Université de Lille

Université de Lille, Ecole graduée ENGSYS-632
Unité de recherche IEMN UMR 8520

Thèse par **Eugénie Dalmas**

Pour l'obtention du grade de **Docteur de l'Université de Lille**

Spécialité Electronique, Microélectronique, Nanoélectronique et Micro-ondes
Domaine Sciences et technologies de l'information et de la communication

Bioinspired Ultra-Low Power Architectures for Sound Source Localization and Recognition

Architectures Bioinspirées Ultra-Faible Consommation pour la Localisation et la Reconnaissance de Sources Sonores

Soutenue le 5 décembre 2025, à Villeneuve-d'Ascq

Membres du jury

Président du jury	**Pierre BOULET**	Professeur des universités, Univ. de Lille / CRIStAL
Rapporteurs	**Blaise YVERT**	Directeur de recherche, Inserm Grenoble Alpes
	Damien QUERLIOZ	Directeur de recherche CNRS, C2N Paris-Saclay
Examinatrice	**Elena Ioana VATAJELU**	Chargée de recherche CNRS, Laboratoire TIMA
Invités	**Michael BOCQUET**	Maître de conférences, UPHF / IEMN
	Fouzia BOUKOUR	Directrice de recherche, Univ. Gustave Eiffel / LEOST
Direction de thèse	**Christophe LOYEZ**	Directeur de recherche CNRS, IEMN
Co-directeur de thèse	**François DANNEVILLE**	Professeur des universités, Univ. de Lille / IEMN


iemn
Institut d'Electronique,
de Microélectronique
et de Nanotechnologie
UMR CNRS 8520
IRCICA
cnrs
anr
agence nationale
de la recherche

Titre: Architectures Bioinspirées Ultra-Faible Consommation pour la Localisation et la Reconnaissance de Sources Sonores

Mots clés: neurones à spikes, technologie neuromorphique, ultra-faible consommation, localisation de sources sonores, reconnaissance acoustique

Résumé: Le biomimétisme est une approche de plus en plus répandue dans les différents domaines scientifiques. Il est régulièrement la source de nouveaux paradigmes et, depuis quelques années, a impulsé le traitement et les technologies neuromorphiques qui portent la promesse d'avancées significatives dans le domaine de la théorie de l'information et une efficacité énergétique sans précédent. Avec cette approche, les systèmes artificiels à impulsions inspirés du traitement neuronal dans le cerveau peuvent traiter des signaux issus de diverses modalités. Dans le contexte de la surveillance acoustique de la biodiversité, cette thèse explore le potentiel d'une technologie neuromorphique analogique intégrant des transistors à effet de champ métal-oxyde-semi-conducteur fonctionnant en régime sous-le-seuil avec une puissance consommée ultra-faible (ULP). En tenant compte des contraintes ULP, des outils de pré-traitement bioinspirés et économes en énergie sont conçus pour la localisation et la reconnaissance des sources sonores et leur performances sont évaluées. Tout d'abord, un extracteur original de différence interaurale de temps (ITD) est modélisé d'après le détecteur de mouvement Hassenstein-Reichardt, choisi pour son faible nombre de neurones, et appliqué aux signaux acoustiques pour estimer la direction d'arrivée des sources sonores. L'extracteur d'ITD est évalué en simulation sur la base d'enregistrements en intérieur en 2D et 3D à des distances comprises entre 24 cm et 10 m, en particulier pour des sons de type clic. Une technique simplifiée de multilatération hyperbolique permet d'analyser les performances de localisation de l'extracteur d'ITD, et qui résulte en des précisions azimutales encourageantes, compte tenu de sa consommation potentielle d'ULP, de 73,9 % (±2,5°) et 77 % (±5°) entre 1 m et 3 m pour les sons de type clic. Ensuite, dans le but de traiter des scénarios multisources, un détecteur de caractéristiques temporelles inspiré du mécanisme de reconnaissance du chant d'appel des criquets femelles a été conçu et intégré sur puce à partir de la technologie neuromorphique sous-le-seuil. Testé sur un banc sous pointes avec des signaux artificiels et des chants d'appel de criquets réels, le détecteur atteint une consommation totale inférieure au nanowatt dans des scénarios calmes ou bruyants, avec une grande précision et un recall encourageant. Finalement, la combinaison de plusieurs instances de ces deux outils de pré-traitement permet d'envisager des applications de suivi et de comptage des sources acoustiques.

Title: Bioinspired Ultra-Low Power Architectures for Sound Source Localization and Recognition

Keywords: spiking neurons, neuromorphic technology, ultra-low power, sound source localization, acoustic recognition

Abstract: Biomimetism is an increasingly widespread approach in various scientific fields. It regularly gives rise to new paradigms and, in recent years, has driven neuromorphic computing and technologies which promise significant advances in the field of information theory and unprecedented energy efficiency. With this approach, artificial spiking systems inspired by the neuronal impulse processing in the brain can process signals from various modalities. In the context of acoustic monitoring of biodiversity, this thesis investigates the potential of an analog neuromorphic technology integrating metal-oxide-semiconductor field-effect transistors operating in the subthreshold regime with ultra-low power (ULP) consumption. Keeping ULP constraints under consideration, bioinspired energy-efficient precomputing tools are designed for sound source localization and recognition and their performances assessed. Firstly, an original interaural time difference (ITD) extractor is modelled after the Hassenstein-Reichardt detector of motion detection, chosen for its low number of neurons, and applied to acoustic signals for estimation of sound sources' direction of arrival. The ITD extractor is evaluated in simulation on the basis of 2-D and 3-D indoor recordings at distances between 24 cm and 10 m of click-like sounds in particular. A simplified hyperbolic multilateration technique enables the analysis of the ITD extractor's localization performances, resulting in encouraging azimuth accuracies, in view of its potential ULP consumption, of 73.9% (±2.5°) and 77% (±5°) between 1 m and 3 m for click-like sounds. Then, with the aim to address multisource scenarios, a detector of temporal characteristics inspired by the calling song recognition mechanism of female field crickets is designed and successfully implemented on chip using the subthreshold neuromorphic technology. Tested under a probe station with artificial and real-world cricket calling songs, the detector reaches a sub-nanowatt total power consumption in quiet or noisy scenarios with high precision and encouraging recall. Finally, combining multiple instances of these two precomputing tools enables one to envision acoustic source tracking and counting applications.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my PhD supervisors, Christophe Loyez and François Danneville for their continuous guidance, trust, and encouragement throughout the course of this research. I have greatly benefited from their experience, expertise, and feedback at every stage of this research.

I thank the researchers with whom I exchanged on my work, especially Michael Bocquet, Fouzia Boukour, Sébastien Paris, and Hervé Glotin, for their insightful advices on how to improve my approach and research. Moreover, Michael Bocquet crafted the 4-microphone wooden stand allowing me to produce 3-D localization recording sets and for which I am thankful.

I thank the laboratory IEMN for providing me access to the probe station, precision instruments, and essential equipment. Furthermore, the research engineers Hicham Larach and Kevin Carpentier of the IEMN team kindly assisted me with the setup and use of the probe station, without which my results under probes would not have been obtained. Especially, I am grateful to Kevin Carpentier who greatly contributed to the integration on-chip of the temporal characteristic detection circuit presented in this thesis.

I am also grateful to the ANR Project ULP SMART 3D COCHLEA for their financial support without which this PhD would not have been possible.

Finally, I wish to thank my entourage for their kind support, and especially Rémi Arbache for his help in refining the language of this manuscript and associated publications.

