this stage as well as from the previous stages of the networks, the spike counts and the shape of the band-pass filters vary across all the LN4 neurons. The deviations arising from the effects of the device mismatch increase at each stage of the network. In this case, the responses are measured at the fourth stage of the network (LN4), which differ significantly across the networks. Despite these deviations, different shapes of the filter can be obtained by simply selecting the response from other networks. These band-pass filter responses can be used to control the motor (or a descending neuron) to guide a robot based on acoustic patterns. The supplementary data displaying the variations across other neuron types are included in the DVD attached to this thesis (refer Chapter 7 for more details).

5.4.1 *Behavioral data of cricket phonotaxis*

To qualitatively compare the results obtained from the neuromorphic hardware with the behavioral data of cricket phonotaxis, we re-plotted the phonotaxis data published in [42] with permission, which is shown in fig. 56. The relative phonotaxis behavior scores plotted were averaged across n number of crickets *Gryllus bimaculatus*, for four different stimuli:

1. Fixing the PI; changing the PD and the duty-cycle for $n = 37$.
2. Fixing the PD; changing the PI and the duty-cycle for $n = 38$.
3. Fixing the duty-cycle to 50%; changing the PD and the PI for $n = 27$.
4. Fixing the PP; changing the PD, the PI and the duty-cycle for $n = 19$.

The experiment was performed by monitoring the movement of the female cricket using the walking compensator. The artificial calling songs were generated by multiplying the sinusoidal signal envelope with a sine wave of 4.5 kHz frequency. The songs were presented through the loudspeakers for a Chirp Duration (CD) that varied from 164 to 280 ms, followed by a Chirp Interval (CI) that ranged from 134 to 200 ms. The deviation of the cricket moving towards the sound source was measured as an indicator of phonotaxis. The relative phonotaxis scores were computed based on the methods described in Sec. 5.2. A subset of the phonotaxis behavioral data (averaged across the crickets), presented in [42] is shown in Fig. 56. The mean relative phonotaxis scores in response to the varying PPs with a constant duty-cycle of 50% are shown in

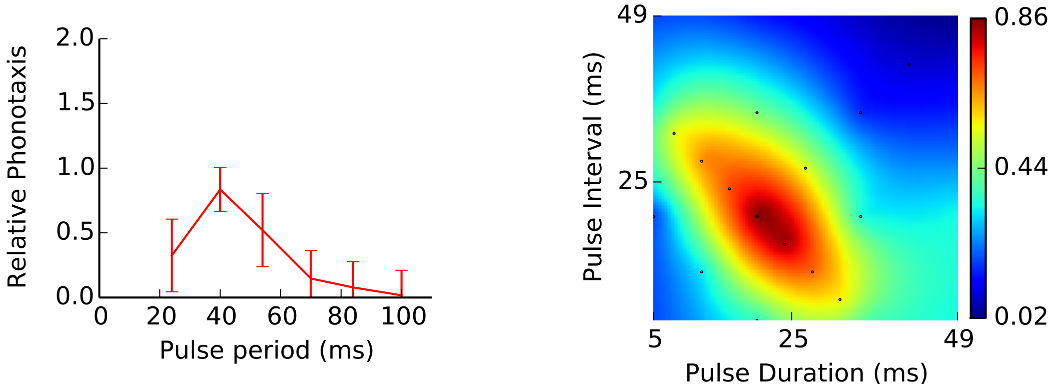


Figure 56: A subset of the cricket phonotaxis behavioral data presented in [42] is re-plotted with permission for qualitative comparison of LN4 responses presented in Fig. 53 and Fig. 54 with the phonotaxis behavior. The mean relative phonotaxis scores of crickets are shown in the left plot, in response to various PPs of the stimulus with the constant duty-cycle. The SDs are represented as the error-bars in the plot. The preference of crickets towards 40 ms PP is visible in the plot. The mean relative phonotaxis scores of the crickets in response to various duty-cycles of the PP are shown in the right plot. The mean responses show a band-pass filter like selectivity. It is visible that the crickets are attracted towards the stimulus that falls within the range of the 20 ms PD and the 20 ms PI in various duty-cycles. The heat-map drawn outside the measured data points (black dots) was interpolated.

the left figure. The error-bars denote the SDs across different animals. The band-pass selectivity towards a specific stimulus is visible in the plot. The peak of the filter exists at 40 ms, which is higher than 34 ms peak of the LN4 neuron modelled in the hardware (see Fig. 53). The heat-map plot of the behavioral data in response to different timescales of the PDs and the PIs is shown in the right figure. The data was obtained for only a few points, mainly on the principal diagonal, in the region of expected maximum responses. Very few data were measured away from the maximum response region. The missing data in between the points were interpolated. The maximum response is centered around the region of 40 ms PPs. This peak is higher compared to the peak activity of the LN4 neuron modelled in the hardware, which occurs in 30-34 ms PP region (see Fig. 54). Furthermore, the region of peak responses is more widespread in behavioral data, compared to the LN4 neuron responses from the hardware. This difference in bandwidth of the filters between the behavior and the neural data is consistent with the observations presented in Kostarakos and his colleagues' work ([58]). The authors showed that the local neuron B-LI4 showed a narrow band-pass region of selectivity, compared to the phonotaxis behavioral data (see Fig. 42). Therefore, the band-pass filter response of our calling song recognition network correlates with the phonotaxis behavior.

5.5 DISCUSSION

The responses of the LN4 neuron presented in Fig. 54 qualitatively matches the behavioral data presented in Fig. 56 and also the fine-tuned band-pass filter like responses of the B-LI4 neuron presented in [58] (shown in Fig. 42). Kostarakos and his colleagues already proposed in [58] that these neural responses correlate with the phonotaxis behavior of crickets. Our network utilizes a bio-realistic computation to recognize calling songs of crickets. The neurophysiological data from the biology [58] explain the functional role of the neurons. The missing data in between were interpolated in the biology, and the actual data points are unknown. However, our results suggest the network responses over the entire range of the stimulus, in steps of 4 ms. It is reported in [58], that no significant change occurs in the neural responses within this duration. Therefore, our model can be useful in predicting the neural activity for the missing data points in the biological experiments.

In [101], Schildberger proposed that the band-pass response profile might be the result of the interplay between the neurons with low-pass and high-pass responses. A more recent work in [58], Kostarakos, and his colleagues identified several local neurons with band-pass filtering characteristics and suggested in [103], that coincidence detection among them might play a role in selecting the attractive stimulus. Nevertheless, we do not rule out the existence of a neuron with high-pass filter characteristics in cricket brain. For instance, the LN3 neuron responded with one spike per PD similar to a pulse-onset detector and the number of the PDs within one chirp became small for long PDs. When the chirp varying in the PP with a constant duty-cycle is presented directly to the standalone LN3 neuron without sending it through the network, one can expect a high-pass filter response from this neuron. The high-pass filter response can also be obtained using the STD synapse as modeled in [97]. However, in both of our implementations, the stimulus plays a significant role in shaping this high-pass filter response. Therefore, temporal screening is the critical element of our computational model. Temporal filtering is one of the well-studied properties of the Short-Term Plasticity (STP) [1]. We modeled the temporal filtering properties of two of the neurons using the STF in their synapses. Along with the SFA, the STF shaped the band-pass filter responses (given the pulse-based chirps) of the neurons. The time-constant of the STF of the LN2 synapse is tuned to recover completely during 17 ms PI. Note that the structure of the chirp plays a crucial rule in tuning these filters, as the first temporal filter is set within the incoming stimulus, and the whole network operated within this window. It is especially true for the LN3 neuron which is tuned to emit one spike per PD, beginning from 17 ms PI. The LN4 neuron however operates on the PP and the

STF of this neuron is tuned to recover after 34 ms PP. Therefore, the temporal selection reaches the maximum at 34 ms PP as evident from Fig. 53. Therefore, the time-constant of the STP plays a crucial role in the network. The designed STP currents offer desired flexibility in tuning their time-constants (see Sec. 4.4 for more details).

In [101], Schildberger mentioned that the onset latency of the AN1 neurons decreases with the increase of the stimulus intensity. In our experiments, the initial onset latency occurred at the AN1 neuron due to its low synaptic weight. The latency increased at each stage of the network due to the STF synapses and a slow integration of the EPSCs by the neurons. The delay can be useful especially during the phonotaxis implementation of both the sides of the auditory system, by tuning one side faster than the other. However, in [86], the authors claim that the useful information for sound localization was the difference in the tuning strengths of the inter-aural responses than the latency itself. Nevertheless, studying the functional role of the delay will be a future direction of this research.