Table of Contents

List of Abbreviations

AI	Artificial Intelligence
ULP	Ultra-Low Power
CMOS	Complementary Metal-Oxide-Semiconductor
Na^+	Sodium
K^+	Potassium
HH	Hodgkin-Huxley
ML	Morris-Lecar
IF	Integrate-and-Fire
LIF	Leaky Integrate-and-Fire
AdEx	Adaptative Exponential Integrate-and-Fire
SNN	Spiking Neural Network
CPU	Central Processing Unit
ASIC	Application-Specific Integrated Circuit
RRAM	Resistive Random-Access Memory
SRAM	Static Random Access Memory
MOSFET	Metal-Oxide-Semiconductor Field-Effect Transistor
FPGA	Field-Programmable Gate Array
NMOS	N-type MOS
PMOS	P-type MOS
DC	Direct Current
STDP	Spike Timing-Dependent Plasticity
SSL	Sound Source Localization
DOA	Direction Of Arrival
ITD	Interaural Time Difference
ILD	Interaural Level Difference
IID	Interaural Intensity Difference
HRTF	Head-Related Transfer Function
MSO	Medial Superior Olive
LSO	Lateral Superior Olive
IC	Inferior Colliculus
WTA	Winner-Takes-All
VLSI	Very-Large-Scale Integration
ANN	Artificial Neural Network
LSM	Liquid State Machine
MAE	Mean Absolute Error
HRD	Hassenstein-Reichardt Detector
SNR	Signal-to-Noise Ratio
2-D	2-Dimensional
3-D	3-Dimensional
GDOP	Geometric Dilution of Precision
FP	False Positive

TP	True Positive
FN	False Negative
RMS	Root Mean Square
DAC	Digital-to-Analog Converter

List of Figures

List of Tables

1
Introduction

The complexity of biological systems has never ceased to be a source of amazement and currently drives many research domains thanks to its efficiency in completing tasks in ideal or challenging conditions. In fact, nature and living beings have always been a source of inspiration.

Bioinspiration can take place at different levels, such as the extraction of data issued from sensors, data encoding, and data processing, but also in a more general approach, in relation to behavioral patterns that an individual may express in a group or in interaction with its environment. Depending on the approach chosen for the creation of a system or subsystem, the design can be completely biomimetic and even bioplausible –as is mostly seen with work derived from neuroscience studies and in the context of biological systems assessments– or, on the contrary, lightly/partially inspired by a biological process or functional aspect. Although originating from a biological inspiration, it sometimes becomes so negligible or diluted that it is no longer mentioned.

Nowadays, one focus of scientific research is turned towards artificial intelligence (AI) with dense neural networks like generative AI, where the massive power consumption is often not questioned. In fact, datacenters specialized in the training and running of AI-powered tools are built to deliver between 1 MW and 1 GW of electric power to numerous servers. The total consumption of datacenters all over the world represents about 2-3% of the world's power consumption, corresponding to about 415 TWh in 2024 and rapidly growing [1]. From a user standpoint, the energy footprint is hidden since computations are performed on remote servers, and yet excessive amounts of energy are consumed annually.

In embedded applications, such overconsumption is not conceivable. Although it can also be considered valid for power-hungry systems like smartphones that are essentially miniature computers, the question of energy efficiency mostly affects battery-powered devices deployed on sites with difficult access or that require extended periods of unperturbed observation. With the difficulty of regularly charging the battery, the autonomy of such remote devices becomes

one of the top priorities. Embedded electronics is now everywhere in our lives, be it basic home controllers or precision tools for research purposes. In this context, but also of the energy transition, greener solutions are sought after. Different approaches may be taken, like further investing in the existing hardware but developing more efficient computing units, from smaller nano-scaled transistors [2], [3] to less fundamental studies on efficient accelerators [4], [5], or in this case by rethinking how information is represented and processed.

For that matter, bioinspiration is often the origin of new paradigms and, few decades ago, has set into motion neuromorphic computing which holds the promise of significant advances in the domain of information theory. It is inspired by the neuronal processing of information in the brain, such that the computing units of the devices are artificial neurons connected by artificial synapses, acting as memory units that weigh the impact of a neuron's output to another neuron, and which use spikes as a common encoding of all sensor modalities (vison, hearing, etc.) for a sparse representation of the information both temporarily and spatially. Projecting the mechanisms observed in biological systems, neuromorphic technologies are brought forward and proposed as solutions to problems also encountered in the living. From bio-plausible models to lighter inspiration, neuromorphic solutions aim to provide energy efficiency similar or even better than its biological counterpart.

Great advances in all applicative fields led to the production of numerous hardware implementations, mostly digital or mixed-signal in the context of computer vision. Yet, despite many research works addressing the question of mimicking biology to various degrees with artificial neuromorphic systems, few were concerned about the development of true ultra-low power (ULP) consuming circuits. While digital neuromorphic processors showcase high accuracy performances and an advanced technological maturity, they fail to reach the ULP consumptions of emerging technologies or fully analog implementations due to generally non-dedicated chips and bulky hardware. Generally consuming several hundreds of microwatts or more, it is still at least a thousand times more power consuming than sub-microwatt systems. Indeed, for a dedicated application, if the circuit power consumption is reduced below the microwatt and with a battery delivering 1000 mA.h, no maintenance would be required for ten years.

Driven by power efficiency considerations, the research team of IEMN supervising this PhD previously developed a highly energy-efficient analog neuromorphic technology, fully asynchronous, based on standard complementary metal-oxide-semiconductor (CMOS) technology and likely able to meet this challenge for a number of applications. Among them, this PhD investigates the potential of the team's analog neuromorphic technology for biodiversity acoustic monitoring through the design and study of bioinspired estimators. This research is funded by the ANR project ULP SMART 3D COCHLEA for which IEMN works in partnership with CRIStAL laboratory in a hardware-software approach for the proposition of neuromorphic solutions for acoustic monitoring.

This research is motivated by bioacousticians' need of non-intrusive and energy-efficient tools for localization and recognition of animal activities, transposable to anthropogenic sounds. In particular, this thesis explores the localization of biodiversity sound sources combined with a recognition mechanism for enhanced tracking. Against the backdrop of stringent energy

constraints, the achievable performance in acoustic monitoring are evaluated through the development of neuromorphic extractors and estimators as precomputing tools tailored to sub-microwatt applications. An emphasis is put on providing ULP spiking systems implementable on-chip that extract critical information from input sound signals for further processing, either conventional or neuromorphic. The focus is determining the possibilities that these systems offer and their potential with these hardware constraints.

Guided by these considerations, the main objectives of this thesis can be articulated as follows:

- Design circuits for acoustic signals processing at the precomputing level compatible with subthreshold neuromorphic technology.
- Localize sounds in real-world conditions using a bioinspired model.
- Perform joint bioinspired recognition for enhanced sound localization in multisource scenarios.
- Implement a circuit using subthreshold neuromorphic technology and achieve state-of-the-art performances in terms of power consumption in the context of acoustic monitoring with a mature technology.

This thesis reports the work that was produced to complete the objectives listed according to the plan thereafter.

Following this general introduction, chapter 2 provides basic knowledge on the core neuronal elements on which any neuromorphic computing or technology is based. Then, the subthreshold analog neuromorphic technology at the center of this thesis is detailed.

In chapter 3, the fundamental mechanisms of the auditive system are described with reference to the mammalian system, and an overview of the underlying neuronal processing is provided. The literature's neuromorphic sound source localization systems are reviewed, categorized by the main binaural cues they rely on to extract the trends and discuss the overall methods, performances, and bioinspiration level reported. Disclosed power consumptions are identified which highlight the compromise between energy efficiency and accuracy.

Then, chapter 4 introduces a neuromorphic circuit issued from coincidence detection in vision, adapted for sound source localization and compatible with the team's neuromorphic technology. Chosen for its simplicity, the proposed time delay extractor is evaluated for 2-D and extended to 3-D direction of arrival estimation using a simplified multilateration technique. Accuracy and potential of such ULP circuit is assessed.

In chapter 5, a detector of temporal patterns inspired by neuronal processing in female field crickets is investigated with the same motivation and implemented in hardware. The resulting demonstrator's detection and power consumption performances are quantified under a probe station.

Finally in chapter 6, the two circuits open on various acoustic monitoring scenarios and applications with ULP consumptions which are overviewed. Especially, the time delay extractor and temporal pattern detector are combined for investigation of multisource scenarios. Overall conclusions are then drawn on this thesis' work that assess the objectives completion and perspectives.

2
Analog Neuromorphic Technology Fundamentals

At the core of all living beings' perception, motricity, understanding, or interpretation, lies the brain. An interconnected ensemble of billions of neurons and trillions of synapses condensed in a volume of about one liter, and more precisely on the cerebral cortex, a thin folded layer of grey matter spreading on around 2500 cm² with thickness 2~3 mm where information is processed.

The general public or lay audience, as much as experts in neurology, always took great interest in the hidden workings of this fine machinery. With nervous connections descending in afferent ends in all parts of the body, every thought, movement, or sensory entry is integrated and analyzed at different levels of abstraction in the cortex. It is no wonder that intensive research following a bioinspired approach is conducted to mimic and surpass the great abilities observed in the living.

From the mathematical formalization of neuronal dynamics, artificial neural networks with bio-plausible responses are simulated in software and implemented in neuromorphic hardware technologies. To better understand what *neuromorphic* means and implies when talking about computing and technology, explanations on fundamental neuronal principles and how that unfold on artificial designs are required.