Evidence of inhibition was found in the local neuron B-LI4 in [58]. Later in [103], the Post-Inhibitory Rebound (PIR) was found to be the cause of the repression. The PIR played a role in band-pass selectivity of the local neuron. A calling song recognition network with PIR was proposed in [103], based on the latencies observed at different local neurons of the brain. This network was based on the coincidence detection between the delayed spikes and the non-delayed spikes from another neuron, resulting in a band-pass response. Such a system cannot be implemented in our neuromorphic hardware because it is not possible to stimulate the synapse with a non-digital signal in the current chip. However, the idea of a non-spiking neuron is promising to be included in the design of the next-generation neuromorphic chips.

5.6 CONCLUSION

The STP has also been reported in other insects such as mushroom bodies of drosophila [106] and honey bees [75]. The STP is also known to play a role in sound localization in avian auditory brain-stem [28]. Since no long-term learning has been observed during cricket phonotaxis, and the recognition occurs at the milliseconds scale, it is worthwhile to investigate the role of the STP in cricket phonotaxis. We demonstrated the significance of the STP in a small network to recognize the calling songs of crickets during phonotaxis. The network is able to select the attractive features of the given stimulus. Since the neurons are modelled based on the biological evidence, the network sugests an educated guess about the connectivity scheme of the cricket phonotaxis network. Our model provides a silicon platform to test the neural responses for complex stimulus conditions, for e.g., various chirp patterns.

The filtering responses of the network are modelled using the STF with different time-constants. These time-constant values are crucial for the implementation of this network in a neuromorphic hardware. The implementation of this network is possible, thanks to the real-time operation of our neuromorphic hardware, that operates in the same scale of biological time-constants. The calling song recognition network modelled in this research is compact, with only four neurons in a simple feed-forward connectivity. This feature ensures the network easier to port to any other neuromorphic chip. We built a dedicated neuromorphic chip to implement the calling song recognition network. The STP circuits proposed in Sec. 4.4 are used to model the STP synapses of the network. The network that is built using the multi-chip hardware can be implemented using a single chip. The chip with the calling song recognition network can be used for acoustic based robotic tasks. Further details of this application specific neuromorphic chip can be found in Chapter 7.

In this thesis, we have presented the response of the networks only to the regular frequencies of the input pulses. However, in a real-world scenario, the input spikes can be randomly distributed, for e.g., Poisson distribution. We simulated our networks using the spikes from the Poisson and the Gamma distribution. The analysis of these response of the networks are beyond the scope of this research. Therefore, we included this data as the supplementary material in the DVD attached with this thesis. This data can be used to characterize the network with the real-time sensor such as the event-based silicon cochlea [65].

6

CONCLUSION

A dedicated STD circuit is available in the neuromorphic chip. However, an explicit control over the recovery-rate of the depression strength was missing in the circuit. We designed a set of STD circuits that offers an independent control over the recovery-rate of the depression voltage. This circuit can be used to obtain specific temporal dynamics of the STP, such as strong depression followed by a fast recovery of the synaptic strength, which was not possible with the existing STD circuit. The STD circuit can also be used to investigate the role of STD in a calling song recognition network of crickets. For example, in the synapse of the LN3 neuron, that implements a high-pass filter like a response, when stimulated out of the network.

No dedicated Short-Term Facilitation (STF) circuit exists in the current hardware. The DPI synapse was used to implement the STF, sharing the time-constant between the synapse and the STF. This constraint limits the synapse to operate with a time-constant different from the STF and vice versa. We designed a dedicated circuit to implement the STF. The output voltage of this circuit did not always reach the steady-state values. Therefore, we redesigned the STF circuit by adding a negative feedback loop. The new STF circuit offers complete control over the recovery-rate of the facilitation strength as well as the output voltage of the circuit reaches the steady-state values. The STF circuit can be used to model long latencies in a calling song recognition network of crickets. It allows implementing both STD and STF at the same synapse, with independent time-constants. We combined the STD and the STF circuits to design the band-pass filter like characteristics of the STP to input frequencies. Please refer Appendix for further details of this circuit. All these STP circuits are fabricated in CMOS, and their responses are tested and characterized.

Alongside the STP circuit design, we demonstrated the computational significance of the STP in a small feed-forward network to recognize calling songs of crickets. We used the STF to model the band-pass filter characteristics of the neurons of the network. The spiking neural network was emulated in the available neuromorphic hardware. The network shows a band-pass filter like selectivity to the calling song stimulus, analogous to the behavioral and neurophysiological evidence. The network responsible for recognizing calling songs of crickets during phonotaxis is unclear in the literature. However, our network model suggests the connectivity scheme of the auditory neuron circuitry in cricket brain. We designed a neuromorphic chip dedicated to implementing this network using the newly developed STP circuits. The details of this chip can be found in the Appendix. This chip is compact and can be used in the acoustic-based robotic tasks.

Therefore, we addressed the two primary goals of neuromorphic engineering through this research:

I. To understand the underlying computational principles of neurobiology using the hardware.

II. To build the silicon circuits inspired by the neuroscience.

6.1 ADVANTAGES OF NEUROMORPHIC APPROACH

Timing is a crucial factor in auditory systems, as the information is precisely encoded in the timing of the spikes. During the implementation of our calling song recognition network, we exploit the real-time operation of our neuromorphic hardware, to faithfully model the latency and the spike-times of the neurons for time-specific features of the chirp. The biologically plausible time-constants in the range of 50-100 milliseconds are crucial to implementing the short-term synaptic dynamics in our network. Thanks to the real-time operation of our neuromorphic chips that the time-constants in this range are achievable. The neuromorphic hardware has the inherent noise due to the device mismatch effects resulting from the fabrication process. This device mismatch is useful in introducing inhomogeneities among the computational elements (the synapses and the neurons) in the chip. These variations can be exploited to model the biologically realistic neural computations. We designed the STP circuits by following the design strategies to minimize the mismatch effects. However, we also used these deviations in the responses to implement various shapes of the band-pass filter like responses of the neurons.

We used a neuromorphic hardware to uncover the puzzles of small-scale neural circuits. Given its small size, the network can be implemented using any digital platform. However, we aim to model the computations in silicon as close as possible to the biology. This goal is relevant, using the neuromorphic chip. Unlike standard processors, the processing speed does not scale with the size of the network. The power consumption is significantly lower than conventional processors. The event-based neuromorphic sensors can be readily integrated to these chips, for real-time interaction with the environment which is useful for online-learning in robots. The neuromorphic chips offer a massively parallel framework, whose units are non-identical and perform real-time computations. These brain-like computing machines can be used to bridge the gap between the machine learning approaches such as deep convolutional neural networks to the bio-inspired approaches such as spiking networks.

6.2 FUTURE WORKS

In this research, we investigated the role of STP in a small neural network. However, the STP is known to influence the neural dynamics in large networks. Examining the impact of the STP in these large systems is the possible future direction of this research. We also designed the STP circuits, which can be tested for more computational properties of the STP. Despite the fact, we built these circuits in the context of the audition; these circuits can be used to implement temporal filters in other modules as well, e.g., vision. For example, our STP circuits can be used to build a spiking model of the Elementary Motion Detection (EMD). EMD is a model that describes the simplified computations to perceive movement from the activity of photoreceptors. The simplest EMD model consists of two photoreceptors, a delay element, and a multiplier. The time delay ensures the two arriving signals at the photoreceptors are correlated in time. The multiplier amplifies the highly correlated activity. In the spiking model of the EMD, events from the pixels of the Dynamic Vision Sensor (DVS) (also known as silicon retina) can be taken into account for the photoreceptors activity. The output of the first pixel can be used to stimulate the STP circuit and the second pixel to the DPI synapse. In this model, the output of the STP circuit can be connected to either the gate of the weight transistor or the threshold transistor of the DPI synapse. The output voltage of the STP circuits can be used as a scaling factor to modify the EPSC of the DPI synapse. This modification occurs on a short-time scale, during which an event or input pulse travels from one synapse to the other. The underlying neuron integrates the incoming EPSC and elicits output spikes depending on the size of the EPSC amplitude. Therefore, the speed of the arriving event is encoded regarding the firing rate of the neuron. This spiking EMD model can be used to navigate insect-inspired robots. Thus, the designed STP circuits can be used as an on-chip velocity encoder in an event-based sensor.

From the implementation of the calling song network, we realize that the insect inspired neural systems require a distinctive architecture of the chip design allowing more flexibility in tuning the synapses individually, than the general purpose hardware. Although available hardware provides an efficient platform to implement small neural networks, more than one chip is required to model distinct synapses. For example, in our implementation, we used two chips to implement four neurons because the parameters are shared among the synapse array of each chip. Another limitation is the portability of the multi-chip setup, to be used in a mobile robotic platform. Therefore, the hardware needs a unique design strategy which allows a large number of independently tunable synapses and neurons, to implement insect-inspired systems. The requirement for a large number of input/output pins might be a limiting factor. However, this constraint can be tackled by implementing on-chip scaling of the parameters such as synaptic weights and switching between these weights. This design technique would offer more flexibility to use the small-scale neuromorphic chips. We designed one such chip to export the calling song recognition network. The architecture of the chip design follows a similar strategy as we discussed above. Each neuron of the chip receives the currents from eight independently tunable synapses. This chip is a prototype with only eight neurons in total (more details of this chip can be found in the Appendix). The chip laid a foundation to build reconfigurable architectures inspired by insect neural structures. This architecture is scalable and can be used to design large-scale neural systems. As we know, reconfigurability is the vital feature of the next generation architecture of the neuromorphic systems.

6.3 OUTLOOK OF NEUROMORPHIC ENGINEERING RESEARCH

Moore's law is reaching its limits, meaning further scaling of the silicon is impossible due to the quantum effects. It opens the door for new computing paradigms such as quantum computing and neuromorphic computing. The quantum computers are aiming to reach high speeds, whereas the neuromorphic processors aim to replicate brain-like processing. Since the foundation in the late 80's by Carver Mead, the neuromorphic computing is an actively growing research area. Many effective approaches have been found by the neuromorphic community since then to build a brain-like-computing machine. The neuromorphic computing field is continually evolving with the advent of new devices, fabrication technologies, and brain imaging techniques. For example, memristors, which have been intensively studied for their usage in neuromorphic computing. The memristors are only a few nanometers in size and show promising capabilities such as long-term storage of synaptic weights. These properties ensure the memristors to be ideal candidates to implement the synapses for large-scale neuromorphic hardware [55]. Integration of these memristive devices with the existing CMOS based neuromorphic circuits will provide a full flexibility in tuning to achieve complex temporal dynamics of the synapse along with the long-term storage of weights. Attempts have been made to culture the neuron cells on top of the memristor arrays, to build a living brain-machine interface. The neuromorphic chips have been used to construct brain-machine interfaces [26]. These chips can locally process the neural recordings and can benefit the society by being used for neuroprosthetics. With the ongoing progress in developing neuromorphic sensors and motor control systems for neuroprosthetic applications [67, 85], the neuromorphic research is gaining attention on bio-medical interfaces. The success of this inter-disciplinary research highly depends on the integration of the researchers from all the communities especially biology, neuroscience, machine-learning, physics, engineering, and robotics.

7 APPENDIX

7.1 FABRICATED CRICKET CALLING SONG RECOGNITION NETWORK

Considering the limitations of the STP circuits that are already existing in the hardware (see Sec. 4.2 for more details), we designed another test chip using the STP circuits we discussed earlier in Sec. 4.3 and Sec. 4.4. We call it a test chip-2 and the schematic of this test chip with the calling song recognition network of cricket is shown in Fig. 57. Block diagram of the test chip-2 is shown in the top, and the monitoring scheme to observe the input and output spikes is shown in the bottom. The calling song recognition block consists of an array of 8 DPI neurons (see right corner). Each neuron receives currents from the excitatory and the inhibitory synapses, STD synapse without feedback, STF synapse with feedback, and a band-pass filter synapse consisting of another set of the STD circuit without feedback and the STF circuit with feedback. The Calcium-based learning synapses ($Ca1$ and $Ca2$) are not used in the calling recognition network, therefore will not be discussed here.

The biases are shared across the rows and are distinct across each block in the column. Two networks of calling song recognition network of crickets can be implemented using this chip, by selecting two sets of four neurons. The first neuron receives the input events through the excitatory synapse and implements the SFA using the inhibitory synapse. The second neuron receives input through the STF synapse with feedback. The third neuron gets the spikes through the synapse from the EMD block. In other words, we borrowed a synapse from the EMD block. The details of this EMD block are beyond the scope of this research. Therefore we will not discuss its features here. The forth neuron receives input through the other STF synapse circuit with feedback, borrowed from the band-pass filter synapse block (explained in the next section). The monitoring scheme is shown in the bottom of Fig. 57. The output voltages of the desired synapses and neurons can be monitored by selecting the appropriate row and column addresses. This way, we could implement two calling song recognition networks in parallel. Neurons with four types of distinct synapses are available in the same chip, and the events can be routed using the inbuilt mapper.

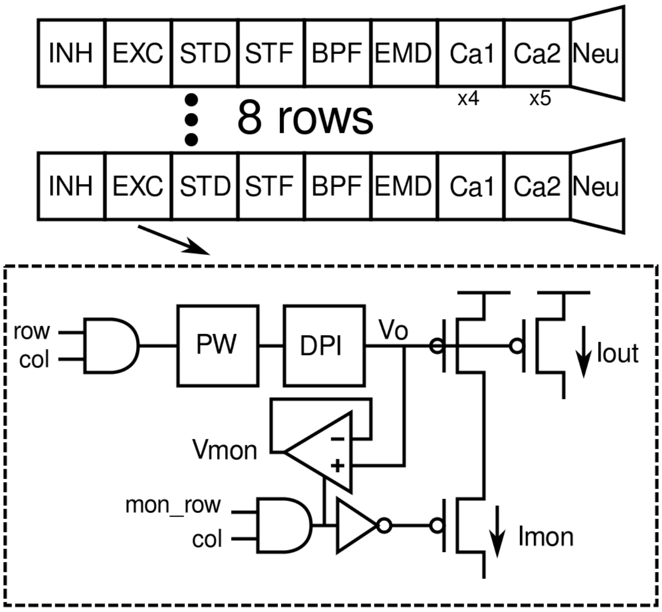


Figure 57: The schematics showing the calling song recognition block of the test chip-2. Top: Block diagram of the test chip showing the array of synapse and neuron blocks along with the newly designed STP circuits. Bottom: Schematic of the monitoring scheme to monitor the output voltages of the desired synapses and neurons by providing appropriate row and column addresses. The architecture is designed in a fashion to allow the implementation of calling song recognition network in a single chip. The top array consists of 8 rows of 8 different synapses, the currents of which are injected into a DPI neuron located at the end of each row. The synaptic arrays include inhibitory (INH), excitatory (EXC), STD without feedback, STF with feedback, Band-pass filter (BPF) and Elementary Motion Detector (EMD). Calcium-based learning synapses (CA1 and CA2) are independent of our network implementation.

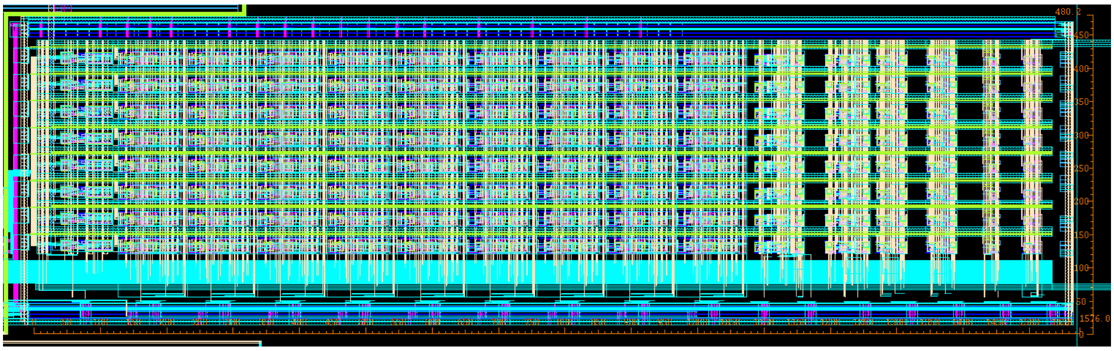


Figure 58: The layout of the testchip2 with the calling song recognition network. The layout consists of arrays of 8 synapses, and eight neurons fabricated using the standard CMOS AMS 180 nm technology. The total design area occupied is 1576.0*480.2 μm^2, including the Calcium-based learning blocks. The actual design area occupied by the calling song recognition block is 1080.0*480.2 μm^2. Further design details are provided in table 2.

The network of cricket calling song recognition discussed in chapter 5 is designed and fabricated in a standard CMOS AMS 180 nm technology. The layout of the calling song recognition network is shown in Fig. 58. The design is compact, and it occupies the silicon area of 1576.0*480.2 μm^2. Omitting the Calcium-based learning blocks, our network design alone occupies 1080.0*480.2 μm^2. This plan can be exported to build large-scale arrays.