Therefore, this chapter describes the main neuronal elements and mathematical formalism of neurons, on which neuromorphic technologies build on. Then, neuromorphic technologies are shortly reviewed for introduction of the analog CMOS neuronal units operating in the subthreshold regime employed in this PhD.

2.1 Basic Neuronal Elements and Mechanisms

Taking a closer look at two but massively connected basic neurological components, we review in this section the essential biological observations and characteristics of these processing units, that are neurons and synapses, starting with the spike representation [6], [7].

2.1.1 Biological Observations

All sentient beings share the use of action potentials, also called spikes, as a key mechanism for encoding and transmitting neuronal information. The nervous system employs a diverse array of methods to encode and process information, but spikes are crucial and ubiquitous. They are sparse and asynchronous events in continuous time, and hold the information temporally and spatially. In fact, the shape of the spikes is relatively the same from one neuron to another, with little variation in duration and amplitude. The information is thus held in the timing and number of the spikes, in other words the spiking activity of the neurons.

Spikes are short electric pulses with typical dynamic amplitude around 100 mV and duration of 1~2 ms. Intracellular electrodes are used to monitor the spiking activity of living neurons by capturing the potential difference between the neuron and its surroundings, referred as the membrane potential. Spikes are characterized by a depolarization of the membrane from its resting potential, at about -70 mV, until a peak is reached around 30 mV, followed by a repolarization and then a hyperpolarization below the resting potential. Finally, the membrane potential slowly returns to this resting potential. This impulse form is the result of an integration (or summation) and non-linear process of electrical currents at the neuron's soma.

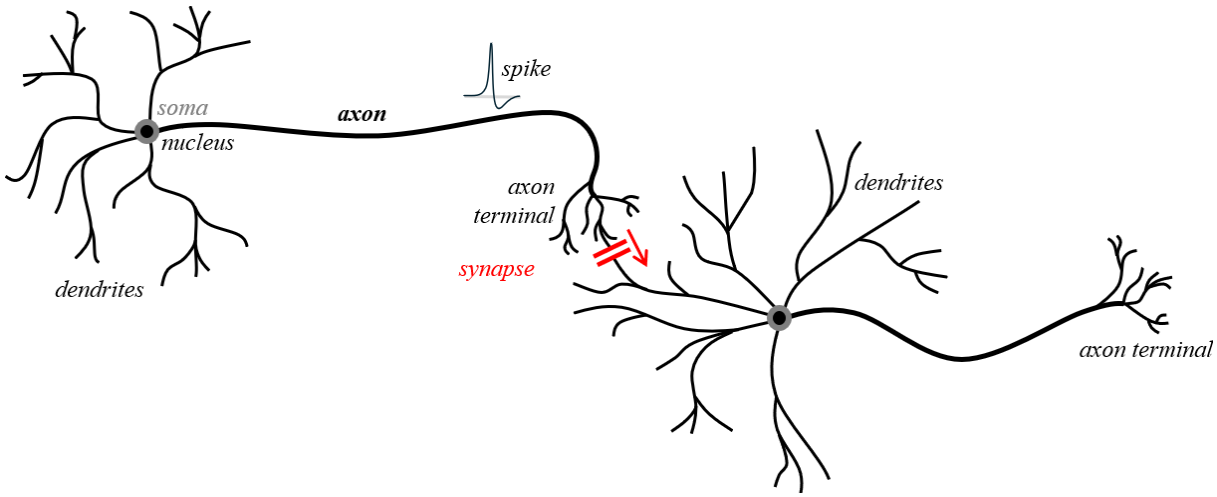


Fig. 2.1 Schematic neuron cell with connection to another neuron. Spikes generated at the soma are propagated along the axon until the axon terminal where they are then transmitted to the other neuron's dendrites through a synapse (in red).

Neurons possess three main parts (Fig. 2.1). The soma holds the nucleus and ensures vital operation of the neuron cell, but also provides the non-linear processing essential to the generation of spikes. Dendrites linked to the soma are tree-structured connections to other

neurons from which spikes are captured to be integrated at the soma. The axon propagates the generated spikes without attenuation, ending in terminal arborization for connection to other neurons' dendrites. These junctions whose role is to transmit the electrical pulses, are called synapses.

In chemical synapses, a tiny gap, the synaptic cleft, separates the terminal arborization of a pre-synaptic neuron (pre-neuron) and the dendritic tree of a post-synaptic neuron (post-neuron). Unlike electrical synapses, chemical synapses are unidirectional but vastly present; electrical synapses allow direct electrical connection between two neurons, but are not found in the cortex. For chemical synapses, a chain of biochemical reactions is triggered when a spike is received, releasing neurotransmitters, that are chemical components which carry information between neurons [8], such as dopamine or adrenaline. Specific receptors on the membrane of the post-neuron capture the neurotransmitters by opening channels, causing an influx of ions from the extracellular to the neuron intracellular fluids and the generation of post-synaptic potentials.

In biology, synapses are either excitatory or inhibitory contributing to the depolarization or hyperpolarization of membrane potentials respectively. These two fundamental types of connections between neurons regulate the flow of spiking activity in the nervous system. Together they provide opposing interactions, maintaining a balance in the brain for proper processing by preventing overstimulation and allowing controlled neuronal communication.

Ionic gates of membrane are crucial for the generation of spikes by regulating the difference in ion concentration in and out of the neuron, thus generating a voltage called the Nernst potential and defined by,

$$E_x = \frac{kT}{q} \ln\left(\frac{n_2}{n_1}\right) \tag{2.1}$$

where E_x is the Nernst potential of ion x, k the Boltzmann constant, T the temperature, q the elementary charge, n_1 and n_2 the concentration of two regions of ions. At room temperature, we note that the thermal voltage $V_T = kT/q$ equals 26 mV. At the neuron, integration of dendritic electrical responses is performed both temporally and spatially as schematized in Fig. 2.2.

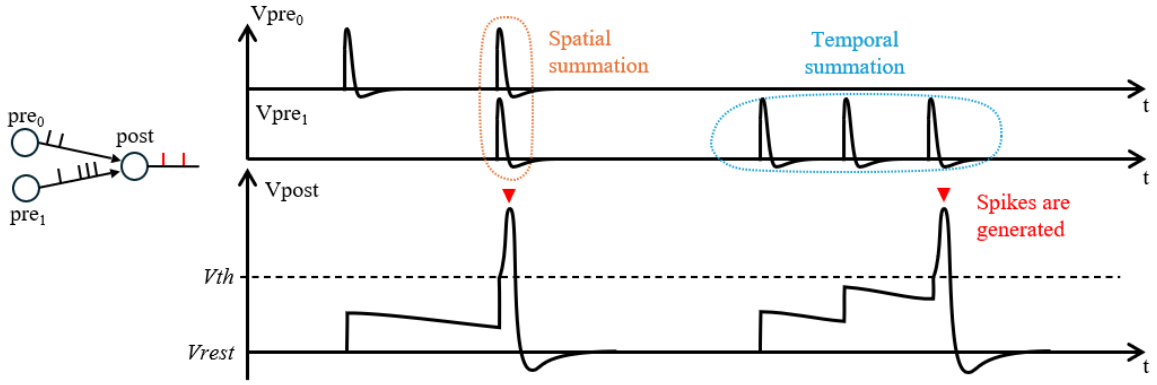


Fig. 2.2 Schematized spike generation at a post-synaptic neuron from two pre-synaptic neurons with leaky behavior, that is a slow decrease over time of the membrane potential $Vpre_0$, $Vpre_1$, or Vpost, towards its rest potential *Vrest*. Upon reaching some threshold potential *Vth*, the membrane generates a spike resulting from the continuous spatial and temporal integration of pre-synaptic spikes.

The voltage-gated ion channels, sodium (Na^+) and potassium (K^+) channels, are responsible for the rapid changes in membrane potential during spikes. When a neuron is stimulated past a threshold, voltage-gated Na^+ channels open, leading to depolarization, followed by the opening of voltage-gated K^+ channels that repolarize the membrane. The selective permeability of ion channels and the active transport of ions by ion pumps, combined with the concentration gradients of ions across the membrane, gives rise to the resting membrane potential V_r.

2.1.2 Mathematical Formalisms

A wide range of mathematical models provide well-defined tools for studying the behavior of a single neuron or a population of neurons. Several parameters may be taken into account to be more or less accurate and faithful to neurological observations previously described.