We designed the architecture of the calling song recognition block based on the model (refer Sec. 5.3.2) we implemented using the multi-chip setup (see Sec. 3.8). Since we used two chips to model four distinct neurons, we aimed to solve this issue with our new architecture which allows us to model four distinct neurons in the same chip. However, the response variability can be characterized for up to two networks with this test chip-2. Nevertheless, the application specific design allows this test chip to be used as the prototype to study the sound based robotic tasks.

Block	Type	Length	Width	Value
Inhibitory Synapse	M5-M6 Cap	10µ	10µ	200.368fF
	Transistor	0.5µ	1µ	-
	Silicon occ.	31.33µ	18.33µ	-
Excitatory Synapse	M5-M6 Cap	10µ	10µ	200.368fF
	Transistor	0.5µ	1µ	-
	Silicon occ.	28.915µ	23.3µ	-
STD simple + Synapse	M5-M6 Cap	7µ	7µ	97.2fF
	Transistor	0.36µ	1µ	-
	Silicon occ.	48.62µ	18.33µ	-
STF f.back + Synapse	M5-M6 Cap	7µ	7µ	97.2fF
	Transistor	0.36µ	1µ	-
	Silicon occ.	47.97µ	18.33µ	-
BPF + Synapse	M1-M4 Cap	7µ	7µ	97.2fF
	Transistor	0.36µ	1µ	-
	Silicon occ.	68.21µ	18.33µ	-
Neuron	Silicon occ.	91.13µ	40.15µ	-

Table 2: Dimensions of the circuit blocks designed in testchip-2.

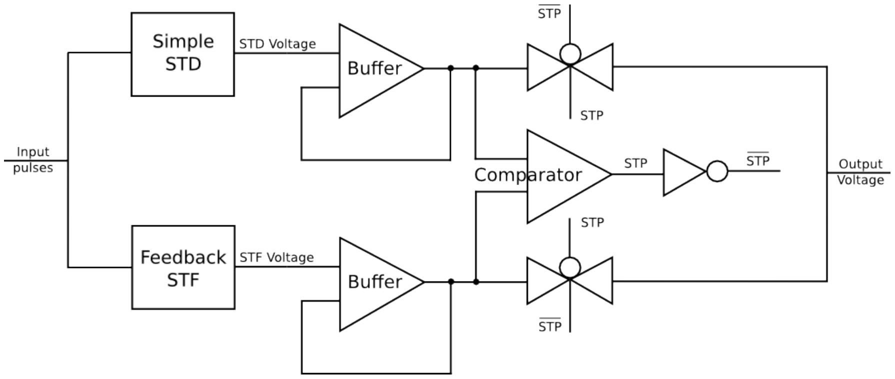


Figure 59: The schematics of the band-pass filter circuit designed in the test chip-2. The circuit consists of the STD (without feedback) and the STF circuits (with the feedback), two opamp buffers, a comparator and a transmission-gate logic. The output voltages of the STD and the STF circuits are compared, and the minimum of these two voltages are selected through the transmission-gate logic.

7.2 STP BASED BAND-PASS FILTER

Evidence in biology [28] suggest that both the STD and the STF synapses can target the same neuron, with different time-constants. This combination of the STP would result in a band-pass filter like selectivity to the incoming pre-synaptic spikes as described in [53]. It is feasible to implement the band-pass characteristics using the STP model proposed by [71]. However, these band-pass filter characteristics cannot be achieved using the current neuromorphic hardware. Therefore, we designed a circuit that implements this phenomenon.

STP based band-pass filter synapse is designed in both the test chip-1 and the test chip-2 (shown as the BPF block in Fig. 57). Considering the similarities in the design of this circuit in both the test chips, we restrict our discussion in this section with the design of the test chip-2. The schematic of the band-pass filter circuit is shown in Fig. 59. The circuit consists of the STD circuit without the feedback and the STF circuit with the feedback. The output voltages of these two STP circuits are connected to the transmission-gate logic. An opamp comparator is used to compare the output voltages of the two STP circuits. The operational amplifier (or opamp) is the voltage amplifier, which is also the most extensively used device in electronics. Various configurations of opamp exits and the comparators are one among them. The comparator compares the two input voltages and outputs a high signal if one of them is larger than the other. The transmission-gates select the minimum of the two compared output voltages. A transmission-gate consists of an nMOS and a pMOS transistor connected in parallel. The voltage applied to the gate of the nMOS is the inverted version of the voltage ap-

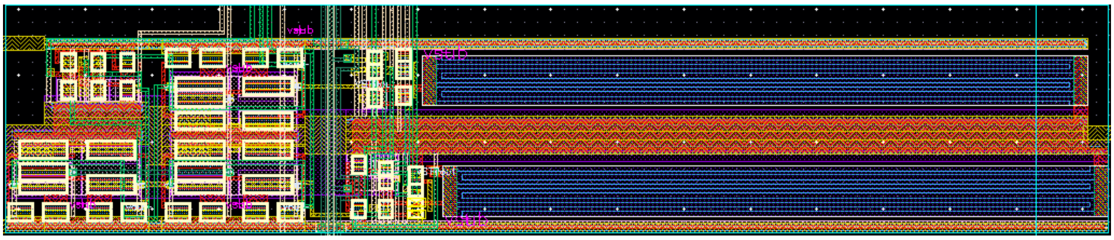


Figure 60: The layout of the band-pass filter circuit designed using the standard CMOS AMS 180 nm technology. The STD circuit and the STF circuits are located in the top-right and the bottom-right regions. Two opamp buffers are situated in the middle (top and bottom). The comparator is located in the bottom-right region, and the transmission-gate logic is situated in the top-left region. As expected, the capacitors in the right corner (top and bottom) occupy most of the silicon area in the design. The capacitors are built using four layers of the Metal-Insulator-Metal(MIM). The layout occupies the design area of 82.91*17.06 μm^2.

plied to the pMOS transistor. The circuit acts a switch with a control voltage supplied through the gates. The operational amplifier, whose output gain is configured to one is called a unity-gain follower and can be used as a buffer. The output voltages of the STD and the STF circuits are sent through the opamp buffer, to decouple the voltages from the transmission-gates.

As already mentioned in Chapter. 4, the STD circuit shows a low-pass filter like response and the STF circuit displays a high-pass filter like profile to the input frequencies. The designed circuit combines these two characteristics of the STD and the STF circuits, to implement the band-pass filter characteristics to input frequencies. The circuits are designed and fabricated using the standard CMOS AMS 180 nm technology. The circuit occupies the silicon area of 82.91*17.06 μm^2.

The layout design of this circuit is shown in fig. 60. The STD circuit without a feedback occupies the top-right region. The STF circuit with the feedback occupies the bottom-right region. The capacitors of the STD (top-left) and the STF circuits (bottom-left) occupy the largest area of the design. The capacitors are designed using four layers of the Metal-Insulator-Metal(MIM). Three opamps used to implement two buffers (top and bottom: middle) and a comparator (bottom-right). The transmission-gate logic is designed in the top-left area of the layout.

The steady-state responses of the fabricated band-pass filter circuit are tested and characterized by the input frequencies. Input pulses of 20 μs duration are provided from 10 Hz up to 150 Hz (in steps of 10) to the circuit, and the output voltages are recorded through the oscilloscope. The mean and the SDs of the steady-state values of the output

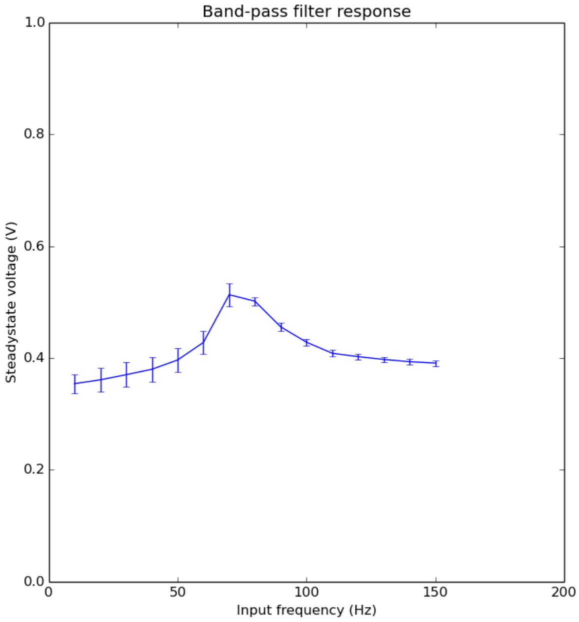


Figure 61: The steady-state output voltage responses of the band-pass filter circuit to the input frequencies are plotted. The input frequencies vary from 10 Hz to 150 Hz, and the corresponding steady-state output voltages are recorded. Each point in the curve represents the mean steady-state output voltages of the circuit, and the error bars represent the SDs. The overall mean steady-state response shows a band-pass filter like profile. The low peak-to-peak amplitude of the steady-state output voltage for high frequencies (due to short ISIs) results in small SD values.

voltages are computed for each input frequency as shown in fig. 61. Each point in the curve denotes the mean of the steady-state output voltages. The error-bars denote the SDs. The resulting profile of the steady-state responses displays a band-pass filter. The shape of the band-pass filter can be modified using different bias settings of the STD and the STF circuits. The peak-to-peak amplitudes of the steady-state output voltage decreases for high-frequency inputs due to the short ISIs. Therefore, the values of the SDs are small for high-frequencies.

The STD responses can be obtained by shutting down the STF (by increasing the V_{low} and V_{up} of the STF higher than corresponding values of the STD circuit) and vice-versa. The STF response alone is used to implement the STF synapse of the calling song recognition network in test chip-2. Therefore, three types of the filter responses can be achieved at a single synapse. The circuit offers a multi-purpose computational building block that can be integrated into the design of large-scale neuromorphic hardware.

Parameter	STD-no feedback (Volts)	STF-no feedback (Volts)	STD-feedback (Volts)	STF-feedback (Volts)
Vwei	0.85	2.225	0.85	2.225
Vtau	0.2	0.52	0.2	0.52
Vup	0.8	2.8	0.8	2.8
Vlow	0.4	0.4	0.4	0.4
Vlim	-	-	0.6	0.3
Frequency	200	200	200	200
Pulsewidth	1µ	1µ	1µ	1µ
Risetime	1n	1n	1n	1n
Falltime	1n	1n	1n	1n

Table 3: Parameters used to simulate the STP circuits to obtain the temporal filters, presented in Fig. 35 of Sec. 4.5.

7.3 SUPPLEMENTARY MATERIAL

The parameters used in the simulations and emulation results obtained in this research are presented in this section as tables. The parameters used to simulate the STP circuits to obtain the temporal filters presented in Fig. 35 of Sec. 4.5 are shown in table 3. The parameters used to test/emulate the fabricated STP circuits to obtain the temporal filters presented in Fig. 40 of Sec. 4.6 are shown in table 4. The parameter sets of the synapses and neurons used to emulate the cricket calling song recognition model presented in Fig. 53 and Fig. 54 of Sec. 5.4 are shown in table 5.

The additional data obtained during the testing of the fabricated STP circuits and also the extra data from the implementation of the calling song recognition network are written to the DVD attached to this thesis. The supplementary data includes the steady-state responses of the output voltages of the STD and the STF circuits, with and without the feedback control. We also added the additional raster plots obtained for the PPs from 10 to 98 ms, with a constant duty-cycle together with the response of all the networks to the PP variants along with the duty-cycle variations (the PD and the PI varied from 5 to 49 ms). We included the response of all the networks to the non-regular frequency of the input pulses such as the 'Poisson' and the 'Gamma' distribution of the input spikes. These responses can be used to understand the operation

Parameter	**STD-no feedback (Volts)**	**STF-no feedback (Volts)**	**STD-feedback (Volts)**	**STF-feedback (Volts)**
Vwei	0.5	0.65	0.5	0.65
Vtau	0.85	0.3	0.81	0.33
Vup	1	1	1	1
Vlow	0.3	0.3	0.3	0.3
Vlim	-	-	0.5	0.5
Frequency	100	100	100	100
Pulsewidth	20μ	20μ	20μ	20μ
Vbuf	1.2	1.2	1.2	1.2

Table 4: Parameters used to emulate the fabricated STP circuits to obtain the temporal filters, presented in Fig. 40 of Sec. 4.6.

Synapse	**Vwei (Volts)**	**Vtau (Volts)**	**Vthr (Volts)**	**Pls.width (Volts)**
AN1 Exc.	0.53	2.95	2.86	-
AN1 Inh.	2.48	0.11	0.8	0.08
LN2 Exc.	0.62	2.86	2.86	-
LN3 Exc.	0.7	2.8	2.85	-
LN4 Exc.	0.55	2.99	2.85	-
Neuron	**Vadap**	**Vleak**	**Vrefr**	
2D-0	0.13	0.17	0.25	-
2D-1	0.2	0.17	0.25	-

Table 5: Parameters used to emulate the cricket calling song recognition model presented in Fig. 53 and Fig. 54 of Sec. 5.4.

of the network (chip) when integrated with a real-time sensor such as event-based silicon cochlea [65]. The scripts used to simulate the neuromorphic hardware and analyze the data from the hardware are also added to the DVD enclosed with this thesis. Please refer to the 'readme.txt' file from the DVD for further details about the scripts and the data attached.

BIBLIOGRAPHY

[1] L.F. Abbott and W.G. Regehr. “Synaptic computation.” In: *Nature* 431.7010 (2004), pp. 796–803.

[2] L.F. Abbott, K. Sen, J. Varela, and S. Nelson. “Synaptic depression and cortical gain control.” In: *Science* 275.5297 (1997), pp. 220–223.

[3] B.E. Alger and T.J. Teyler. “Long-term and short-term plasticity in the CA1, CA3, and dentate regions of the rat hippocampal slice.” In: *Brain research* 110.3 (1976), pp. 463–480.

[4] F. Alibart, S. Pleutin, D. Guérin, C. Novembre, S. Lenfant, K. Lmimouni, C. Gamrat, and D. Vuillaume. “An organic nanoparticle transistor behaving as a biological spiking synapse.” In: *Advanced Functional Materials* 20.2 (2010), pp. 330–337.

[5] J.V. Arthur and K. Boahen. “Recurrently connected silicon neurons with active dendrites for one-shot learning.” In: *IEEE International Joint Conference on Neural Networks*. Vol. 3. 2004, pp. 1699–1704.

[6] A.I. Bain and D.M. Quastel. “Multiplicative and additive Ca (2+)-dependent components of facilitation at mouse endplates.” In: *The Journal of physiology* 455.1 (1992), pp. 383–405.

[7] C. Bartolozzi. “Design concepts for a novel asynchronous space-variant vision sensor.” In: *21th Italian Workshop on Neural Networks, WIRN 2011*. 2011.

[8] C. Bartolozzi and G. Indiveri. “Silicon synaptic homeostasis.” In: *Brain Inspired Cognitive Systems, BICS 2006*. 2006, pp. 1–4.

[9] C. Bartolozzi and G. Indiveri. “Synaptic dynamics in analog VLSI.” In: *Neural Computation* 19.10 (2007), pp. 2581–2603. DOI: 10.1162/neco.2007.19.10.2581.

[10] A. Basu, S. Brink, C. Schlottmann, S. Ramakrishnan, C. Petre, S. Koziol, F. Baskaya, C. M. Twigg, and P. Hasler. “A floating-gate-based field-programmable analog array.” In: *IEEE Journal of Solid-State Circuits* 45.9 (2010), pp. 1781–1794.

[11] J. Benda and R.M. Hennig. "Spike-frequency adaptation generates intensity invariance in a primary auditory interneuron." In: *Journal of computational neuroscience* 24.2 (Apr. 2008), pp. 113–36. ISSN: 0929-5313. DOI: 10.1007/s10827-007-0044-8. URL: http://www.ncbi.nlm.nih.gov/pubmed/17534706.

[12] B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J.-M. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, and K. Boahen. "Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations." In: *Proceedings of the IEEE* 102.5 (2014), pp. 699–716. ISSN: 0018-9219. DOI: 10.1109/JPROC.2014.2313565.

[13] T.V.P. Bliss and G.L. Collingridge. "A synaptic model of memory: long-term potentiation in the hippocampus." In: *Nature* 361.6407 (1993), pp. 31–39.

[14] K.A. Boahen. "Retinomorphic Vision Systems: Reverse Engineering the Vertebrate Retina." Ph.D. thesis. Pasadena, CA: California Institute of Technology, 1997.

[15] M. Boegerhausen, P. Suter, and S. C. Liu. "Modeling short-term synaptic depression in silicon." In: *Neural Computation* 15.2 (2003), pp. 331–348.

[16] S. Brink, S. Nease, P. Hasler, S. Ramakrishnan, R. Wunderlich, A. Basu, and B. Degnan. "A learning-enabled neuron array IC based upon transistor channel models of biological phenomena." In: *IEEE transactions on biomedical circuits and systems* 7.1 (2013), pp. 71–81.

[17] S.L. Bush and J. Schul. "Pulse-rate recognition in an insect: evidence of a role for oscillatory neurons." In: *Journal of Comparative Physiology A* 192.2 (2006), pp. 113–121.

[18] J.D. Castillo and B. Katz. "Statistical factors involved in neuromuscular facilitation and depression." In: *The Journal of Physiology* 124.3 (1954), p. 574.