Hodgkin-Huxley Biological Model

In 1952, Hodgkin and Huxley investigated the giant axon of the squid in numerous experiments, and measured currents passing through ion channels [9]. They introduced non-linear differential equations describing the ionic mechanisms behind the generation and propagation of spikes, which provided an essential base for the formalization of more complex or simplified neuron models.

The Hodgkin-Huxley (or HH) model is described by the following equations with consideration of ion currents Na^+, K^+ and chlorine (Cl^-),

$$I_{ex} = C_m \frac{dV_m}{dt} + G_K n^4 (V_m - E_K) + G_{Na} m^3 h (V_m - E_{Na}) + G_L (V_m - E_L) \tag{2.2}$$

$$\frac{dn}{dt} = \alpha_n (1 - n) - \beta_n n = \frac{1}{\tau_n} (n_{ss} - n) \tag{2.3}$$

$$\frac{dm}{dt} = \alpha_m (1 - m) - \beta_m m = \frac{1}{\tau_m} (m_{ss} - m) \tag{2.4}$$

$$\frac{dh}{dt} = \alpha_h (1 - h) - \beta_h h = \frac{1}{\tau_h} (h_{ss} - h) \tag{2.5}$$

where V_m is the membrane potential, n, m, and h are activation coefficients (or variables defining the ratio of open channels), α_n, α_m, and α_h the frequency at which the channels open, and conversely β_n, β_m, and β_h the rate at which the channels close. Rates α and β are functions of V_m with exponential variations. Linked to these rates, n_{ss}, m_{ss}, and h_{ss} represent the stationary values of the corresponding coefficient defined as functions of V_r, while τ_n, τ_m, and τ_h are the time constant with which the system converges to the stationary values. G_{Na}, G_K, and G_L are conductances. The electrical equivalent circuit is shown in Fig. 2.3.

The HH model provides a well-defined system for the study of spike generation. For higher biomimicry, it is possible to extend this model. For example, Wei [10] proposed a model that introduces biological effects not taken into account in the HH model for better analysis of pathological phenomena of seizure and spreading depression. However, we do not take interest in making the HH model more complex but rather in simplifying it.

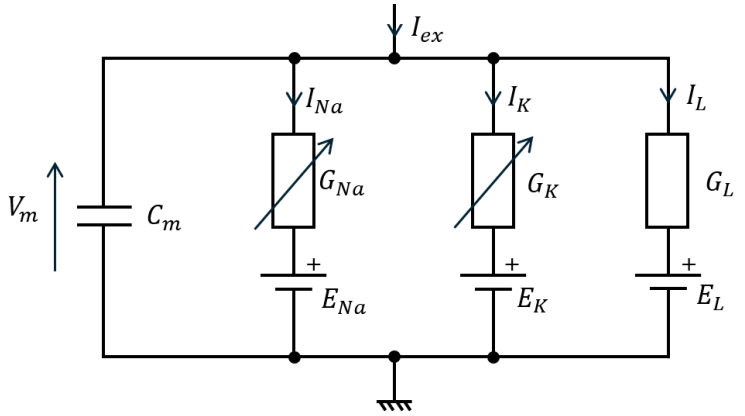


Fig. 2.3 Electrical equivalent circuit of the HH model.

Fitzhugh-Nagumo Model

In 1960 and 1961, Fitzhugh proposed two approximations [11], [12]. Firstly, the time constant τ_m of Na+ channels are disregarded since they open and close at a very fast pace ($\tau_m < 1\,ms$), such that m is always close to its stationary value m_{ss} and we can note $m = m_{ss}$ to remove one differential equation. Secondly, it is observed that the variables n and h are dependent such that $n + h = 0.85$, enabling the suppression of one these variables. In short, these simplifications lead to the new set of equations,

$$I_{ex} = C_m \frac{dV_m}{dt} + G_K n^4 (V_m - E_K) + G_{Na} m_{ss}^3 (0.85 - n)(V_m - E_{Na}) + G_L (V_m - E_L) \quad (2.6)$$

$$\frac{dn}{dt} = \alpha_n (1 - n) - \beta_n n = \frac{1}{\tau_n} (n_{ss} - n) \quad (2.7)$$

Having now two differential equations and two variables, a graphic mathematical technique can be used for a coarse study of the non-linear parameters of the approximated model, namely phase plane analysis. Phase plane analysis gives a global picture of how voltage and recovery interact, essential for understanding spiking and threshold behavior in neuron models.

It supposes two state variables x and y governed by coupled differential equations $\dot{x} = f(x, y)$ and $\dot{y} = g(x, y)$. A flow field of $\dot{x}$ and $\dot{y}$ (or phase portrait) may be traced by determining the direction of small displacements for small time steps, plotted as a vector field $(\Delta x, \Delta y)^T$ with values determined by integration of the differential equations. Necessary for the construction of the phase portrait, nullclines $f(x, y) = 0$ and $g(x, y) = 0$ are traced to follow the evolution in time of the system composed by the two variables. Nullclines divide the space into two regions where the derivation of the state variables are positive or negative, enabling the establishment of a trajectory. Then, intersections of nullclines provide fixed points, stable or unstable, informing on the system dynamics.

In the context of neuron membrane models, it provides a graphical tool for analysis and comparison of models. By looking at where trajectories go near a fixed point, it is possible to tell if the membrane returns to rest (stable) or the membrane departs, like firing a spike (unstable). Trajectory shape and behavior also inform on how the neuron spikes, how it recovers, and whether it oscillates (limit cycle) or settles quietly at rest. Without solving the differential equations, it may predict when the system will fire a spike, how changing

parameters (that are external current or time constants) shifts the dynamics, and also predict the thresholds and excitability properties.

In the end, these approximations provided a great simplification of the HH model and enabled an easier and computationally less expensive model. In 1962, a model combining approximation from Fitzhugh and Nagumo [13] was introduced as a system describing a Van der Pol relaxation oscillator [12], where the variables are V_m and u, a recovery variable,

$$\frac{dV_m}{dt} = I_{ex} + V_m - \frac{V_m^3}{3} - u \tag{2.8}$$

$$\tau \frac{du}{dt} = V_m + a - bu \tag{2.9}$$

with τ a time constant, a is the x-intercept and $-1/b$ is the slope of the y nullcline on the (x, y) phase plane. For relation with the HH model, the latter is split into two reduced systems of variables such that x corresponds to the pair (V_m, m) and y to the pair (h, n).

Morris-Lecar Biological Model

After validating of Fitzhugh approximations by comparing both models' dynamics, other researchers were encouraged to propose modifications of the equation for adaptation to other neuron types. Indirectly derived from the HH model of ionic currents, Morris and Lecar [14] further simplified Fitzhugh's approximations. The authors approximate n_{ss} and m_{ss} as hyperbolic tangents and the variation of τ_n as a hyperbolic cosine (rewritten λ), all functions of V_m. Functions in powers of 3 and 4 are removed to be included in the conductance terms. This model originates from studies of giant barnacles' muscle cells, where Ca^+ and K^+ ions are mostly present.

The Morris-Lecar (or ML) formalism is expressed as a two non-linear differential equations system defined as follows,

$$C_m \frac{dV_m}{dt} = I_{ex} - G_{Ca} m_{ss}(V_m).(V_m - E_{Ca}) - G_K n(V_m - E_K) - G_L(V_m - E_L) \tag{2.10}$$

$$\frac{dn}{dt} = \lambda\,(V_m).(n_{ss}(V_m) - n) \tag{2.11}$$

$$m_{ss}(V_m) = \frac{1}{2}\left[1 + tanh\left(\frac{V_m - V_1}{V_2}\right)\right] \tag{2.12}$$

$$n_{ss}(V_m) = \frac{1}{2}\left[1 + tanh\left(\frac{V_m - V_3}{V_4}\right)\right] \tag{2.13}$$

$$\lambda(V_m) = \lambda_0 \cosh\left(\frac{V_m - V_3}{2V_4}\right) \tag{2.14}$$

where λ_0, V_1, V_2, V_3, V_4 are constants. V_1 and V_3 the potentials at which m_{ss} and n_{ss} equal 0.5 mV, respectively. V_2 and V_4 are the reciprocals of the slope of voltage dependence of m_{ss} and n_{ss}. λ is defined as a rate constant for opening K^+ channels.