[19] F.S. Chance, S.B. Nelson, and L.F. Abbott. "Synaptic Depression and the Temporal Response Characteristics of V1 Cells." In: *The Journal of Neuroscience* 18.12 (1998), pp. 4785–99.

[20] T. Chang, S. Jo, and W. Lu. "Short-term memory to long-term memory transition in a nanoscale memristor." In: *ACS nano* 5.9 (2011), pp. 7669–7676.

[21] A. Chiang, C. Lin, C. Chuang, H. Chang, C. Hsieh, C. Yeh, C. Shih, J. Wu, G. Wang, Y. Chen, C. Wu, G. Chen, Y. Ching, P. Lee, C. Lin, H. Lin, C. Wu, H. Hsu, Y. Huang, J. Chen, H. Chiang, C. Lu, R. Ni, C. Yeh, and J. Hwang. "Three-dimensional reconstruction of brain-wide wiring networks in Drosophila at single-cell resolution." In: *Current Biology* 21.1 (2011), pp. 1–11.

[22] E. Chicca, D. Badoni, V. Dante, M. D'Andreagiovanni, G. Salina, L. Carota, S. Fusi, and P. Del Giudice. "A VLSI recurrent network of integrate–and–fire neurons connected by plastic synapses with long–term memory." In: *IEEE Transactions on Neural Networks* 14.5 (2003), pp. 1297–1307. DOI: 10.1109/TNN.2003.816367.

[23] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri. "Neuromorphic electronic circuits for building autonomous cognitive systems." In: *Proceedings of the IEEE* 102.9 (2014), pp. 1367–1388. ISSN: 0018-9219. DOI: 10.1109/JPROC.2014.2313954.

[24] L. Chua. "Memristor-The missing circuit element." In: *IEEE Transactions on Circuit Theory* 18.5 (1971), pp. 507–519. ISSN: 0018-9324. DOI: 10.1109/TCT.1971.1083337.

[25] J. Clemens, C.C Girardin, P. Coen, X.-. Guan, B.J. Dickson, and M. Murthy. "Connecting neural codes with behavior in the auditory system of Drosophila." In: *Neuron* 87.6 (2015), pp. 1332–1343.

[26] F. Corradi and G. Indiveri. "A neuromorphic event-based neural recording system for smart brain-machine-interfaces." In: *IEEE transactions on biomedical circuits and systems* 9.5 (2015), pp. 699–709.

[27] A. Destexhe, Z.F. Mainen, and T.J. Sejnowski. "Methods in Neuronal Modelling, from ions to networks." In: The MIT Press, Cambridge, Massachussets, 1998. Chap. Kinetic Models of Synaptic Transmission, pp. 1–25.

[28] L.A. Grande D.L. Cook P.C. Schwindt and W.J. Spain. "Bursting neurons and ultrasound avoidance in crickets." In: *Nature* 421 (2003), pp. 66–70.

[29] T. Dowrick, S. Hall, and L.J. McDaid. "Silicon-based dynamic synapse with depressing response." In: *IEEE transactions on neural networks and learning systems* 23.10 (2012), pp. 1513–1525.

[30] *Echolocation in bats.* Scholarpedia website. 2015. URL: http://www.scholarpedia.org/article/Echolocation_in_bats.

[31] R.T. Edwards and G. Cauwenberghs. "Synthesis of Log-Domain Filters from First-Order Building Blocks." In: *International Journal of Analog Integrated Circuits and Signal Processing* 22 (2000), pp. 177–186.

[32] S. K. Esser, P. A. Merolla, J. V. Arthur, A. S. Cassidy, R. Appuswamy, A. Andreopoulos, D. J. Berg, J. L. McKinstry, T. Melano, D. R. Barch, C. di Nolfo, P. Datta, A. Amir, B. Taba, M. D. Flickner, and D. S. Modha. "Convolutional Networks for Fast, Energy-Efficient Neuromorphic Computing." In: *CoRR* abs/1603.08270 (2016). URL: http://arxiv.org/abs/1603.08270.

[33] F. Farkhooi, E. Muller, and M.P. Nawrot. "Adaptation reduces variability of the neuronal population code." In: *Physical Review E* 83.5 (2011), p. 050905.

[34] F. Farkhooi, A. Froese, E. Muller, R. Menzel, and M.P. Nawrot. "Cellular adaptation facilitates sparse and reliable coding in sensory pathways." In: *PLoS Comput Biol* 9.10 (2013), e1003251.

[35] F. Folowosele, R. Etienne-Cummings, and T.J. Hamilton. "A CMOS Switched Capacitor Implementation of the Mihalas-Niebur Neuron." In: *Biomedical Circuits and Systems Conference, BIOCAS 2009*. IEEE. 2009, pp. 105–108. DOI: 10.1109/BIOCAS.2009.5372048.

[36] F. Folowosele, A. Harrison, A. Cassidy, A.G. Andreou, R. Etienne-Cummings, S Mihalas, Niebur, and T.J. Hamilton. "A Switched Capacitor Implementation of the Generalized Linear Integrate-And-Fire Neuron." In: *International Symposium on Circuits and Systems, ISCAS 2009*. IEEE. 2009, pp. 2149–2152.

[37] Simon Friedmann, Johannes Schemmel, Andreas Grübl, Andreas Hartel, Matthias Hock, and Karlheinz Meier. "Demonstrating hybrid learning in a flexible neuromorphic hardware system." In: *IEEE transactions on biomedical circuits and systems* 11.1 (2017), pp. 128–142.

[38] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana. "The SpiNNaker Project." In: *Proceedings of the IEEE* 102.5 (2014), pp. 652–665. ISSN: 0018-9219. DOI: 10.1109/JPROC.2014.2304638.

[39] S. Fusi, M. Annunziato, D. Badoni, A. Salamon, and D. J. Amit. "Spike-driven synaptic plasticity: Theory, simulation, VLSI implementation." In: *Neural Computation* 12 (2000), pp. 2227–2258.

[40] D.F.M. Goodman and R. Brette. "The brian simulator." In: *Frontiers in neuroscience* 3 (2009), p. 26.

[41] R.M. Hennig. "Acoustic feature extraction by cross-correlation in crickets?" In: *Journal of Comparative Physiology A* 189.8 (2003), pp. 589–598.

[42] R.M. Hennig. "Walking in Fourier's space: algorithms for the computation of periodicities in song patterns by the cricket Gryllus bimaculatus." In: *Journal of Comparative Physiology A* 195.10 (2009), pp. 971–987.

[43] R.M. Hennig, K. Heller, and J. Clemens. "Time and timing in the acoustic recognition system of crickets." In: *Frontiers in Physiology* 5.286 (2014). ISSN: 1664-042X. DOI: 10.3389/fphys.2014.00286. URL: http://www.frontiersin.org/integrative_physiology/10.3389/fphys.2014.00286/abstract.

[44] R.R. Hoy. "Acoustic communication in crickets: a model system for the study of feature detection." In: *Federation proceedings*. Vol. 37. 10. 1978, pp. 2316–2323.

[45] B. Hutcheon and Y. Yarom. "Resonance, oscillation and the intrinsic frequency preferences of neurons." In: *Trends in Neurosciences* 23.5 (2000), pp. 216 –222. ISSN: 0166-2236. DOI: http://dx.doi.org/10.1016/S0166-2236(00)01547-2. URL: http://www.sciencedirect.com/science/article/pii/S0166223600015472.

[46] G. Indiveri. "A low-power adaptive integrate-and-fire neuron circuit." In: *International Symposium on Circuits and Systems, ISCAS 2003*. IEEE. 2003, pp. IV–820–IV–823. DOI: 10.1109/ISCAS.2003.1206342.

[47] G. Indiveri. "Modeling Selective Attention using a Neuromorphic analog VLSI Device." In: *Neural Computation* 12.12 (2000), pp. 2857–2880. DOI: 10.1162/089976600300014755.

[48] G. Indiveri. "Neuromorphic VLSI models of selective attention: from single chip vision sensors to multi-chip systems." In: *Sensors* 8.9 (2008), pp. 5352–5375. ISSN: 1424-8220. DOI: 10.3390/s8095352. URL: http://www.mdpi.com/1424-8220/8/9/5352.

[49] G. Indiveri, E. Chicca, and R.J. Douglas. "A VLSI array of low-power spiking neurons and bistable synapses with spike–timing dependent plasticity." In: *IEEE Transactions on Neural Networks* 17.1 (2006), pp. 211–221. DOI: 10.1109/TNN.2005.860850. URL: http://ncs.ethz.ch/pubs/pdf/Indiveri_etal06.pdf.