The ML model provides a compromise between bio-plausibility and easiness of implementation. For an even higher implementation friendliness and easier dynamics analysis, mimicry of the electrophysiological principles in conductance-based models was further put aside in simplified phenomenological models.

Leaky-Integrate-and-Fire model

In its simplest form, the modelling of currents integration and spike generation is described as a single linear differential equation and threshold condition, where spikes are simple events instead of complex ionic gates dynamics. The biophysical mechanism and shape of spikes are put aside in favor of simple "Integrate-and-Fire" (IF) dynamics focusing on precise firing timing [15], [16]. In fact, reset to the rest potential is immediate after the generation of the spike and done through an algorithm (discontinuous jump when the firing threshold is crossed). They are largely used for the prediction of spike timings and construction of large neuron networks.

The most common and simple model of the IF family is the leaky IF (LIF), which is linear and apparent to a resistor R in parallel with the capacitance C_m (Fig. 2.4a). The equation of the LIF model referred as the passive membrane by neuroscientist is defined as,

$$\tau \frac{dV_m}{dt} = -(V_m - V_r) + RI_{ex} \tag{2.15}$$

with $\tau = RC_m$ a time constant, u the potential, u_r the resting potential, and I_{ex} the excitatory current. The leaky term $-(V_m - V_r)$ allows return to the rest potential when no stimuli is received.

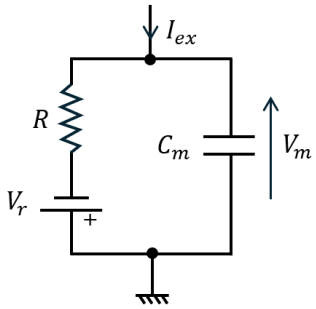


Fig. 2.4 Electrical equivalent circuit of the LIF model.

Variants of the IF neuron introduce more range of neuronal dynamics by complexifying the architecture with the introduction of exponential and adaptation variables in non-linear equations. The leaky return to rest potential is replaced by a non-linear function $f(V_m)$ for the general equation,

$$\tau \frac{dV_m}{dt} = f(V_m) + R(V_m)I_{ex}\,. \tag{2.16}$$

Among the non-linear IF, the adaptative exponential IF model (AdEx) accounts for rich spiking patterns found in biological neurons. It arose from the lack of adaptation in exponential [17], [18] or quadratic [19] IF models. Thus, with an additional adaptation variable u, the AdEx model is defined by the following two equations,

$$\tau \frac{dV_m}{dt} = -(V_m - V_r) + \Delta_T \exp\left(\frac{V_m - \theta_{rh}}{\Delta_T}\right) - Ru + RI_{ex}\,, \tag{2.17}$$

$$\tau \frac{du}{dt} = a(V_m - V_r) - u + b\tau_u \sum_{t^f} \delta(t - t^f)\,. \tag{2.18}$$

where $f(V_m) = -(V_m - V_r) + \Delta_T \exp\left(\frac{V_m - \theta_{rh}}{\Delta_T}\right)$ is issued from the exponential model with θ_{rh} indicating the firing threshold voltage for repetitive firing with a constant current injection and

Δ_T a sharpness parameter of the exponential term. a and b are the adaptation parameters that define the firing pattern of the neuron ranging from tonic to adapting and bursting spiking found in biology.

In a same approach and taking as reference the work from Fitzhugh and Nagumo, Izhikevich integrates quadratic non-linearity to propose a bio-plausible model that is able to showcase spiking and bursting dynamics similar to the AdEx [20]. It is formulated according to the following differential equation where u also depends on (2.18),

$$\tau \frac{dV_m}{dt} = (V_m - V_r)(V_m - \theta) - Ru + RI_{ex}\,, \tag{2.19}$$

where $f(V_m) = (V_m - V_r)(V_m - \theta)$ is the quadratic term with θ the voltage threshold. Similarly, a simple LIF can also be combined with an adaptation term for more range of spiking patterns. In fact, other models exist that build upon or are variants of the aforementioned models.

The AdEx or Izhikevich models, as well as the ML formalism, provide a compromise between biomimicry and easiness of implementation. A comparison table of most found neuron models is provided in Table 2.1 which summarizes characteristics reported in [21]. Nevertheless, the ML formalism is favored for implementation in the technology used in this thesis, considering its non-algorithmic equations, low dimension, and stronger biophysical plausibility as a conductance-based formalism.

Although analog design of IF-based artificial neurons is possible, this is not the direction taken in this thesis. In fact, IEMN team's experience and mastery of ML-based artificial neurons provides a well-grounded basis on which to build analog neural networks, and whose low power consumption and potential are concurrently being confirmed in health-related applications. In the end, the choice of the neuron model can be summarized as taking inspiration from the ML formalism, where further simplifications in the circuit design allows a suitable compromise between simplicity and power consumption.

Table 2.1 Comparison of the neuron models.

Model	Hodgkin-Huxley	Fitzhugh-Nagumo	**Morris-Lecar**	LIF	AdEx	Izhikevich
Dimension	4	2	**2**	1	2	2
Biophysical plausibility	Yes	No	**Yes**	No	No	No
Adaptation	Yes	No	**No**	No	Yes	Yes
Firing pattern range *	20	12	**15**	1	20	20

* According to 20 patterns identified in [22] and reports in [21].

The transition from mathematical neuron models to their physical realization leads naturally to neuromorphic technology. This approach aims to replicate neural behavior in hardware, enabling scalable and energy-efficient computation. The following section presents an overview of this paradigm with an emphasis on recent developments in hardware implementations and introduces the specific technology employed in this thesis.

2.2 Neuromorphic Technology

Building on principles of neural information processing, neuromorphic technology represents a paradigm shift from traditional computing architectures such as von Neumann or Harvard [23] commonly found in computers and microcontrollers for embedded applications respectively. Unlike these conventional systems which separate memory and processing units and rely on sequential operations on binary data (Fig. 2.5a,b), neuromorphic systems integrate memory and asynchronous processing in a highly parallel, event-driven manner. Neuromorphic architectures are essentially spiking neural networks (SNN) where the neurons and synapses are computing and memory units respectively (Fig. 2.5a). There is no separation between these units, unlike von Neumann architectures in which the central processing unit (CPU) and memory (data and program) are separated.

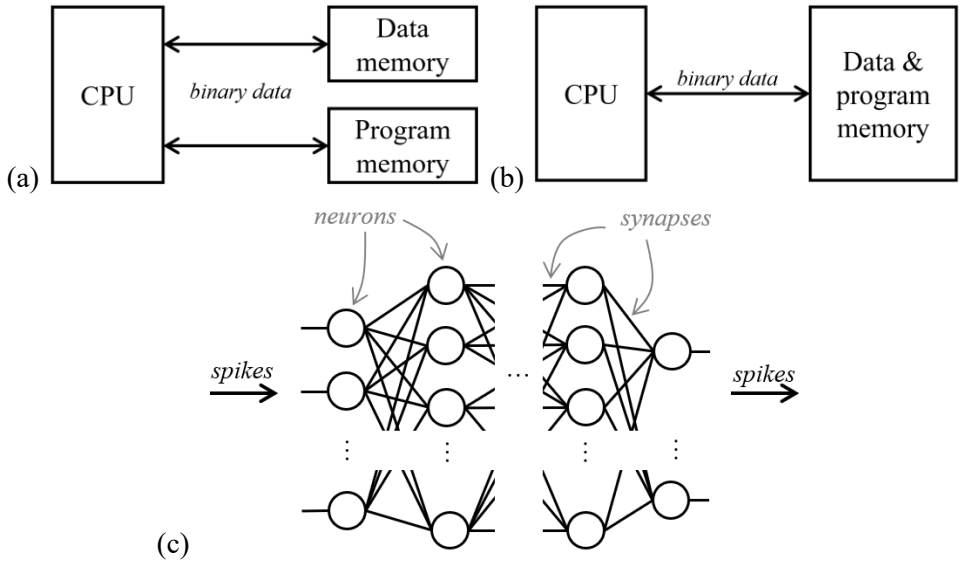


Fig. 2.5 Architectures (a) Von Neumann, (b) Harvard, and (c) neuromorphic.

SNNs revolve around the use of spikes to encode, process, and decode the information, initially introduced for the study of neural systems and the underlying learning mechanisms in the brain. As in biology, populations of neurons are connected by synapses to form neural networks. Unlike conventional artificial neural networks that only implement the concept of neuron weighted integration, SNNs thus fully reproduce the asynchronous event-based processing of biological neurons, more or less plausible from a biological standpoint depending on the complexity of the artificial neuron model.

Neuromorphic architectures are developed in digital hardware where spikes are represented as '0' and '1' binary events processed in clocked algorithms, or in analog or mixed-signal (partial digital computing) systems, as electric pulses for continuous asynchronous signal processing. Building on the more or less complex neuron models formalized over the years, diverse hardware mediums are researched today to successfully implement neuromorphic architectures with energy efficiencies surpassing those of conventional architectures or mimicking the brain in the context of artificial intelligence.

Concurrently, software frameworks were developed that enabled the simulation of complex bio-plausible SNNs or completed existing deep learning coding libraries to process spikes [24]. These tools greatly contributed to the creation of new learning rules or network architectures that are either inspired by biology or conventional deep learning. Among them, the Python library Brian2 [25] is a popular neuromorphic simulator and compiler with built-in tools for neuronal dynamics studies, which was used throughout this PhD for the emulation of the artificial neurons and synapses in software.

Overall, bioinspiration is favored in front of biomimetism in the neuromorphic computing community that target higher performances since the coupled accuracy and energy efficiency is the main objective in applications of the technology.

2.2.1 Current Advances

Given the growing interest in event-driven computing, large-scale neuromorphic architectures have been developed for the implementation of SNNs on programmable chips [24]. Among them, digital accelerators were designed for neuroscience simulation like BrainScaleS commissioned by the Human Brain Project [26], and SpiNNaker [27] which put the emphasis on understanding biology. Industries also have developed neuromorphic digital chips, including TrueNorth by IBM [28] and Loihi by Intel [29], [30], with the aim of commercializing the technology while promoting a high energy efficiency. Focusing on edge applications with low-dimension data are also the Xylo family of application-specific integrated circuit (ASIC) chips with a power consumption in the order of few hundreds of microwatts for their Xylo Audio 2 model according to [31]. The latter is described as an ULP chip as other digital chips consume above the milliwatt if not above the watt. For example, TrueNorth reports a power consumption of about 65 mW for a very large neuron capacity of 1 M neurons (256 M synapses). In comparison, SpiNNaker reaches around 30 W with 1 k neurons and 1 M synapses. Few benchmarks were made and run to accurately compare these chips different in size, CMOS process, and target use, but which gives a great idea of their full potential [32], [33].

Except TrueNorth that implements augmented IF neurons, the digital chips use LIF models to run the inference of SNNs, and in rare cases on-chip training. Working with clocked systems where spikes are efficiently reduced to binary events naturally leads to the use of IF models and prevent complex implementations of conductance-base models for example. Yet, TrueNorth indicates the ability to obtain the bio-plausible spiking dynamics described by Izhikevich in [22] by combining multiple augmented IF neurons.

To put it in a nutshell, digital chips are coming along nicely to provide mature and commercialized mediums for SNN hardware implementation. However, they do not address the actual ULP consumption required for emulating 10^{11} neurons with only 20 W like the human brain. While digital neuromorphic processors provide modular neural networks tunable in software for easier on-field testing, emerging technologies address in-depth the issue of energy efficiency. Compared to the *ULP* consideration in digital chips describing sub-milliwatt power consumptions, this term is used for sub-microwatt or sub-nanowatt orders of magnitudes in analog or novel technologies.

Since a high interest is given to energy and computation efficiency, efforts are made to bring analog computing in neuromorphic processors. Mixed-signal processors take advantage of an analog processing while keeping overall a digital routing, programming, and digital memory. For example, the processor DYNAP-SE(2) [34], [35] implements analog neurons and synapses with CMOS transistors working in the subthreshold regime. Nevertheless, digital parts of mixed-signal systems still depend on conventional supply voltage, which impacts the static consumption of the whole.

Ongoing research is being conducted to design analog synaptic and neuron arrays with memristive devices, such as memristor or resistive random-access memory (RRAM) based arrays [21], [36], [37], [38]. Memristors are a type of non-volatile memories acting as variable resistors either bistable or continuous. Application of positive or negative voltage causes ions to migrate, forming, or breaking conductive filaments or altering electron trap densities. More focused on synaptic plasticity, they are researched as analog solutions to large-scale trainable networks replacing discrete and more power consuming static random access memories (SRAM) used in complementary CMOS-based processors. Memristive devices are very promising but maturity level for commercialization is still insufficient as challenges regarding scalability, variability, and reliability must be overcome [39].

Also among emerging technologies, novel photonics components have opened the path for the creation of neuromorphic photonic accelerators, which are still in the early stages of development. Interaction between modes of a fiber are used, among other photonic properties, to produce a weighted summation of light pulses inputs as in neural networks [40], [41].

In the end, CMOS are still predominant even in emerging technologies as a mature technology to design artificial neurons, to which memristive components are combined with, or new approaches are taken. In order to reduce the consumption of transistors, a direct way is reducing the power supply in an ASIC system (bringing transistors in the subthreshold regime) which is thus confronted with variability challenges and limited processing range but enables true ULP consumptions.

2.2.2 Subthreshold CMOS Neuromorphic Technology

CMOS is a fabrication process and technology which forms the foundation of nearly all modern digital electronics. Its key principle is the combination of complementary and symmetrical pairs of p-type and n-type metal-oxide-semiconductor field-effect transistors (MOSFET) to implement logic functions in integrated circuits. This pairing allows CMOS circuits to consume extremely low power, as dynamic current only flows during switching between logic states rather than in the static state (for which only leaky DC currents remain). Because of this property, CMOS circuits have become the dominant technology for constructing large-scale integrated circuits, such as microprocessors, microcontrollers, SRAM, field-programmable gate arrays (FPGA), and digital signal processors. CMOS technology also offers high noise immunity and scalability, enabling the design of circuits with millions or even billions of transistors operating reliably on a single chip. Its low power consumption makes it

especially suitable for battery-operated devices such as smartphones, laptops, and embedded systems.

Most neuromorphic technologies rely on CMOS processes to build digital or mixed-signal basic neural components and SNNs, embedded in energy efficient processors. A greater care is given to match with biological observations, and using the subthreshold operation mode of MOSFETs enables even greater energy efficiency, especially in fully analog implementations. Before the neuromorphic technology used in this thesis is introduced, MOSFETs and the two main operation modes are described. Descriptions in this section are largely according to explanations provided in [7], [42], [43], [44].

MOSFET Transistors

At the heart of CMOS technology lies the MOSFET, a type of transistor that controls current flow through an electric field. Transistors are three-terminal semiconductor devices where two terminals are used to control the current flow in the third terminal. The two main types of transistors are bipolar junction transistors, and MOSFETs, standing as the most prevalent device in electronic circuits for their simpler manufacturing process, smaller size, and higher energy efficiency.

A MOSFET has four terminals: gate, source, drain, and bulk connected to the body. The gate is separated from the semiconductor body by a thin oxide layer, allowing the gate voltage to control the formation of a conductive channel between the source and drain without direct current flow into the gate itself. MOSFETs come in two polarities: n-type (NMOS), where current flows when the gate voltage is high relative to the source, and p-type (PMOS), where current flows when the gate voltage is low (Fig. 2.6). PMOS are similar and complementary to NMOS, operating with reversed polarities and charge carriers. The operation of a MOSFET depends on the applied gate voltage relative to the source. When this gate-source voltage exceeds a certain threshold voltage V_{th}, a conductive channel forms, allowing current to flow, provided a voltage is applied between the drain and the source. This field-effect operation enables MOSFETs to act as voltage-controlled switches or amplifiers, constituting the basis of digital logic and analog circuits. The transistor's mode of operation determines its behavior and relation to the current passing through.