[50] G. Indiveri, B. Linares-Barranco, T.J. Hamilton, A. van Schaik, R. Etienne-Cummings, T. Delbruck, S.-C. Liu, P. Dudek, P. Häfliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. Saighi, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, and K. Boahen. "Neuromorphic silicon neuron circuits." In: *Frontiers in Neuroscience* 5 (2011), pp. 1–23. ISSN: 1662-453X. DOI: 10.3389/fnins.2011.00073.

[51] Giacomo Indiveri, Federico Corradi, and Ning Qiao. “Neuromorphic architectures for spiking deep neural networks.” In: *Electron Devices Meeting (IEDM), 2015 IEEE International*. IEEE. 2015, pp. 4–2.

[52] J.S. Isaacson and B. Walmsley. “Amplitude and time course of spontaneous and evoked excitatory postsynaptic currents in bushy cells of the anteroventral cochlear nucleus.” In: *Journal of Neurophysiology* 76.3 (1996), pp. 1566–1571.

[53] E. M. Izhikevich. “Simple model of spiking neurons.” In: *IEEE Transactions on Neural Networks* 14.6 (2003), pp. 1569–1572. DOI: 10.1109/TNN.2003.820440.

[54] E.M. Izhikevich, N.S. Desai, E.C. Walcott, and F.C. Hoppensteadt. “Bursts as a unit of neural information: selective communication via resonance.” In: *Trends in Neurosciences* 26.3 (2003), pp. 161–167. DOI: 10.1016/S0166-2236(03)00034-1. URL: http://dx.doi.org/10.1016/S0166-2236(03)00034-1.

[55] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu. “Nanoscale memristor device as synapse in neuromorphic systems.” In: *Nano letters* 10.4 (2010), pp. 1297–1301.

[56] E.R. Kandel, J.H. Schwartz, and T.M. Jessell. *Principles of Neural Science*. Mc Graw Hill, 2000.

[57] K. Kostarakos and B. Hedwig. “Calling song recognition in female crickets: temporal tuning of identified brain neurons matches behavior.” In: *The Journal of Neuroscience* 32.28 (2012), pp. 9601–9612.

[58] K. Kostarakos and B. Hedwig. “Calling song recognition in female crickets: temporal tuning of identified brain neurons matches behavior.” In: *The Journal of Neuroscience* 32.28 (2012), pp. 9601–9612.

[59] M. Lau. “Characterizing the Dynamical Response Properties of Silicon Neurons.” MA thesis. Universität Bielefeld, 2016.

[60] J. Lazzaro and J. Wawrzynek. *Low-power silicon neurons, axons, and synapses*. Tech. rep. UC Berkley, 1992.

[61] J.P. Lazzaro. “Silicon Implementation of Pulse Coded Neural Networks.” In: ed. by M.E. Zaghloul, J.L. Meador, and R.W. Newcomb. Kluwer Academic Publishers, 1994. Chap. Low-power silicon axons, neurons, and synapses, pp. 153–164.

[62] P. Lichtsteiner, C. Posch, and T. Delbruck. “An 128x128 120dB 15μs-latency temporal contrast vision sensor.” In: *IEEE J. Solid State Circuits* 43.2 (2008), pp. 566–576.

[63] H. Lim, I. Kim, J. Kim, C.S. Hwang, and D.S. Jeong. “Short-term memory of TiO2-based electrochemical capacitors: empirical analysis with adoption of a sliding threshold.” In: *Nanotechnology* 24.38 (2013), p. 384005.

[64] S.-C. Liu, J. Kramer, G. Indiveri, T. Delbruck, and R.J. Douglas. *Analog VLSI:Circuits and Principles.* MIT Press, 2002.

[65] S.-C. Liu, A.V. Schaik, B.A. Mincti, and T. Delbruck. “Event-based 64-channel binaural silicon cochlea with Q enhancement mechanisms.” In: *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on.* IEEE. 2010, pp. 2027–2030.

[66] S.-C. Liu, J. Kramer, G. Indiveri, T. Delbruck, T. Burg, and R.J. Douglas. “Orientation-selective aVLSI spiking neurons.” In: *Neural Networks* 14.6/7 (2001), pp. 629–643.

[67] H. Lorach, R. Benosman, O. Marre, S.-H. Ieng, J.A. Sahel, and S. Picaud. “Artificial retina: the multichannel processing of the mammalian retina achieved with a neuromorphic asynchronous light acquisition device.” In: *Journal of neural engineering* 9.6 (2012), p. 066004.

[68] S. Luck, L. Chelazzi, S.A. Hillyard, and R. Desimone. “Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex.” In: *Journal of neurophysiology* 77 (1997), pp. 24–42.

[69] K.M. MacLeod, T.K. Horiuchi, and C.E. Carr. “A role for short-term synaptic facilitation and depression in the processing of intensity information in the auditory brain stem.” In: *Journal of neurophysiology* 97.4 (2007), pp. 2863–2874.

[70] M. Mahowald and C. Mead. “Analog VLSI and Neural Systems.” In: Reading, MA: Addison-Wesley, 1989. Chap. Silicon Retina, pp. 257–278.

[71] H. Markram, D. Pikus, A. Gupta, and M. Tsodyks. “Potential for multiple mechanisms, phenomena and algorithms for synaptic plasticity at single synapses.” In: *Neuropharmacology* 37.4 (1998), pp. 489–500.

[72] C.A. Mead. *Analog VLSI and Neural Systems.* Reading, MA: Addison-Wesley, 1989.

[73] G. Meckenhäuser, R.M. Hennig, and M.P. Nawrot. “Critical song features for auditory pattern recognition in crickets.” In: *PloS one* 8.2 (2013), e55349.

[74] G. Meckenhäuser, S. Krämer, F. Farkhooi, B. Ronacher, and M.P. Nawrot. "Neural representation of calling songs and their behavioral relevance in the grasshopper auditory system." In: *Name: Frontiers in Systems Neuroscience* 8 (2014), p. 183.

[75] R. Menzel and G. Manz. "Neural plasticity of mushroom body-extrinsic neurons in the honeybee brain." In: *Journal of Experimental Biology* 208.22 (2005), pp. 4317–4332. ISSN: 0022-0949. DOI: 10.1242/jeb.01908. eprint: http://jeb.biologists.org/content/208/22/4317.full.pdf. URL: http://jeb.biologists.org/content/208/22/4317.

[76] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha. "A million spiking-neuron integrated circuit with a scalable communication network and interface." In: *Science* 345.6197 (2014), pp. 668–673. ISSN: 0036-8075. DOI: 10.1126/science.1254642. eprint: http://science.sciencemag.org/content/345/6197/668.full.pdf. URL: http://science.sciencemag.org/content/345/6197/668.

[77] Y. Miyashita. "Neuronal correlate of visual associative long-term memory in the primate temporal cortex." In: *Nature* (1988).

[78] *Mouse Brain Atlas: DBA/2J Coronal.* 2003. URL: http://www.mbl.org/atlas165/atlas165_start.html.

[79] A.F. Murray. "Pulse-based computation in VLSI neural networks." In: *Pulsed Neural Networks.* Ed. by W. Maass and C.M. Bishop. MIT Press, 1998. Chap. 3, pp. 87–109.

[80] E. Neftci. "Towards VLSI Spiking Neuron Assemblies as General-Purpose Processors." Ph.D. thesis. ETH Zürich, 2010.

[81] E. Neftci and G. Indiveri. "A Device Mismatch Compensation Method for VLSI Spiking Neural Networks." In: *Biomedical Circuits and Systems Conference BIOCAS 2010.* IEEE. 2010, pp. 262–265. DOI: 10.1109/BIOCAS.2010.5709621.

[82] M. Noack, C. Mayr, J. Partzsch, M. Schultz, and R. Schüffny. "A switched-capacitor implementation of short-term synaptic dynamics." In: *Mixed Design of Integrated Circuits and Systems (MIXDES), 2012 Proceedings of the 19th International Conference.* IEEE. 2012, pp. 214–218.

[83] M. Noack, J. Partzsch, C. Mayr, S. Hänzsche, S. Scholze, S. Höppner, G. Ellguth, and R. Schüffny. "Switched-capacitor realization of presynaptic short-term-plasticity and stop-learning synapses in 28 nm CMOS." In: *arXiv preprint arXiv:1412.3243* (2014).

[84] M.J.M. Pelgrom, A.C.J. Duinmaijer, and A.P.G. Welbers. "Matching properties of MOS transistors." In: *IEEE Journal of Solid-State Circuits* 24.5 (1989), pp. 1433–1440.