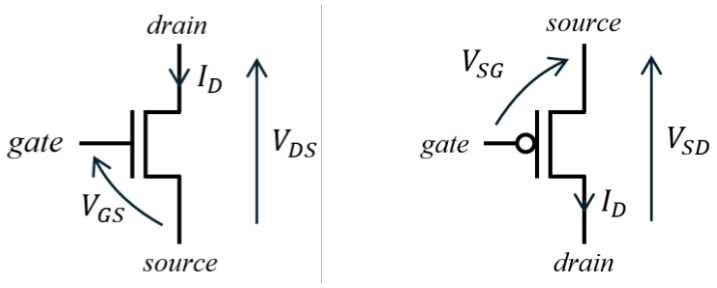
Saturated Mode of Operation

The primary mode of MOSFET operation is the saturation mode. In this mode, the MOSFET behaves like a controlled current source. For a long channel NMOS transistor, saturation occurs when the drain-source voltage V_{DS} exceeds $V_{GS} - V_{th}$, where V_{GS} is the gate-source voltage. In this regime, the channel is pinched off near the drain, and the current no longer increases significantly with further increases in V_{DS}. Instead, the drain current I_D primarily depends on V_{GS} and is given, for $V_{GS} > V_{th}$, by the quadratic relationship

$$I_D = \frac{1}{2}\mu_n C_{OX}\frac{W}{L}(V_{GS} - V_{th})^2 \tag{2.20}$$

where μ_n is the electron mobility, C_{OX} is the oxide capacitance per unit area, and $\frac{W}{L}$ is the transistor's width-to-length ratio. Whereas an NMOS turns on when the gate voltage is positive

with respect to the source (and conducts via electrons), a PMOS turns on when the gate voltage is negative with respect to the source (and conducts via holes). Then, saturation occurs for $V_{SD} > V_{SG} - |V_{th}|$, and $I_D = \frac{1}{2}\mu_n C_{OX}\frac{W}{L}(V_{SG} - |V_{th}|)^2$, where V_{th} is negative. Fig. 2.6 depicts the electrical schema of the two MOSFET types.



(a) (b)

Fig. 2.6 Schema of (a) an n-type, and (b) a p-type MOSFETs.

In digital applications, this mode corresponds to the high (or "on") state of the transistor, where it provides a low-resistance path and can drive significant current to switch logic levels. Transitions from a low (or "off") to a high state (and vice versa) occur with little transient period in the undetermined state of the transistor that is intermediary between the on/off states.

A nominal supply voltage around 1.2 V is required to saturate transistors. Such a value greatly impacts the static and dynamic power consumption of the system, especially if one targets ULP applications. Further reducing the supply voltage changes the behavior of the transistors, then falling into the subthreshold regime when $V_{GS} < V_{th}$.

Subthreshold Mode of Operation

Reducing the supply voltage below the threshold voltage to few hundreds of mV (usually inferior to V_{th}) leads transistors to operate in sub-threshold region, occurring when V_{GS} is below the threshold voltage V_{th} and the transistor is considered nominally "off". Even under these conditions, a small but non-zero leaky current flows through the device, in the order of few fA to few nA. This current arises from the diffusion of minority carriers (that are charges, like electrons and holes) across the weakly inverted channel, rather than the drift-dominated conduction seen above threshold (arising from the electrical field created by the voltage difference between the source and drain). In the subthreshold, electrons cannot fully flow from the source to the drain, so only a small current is obtained.

The subthreshold drain current has an exponential dependence on V_{GS}. Mathematically, the drain current increases exponentially as a function of the source-to-gate voltage $I_D \propto \exp\left(\frac{V_{GS}}{nV_T}\right)$, where n is the subthreshold slope factor or ideality coefficient (n~1.5 in practice). For transistors with few nanometers of gate length, and with the source tied to the body, the drain current can be approximated for NMOS and PMOS in subthreshold operation respectively as,

$$I_{D_{NMOS}} = I_n \exp\left(\frac{V_{GS}}{nV_T}\right)\left(1 - \exp\left(-\frac{V_{DS}}{V_T}\right)\right)\left(1 + \frac{V_{DS}}{V_a}\right) \tag{2.21}$$

$$I_{D_{PMOS}} = I_p \exp\left(-\frac{V_{GS}}{nV_T}\right)\left(1 - \exp\left(\frac{V_{DS}}{V_T}\right)\right)\left(1 - \frac{V_{DS}}{V_a}\right) \tag{2.22}$$

where I_n and I_p are constant currents deduced from the physical parameters of the transistors, and V_a is Early's voltage (few hundreds of millivolts) used to take into account the imperfect drain-source saturation ($V_{DS} \gg V_T$). V_a is supposed identical for NMOS and PMOS to simplify equations.

Focusing on the region $V_{DS} \gg V_T$ (but still $V_{GS} < V_{th}$) corresponding to a "saturation regime" of the subthreshold operation, I_D becomes

$$I_{D_{NMOS}} = I_n \exp\left(\frac{V_{GS}}{nV_T}\right)\left(1 + \frac{V_{DS}}{V_a}\right) \tag{2.23}$$

$$I_{D_{PMOS}} = I_p \exp\left(-\frac{V_{GS}}{nV_T}\right)\left(1 - \frac{V_{DS}}{V_a}\right). \tag{2.24}$$

If the Early effect would be negligeable, which amounts to disregarding V_a, I_D becomes independent of V_{DS}, such that (2.23) and (2.24) can then be rewritten as

$$I_{D_{NMOS}} = I_n \exp\left(\frac{V_{GS}}{nV_T}\right) \tag{2.25}$$

$$I_{D_{PMOS}} = I_p \exp\left(-\frac{V_{GS}}{nV_T}\right). \tag{2.26}$$

In conventional digital circuits, subthreshold current contributes to static power dissipation as leakage when transistors are supposed to be off. However, subthreshold operation is judiciously exploited in ULP designs, where circuits operate with minimal energy consumption. With currents of few nanoamperes and supply voltage of few hundreds of millivolts, sub-nanowatt power consumption can be attained. Applications involve medical implants or low-power internet-of-things devices, but also, as is the case in this thesis, subthreshold operation is used for the implementation of analog neuromorphic processing units.

2.3 Subthreshold Neuromorphic Elements

Transistor behavior in subthreshold mode of operation allows the non-linear and asynchronous nature of biological neurons to be reproduced, formalized through the generation of spikes in the time domain when excited with an input current.

Subthreshold operating current-mode (currents as state variables) designs were proposed that used current mirrors, based on the relations (2.25) and (2.26), to mimic the dynamics of biological neurons and synapses. Compact generalized IF circuits, and variants using differential pair integrators, are also investigated in the context of very large-scale integration circuits to build SNNs on a chip [35], [44]. However, these circuits introduce a large number of transistors in complex digital architectures with event-address representation. The resulting energy efficiency per spike falls in the order of hundreds of pJ, spikes amplitudes in the order of 1 V, and supply voltages higher than the threshold V_{th} of the transistors in order to enable current mirrors.

In that manner, the subthreshold technology we take interest into relies on a voltage-mode (voltages as state variables), where voltages have unit orders more similar to biological observations. Moreover, it is well suited for ULP designs. Based on a well-mastered 65 nm CMOS process, artificials neurons and synapses were designed, implemented on chip, and tested by the IEMN team [45].

In this section, the different neuromorphic elements developed by the team that constitute the toolbox used throughout this thesis are described.

2.3.1 Morris-Lecar Artificial Neurons

As previously mentioned, taking inspiration from the ML formalism was favored as a trade-off between bio-plausibility and easiness of implementation in the analog technology, but also for its discernible connection between the equations and the transistor physics, and the small silicon surface required, besides being well-mastered by IEMN team. In 2017, the team reported the successful implementation in subthreshold CMOS of artificial neurons approximating the ML formalism. A base and a simplified design of artificial neurons were developed, here referred to as the ML *Base* and ML *Fast* neurons respectively. Additionally, this thesis introduces a third neuron we will call the ML *Slow* neuron based on the ML *Base* with a different dimensioning for wider range of temporal dynamics in the toolbox.

Neuron ML *Base*

The base design of the ML neuron is described in [45] as a *biomimetic* model. Two feedback loops, negative and positive resulting from a pull-up and pull-down network respectively, charge and discharge the membrane capacitance for the generation of the spike form. In [45], the base model is characterized with its circuit biased according to the Nernst potentials 55 mV and -92 mV for a highest resemblance to biological membrane potential dynamics. In a bioinspired approach as opposed to biomimetism, several modifications are applied to the circuit in order to comply with ULP applications' needs.