[85] F. Perez-Peña, A. Linares-Barranco, and E. Chicca. "An approach to motor control for spike-based neuromorphic robotics." In: *2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings*. 2014, pp. 528–531. DOI: 10.1109/BioCAS.2014.6981779.

[86] G.S. Pollack. "Sensory cues for sound localization in the cricket Teleogryllus oceanicus: interaural difference in response strength versus interaural latency difference." English. In: *Journal of Comparative Physiology A* 189.2 (2003), pp. 143–151. ISSN: 0340-7594. DOI: 10.1007/s00359-003-0388-0. URL: http://dx.doi.org/10.1007/s00359-003-0388-0.

[87] G.D. Puccini, M.V. Sanchez-Vives, and A. Compte. "Integrated Mechanisms of Anticipation and Rate-of-Change Computations in Cortical Circuits." In: *PLoS Comput Biol* 3.5 (May 2007), pp. 1–13. DOI: 10.1371/journal.pcbi.0030082. URL: http://dx.plos.org/10.1371%2Fjournal.pcbi.0030082.

[88] N. Qiao, H. Mostafa, F. Corradi, M. Osswald, F. Stefanini, D. Sumislawska, and G. Indiveri. "A Re-configurable On-line Learning Spiking Neuromorphic Processor comprising 256 neurons and 128K synapses." In: *Frontiers in Neuroscience* 9.141 (2015). ISSN: 1662-453X. DOI: 10.3389/fnins.2015.00141.

[89] B.N. Ralston, L.Q. Flagg, E. Faggin, and J.T. Birmingham. "Incorporating spike-rate adaptation into a rate code in mathematical and biological neurons." In: *Journal of neurophysiology* 115.5 (2016), pp. 2501–2518.

[90] H. Ramachandran, S. Weber, S.A. Aamir, and E. Chicca. "Neuromorphic Circuits for Short-term Plasticity with Recovery Control." In: *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. 2014, pp. 858–861. DOI: 10.1109/ISCAS.2014.6865271.

[91] C. Rasche and R. Hahnloser. "Silicon Synaptic Depression." In: *Biological Cybernetics* 84.1 (2001), pp. 57–62.

[92] C.A. Rasche, R.J. Douglas, and K.A.C. Martin. “Analog VLSI circuits for emulating computational features of pyramidal cells.” PhD thesis. Zürich: Naturwissenschaften ETH Zürich, 1999. URL: http://cds.cern.ch/record/893600.

[93] W.G. Regehr. “Short-term presynaptic plasticity.” In: *Cold Spring Harbor Perspectives in Biology* 4.7 (2012), a005702.

[94] W.G. Regehr, K.R. Delaney, and D.W. Tank. “The role of presynaptic calcium in short-term enhancement at the hippocampal mossy fiber synapse.” In: *The Journal of neuroscience* 14.2 (1994), pp. 523–537.

[95] A. Reyes, R. Lujan, A. Rozov, N. Burnashev, P. Somogyi, and B. Sakmann. “Target-cell-specific facilitation and depression in neocortical circuits.” In: *Nature neuroscience* 1.4 (1998), pp. 279–285.

[96] T. Rost. “Modelling Pattern Recognition in Cricket Phonotaxis.” Masters thesis. Freie Universität Berlin, 2011. URL: http://edocs.fu-berlin.de/docs/receive/FUDOCS_document_000000015361.

[97] T. Rost, H. Ramachandran, M.P. Nawrot, and E. Chicca. “A neuromorphic approach to auditory pattern recognition in cricket phonotaxis.” In: *2013 European Conference on Circuit Theory and Design (ECCTD)*. IEEE. 2013, pp. 1–4. DOI: 10.1109/ECCTD.2013.6662247.

[98] S. Saïghi, C. Mayr, T. Serrano-Gotarredona, H. Schmidt, G. Lecerf, J. Tomas, J. Grollier, S. Boyn, A.F. Vincent, D. Querlioz, S. La Barbera, F. Alibart, D. Vuillaume, O. Bichler, C. Gamrat, and B. Linares-Barranco. “Plasticity in memristive devices for spiking neural networks.” In: *Frontiers in neuroscience* 9 (2015), p. 51.

[99] J. Schemmel, D. Bruderle, A. Grubl, M. Hock, K. Meier, and S. Millner. “A wafer-scale neuromorphic hardware system for large-scale neural modeling.” In: *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*. IEEE. 2010, pp. 1947–1950.

[100] K. Schildberger. “Behavioral and neuronal mechanisms of cricket phonotaxis.” In: *Motor Control* 44 (1988), pp. 408–415.

[101] K. Schildberger. “Temporal selectivity of identified auditory neurons in the cricket brain.” In: *Journal of Comparative Physiology A* 155.2 (1984), pp. 171–185. ISSN: 0340-7594. DOI: 10.1007/BF00612635. URL: http://www.springerlink.com/index/10.1007/BF00612635.

[102] M. Schmuker, T. Pfeil, and M.P. Nawrot. “A neuromorphic network for generic multivariate data classification.” In: *Proceedings of the National Academy of Sciences* 111.6 (2014), pp. 2081–2086.

[103] S. Schöneich, K. Kostarakos, and B. Hedwig. “An auditory feature detection circuit for sound pattern recognition.” In: *Science advances* 1.8 (2015), e1500325.

[104] S. Sheik, E. Chicca, and G. Indiveri. “Exploiting Device Mismatch in Neuromorphic VLSI Systems to Implement Axonal Delays.” In: *The 2012 International Joint Conference on Neural Networks (IJCNN).* IEEE. 2012, pp. 1940–1945. DOI: 10.1109/IJCNN.2012.6252636. URL: http://ncs.ethz.ch/pubs/pdf/Sheik_etal12b.pdf.

[105] J. Thornson, T. Weber, and F. Huber. “Auditory behaviour of the cricket: II. Simplicity of calling song recognition in Gryllus and anomalous phonotaxis at abnormal carrier frequency.” In: *Journal of Comparative Physiology A* 146 (1982), pp. 361–378.

[106] S. Trannoy, C. Redt-Clouet, J. Dura, and T. Preat. “Parallel Processing of Appetitive Short- and Long-Term Memories In Drosophila.” In: *Current Biology* 21.19 (2011), pp. 1647 –1653. ISSN: 0960-9822. DOI: http://dx.doi.org/10.1016/j.cub.2011.08.032. URL: http://www.sciencedirect.com/science/article/pii/S0960982211009389.

[107] M.V. Tsodyks and H. Markram. “The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability.” In: *Proceedings of the National Academy of Sciences of the USA* 94.2 (1997), pp. 719–723.

[108] J.A. Varela, K. Sen, J. Gibson, J. Fost, L.F. Abbott, and S.B. Nelson. “A quantitative description of short-term plasticity at excitatory synapses in layer 2/3 of rat primary visual cortex.” In: *The Journal of neuroscience* 17.20 (1997), pp. 7926–7940.

[109] R.J. Vogelstein, U. Mallik, J.T. Vogelstein, and G. Cauwenberghs. “Dynamically Reconfigurable Silicon Array of Spiking Neurons With Conductance-Based Synapses.” In: *IEEE Transactions on Neural Networks* 18.1 (2007), pp. 253–265.

[110] D.M. Walsh, I. Klyubin, J.V. Fadeeva, W.K. Cullen, R. Anwyl, M.S. Wolfe, M.J. Rowan, and D.J. Selkoe. “Naturally secreted oligomers of amyloid β protein potently inhibit hippocampal long-term potentiation in vivo.” In: *Nature* 416.6880 (2002), pp. 535–539.

[111] Z.Q. Wang, H.Y. Xu, X.H. Li, H. Yu, Y.C. Liu, and X.J. Zhu. "Synaptic learning and memory functions achieved using oxygen ion migration/diffusion in an amorphous InGaZnO memristor." In: *Advanced Functional Materials* 22.13 (2012), pp. 2759–2765.

[112] P. Xu, T.K. Horiuchi, A. Sarje, and P. Abshire. "Stochastic synapse with short-term depression for silicon neurons." In: *2007 IEEE Biomedical Circuits and Systems Conference*. IEEE. 2007, pp. 99–102.

[113] M. Zorović and B. Hedwig. "Processing of species-specific auditory patterns in the cricket brain by ascending, local, and descending neurons during standing and walking." In: *Journal of Neurophysiology* 105.5 (2011), pp. 2181–2194. ISSN: 0022-3077. DOI: 10.1152/jn.00416.2010.

[114] R.S. Zucker. "Short-term synaptic plasticity." In: *Annual review of neuroscience* 12.1 (1989), pp. 13–31.

[115] R.S. Zucker and W.G. Regehr. "Short-term synaptic plasticity." In: *Annual Review of Physiology* 64 (2002), pp. 355–405.