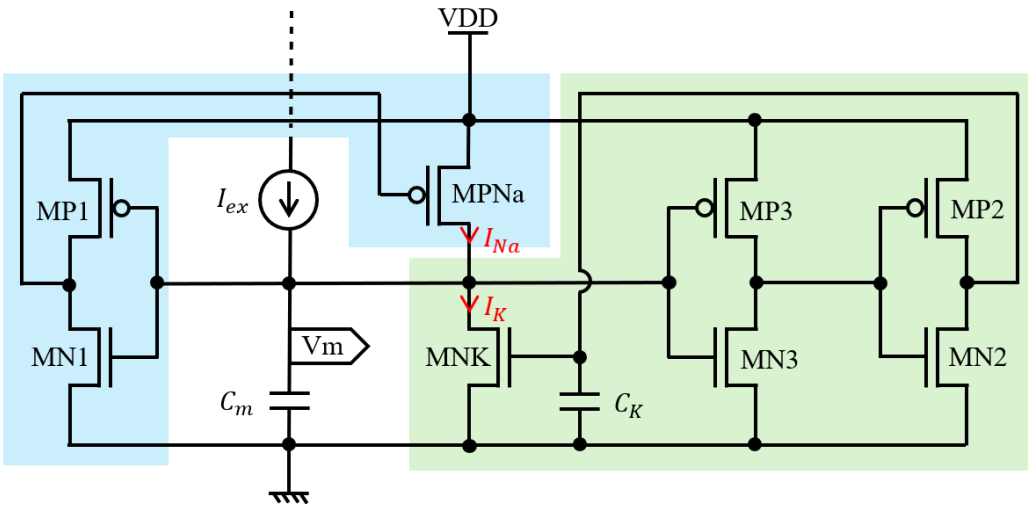


Fig. 2.7 Circuits of the ML *Base* neuron. Negative and positive feedback circuits are highlighted in green and blue respectively.

The circuit of the ML *Base* considered here is depicted in Fig. 2.7. In this thesis, the leakage conductance G_L is not used in the design of the ML *Base*. When necessary, a leak is added as a voltage controlling an NMOS transistor gate (that is, a transconductance) whose drain is connected to the neurons' membrane (at V_m). Moreover, a unique supply voltage V_{DD} is used to reduce power consumption. Thus, the neuron is connected to the ground such that the membrane potential V_m is kept between 0 V and V_{DD}. It is to be noted that all transistors are supplied with this single supply source. As a result, it warrants that all absolute source-to-gate voltage keeps lower than the threshold voltage, ensuring that all transistors operate in the subthreshold regime. Besides, with a resting membrane potential pulled-down to the low supply 0 V (source to the ground), the direct current (DC) power consumption of the neurons is greatly reduced.

According to the ML *Base* circuit, the ML formalism defined by (2.10) to (2.14) then translates in the electrical model to the following set of equations that can be coded in Python for simulation purposes,

$$C_m \frac{dV_m}{dt} = I_{ex} + I_{Na} - I_K \tag{2.27}$$

$$I_{Na} = I_{Na0} \exp\left(\frac{V_{DD} - V_{Na}}{\eta V_T}\right)\left(1 - \exp\left(-\frac{V_{DD} - V_m}{V_T}\right)\right)\left(1 + \left|\frac{V_{DD} - V_m}{V_a}\right|\right) \tag{2.28}$$

$$I_K = I_{K0} \exp\left(\frac{V_K}{\eta V_T}\right)\left(1 - \exp\left(\frac{-V_m}{V_T}\right)\right)\left(1 + \left|\frac{V_m}{V_a}\right|\right) \tag{2.29}$$

$$V_{Na} = \frac{V_{DD}}{2}\left(1 - \tanh\left(\frac{2V_m - V_{DD}}{2\eta V_T} + \frac{1}{2} log\left(\frac{G_{N1}}{G_{P1}}\right)\right)\right) \tag{2.30}$$

$$C_K \frac{dV_K}{dt} = I_{P2} - I_{N2} \tag{2.31}$$

$$I_{P2} = I_{P2_0} \exp\left(\frac{V_{DD} - V_{Na}}{\eta V_T}\right)\left(1 - \exp\left(-\frac{V_{DD} - V_K}{V_T}\right)\right)\left(1 + \left|\frac{V_{DD} - V_K}{V_a}\right|\right) \tag{2.32}$$

$$I_{N2} = I_{N2_0} \exp\left(\frac{V_{Na}}{\eta V_T}\right)\left(1 - \exp\left(\frac{-V_K}{V_T}\right)\right)\left(1 + \left|\frac{V_K}{V_a}\right|\right) \tag{2.33}$$

where I_{Na0}, I_{K0}, I_{P2_0}, and I_{N2_0} are bias currents. I_{N1}, I_{N2}, I_{P1}, I_{P2} are the drain current of the transistors MN1, MN2, MP1, MP2 in Fig. 2.7 similarly to I_{Na} and I_K. V_{Na} and V_K are the gate voltage of the transistors MPNa and MNK. G_{N1} and G_{P1} are device conductances of MN1 and MP1.

In electronics, excitation can be directly applied with a current source, which would be I_{ex} in (2.27), or using a transconductance controlled by a voltage source. Usually being a PMOS with its source connected to V_{DD}, a voltage pulled down to the ground at its gate increases the drain current flowing to the membrane of the neuron.

The ML *Base*, whose parameters are summarized in Table 2.2, provides a spike frequency up to about 25 kHz at $V_{DD} = 200$ mV for 78.3 fJ/spike energy efficiency. Lower V_{DD} values can be used but would make the artificial neuron more subject to variability and instability. A particular interest is given to very slow and fast spiking activity for integration of slow inputs and for fast spike encoding respectively. In that manner, the ML *Base* as it is presented in [45] is not used in this thesis but rather its slower and faster variants.

Neuron ML *Slow*

To obtain an even slower spiking activity, the transistors and capacitances in the base circuit were dimensioned according to the values reported in the summary Table 2.2. We refer to this variant of the base neuron as ML *Slow*. This neuron is more energy demanding than the initial dimensioning, but still reaches energy efficiency lower than 100 fJ per spike. The ML *Slow* reaches at most a spiking frequency of about 1.5 kHz at $V_{DD} = 300$ mV and for an appropriate dimensioning of the input transconductance. This low spiking frequency is desirable as the purpose of this neuron is to generate spikes at the lowest rate possible.

Neuron ML *Fast*

Simplifying the base circuit provides a more compact and energy efficient system. Fig. 2.8 depicts the circuit of the ML *Fast* for a given step voltage in input. By reducing the capacitance value and transistors width, shorter time constants are obtained, and together with a smaller number of components its energy consumption reaches 4 fJ per spike at $V_{DD} = 200$ mV [45]. At $V_{DD} = 300$ mV, the spiking frequency of the neuron can go up to 270 kHz approximately for a spike duration of about 1 µs, making it particularly suitable for fast processing. Its spiking frequency could even attain 1 MHz with $V_{DD} = 400$ mV according to the simulation, but this would require on-chip experiments under a probe station to determine to which extent such rate can be reached. Until then, this high value is the supposed ideal maximal spiking rate of the ML *Fast* neuron, as simulations are usually in concordance with hardware observations.

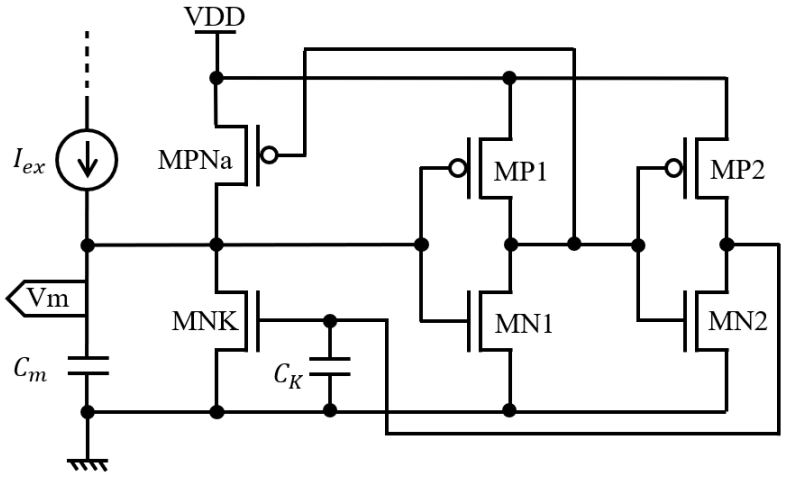


Fig. 2.8 Circuit of the ML *Fast* neuron, also called simplified ML neuron.

Table 2.2 reports the dimensions of the transistors and capacitances composing the different types of artificial ML neurons. Moreover, maximum spiking frequency and energy efficiency are summarized. While some performances are estimated from Brian2 and LTspice XVII simulations for $V_{DD} = 300$ mV, measurements available in [45] are reminded (valid for $V_{DD} = 200$ mV). The ML Slow neuron developed for this thesis has not been characterized alone in hardware but has been directly implemented in a small neural network, therefore its power consumption is unavailable and an upper bound of its energy efficiency is provided.

Table 2.2 Value of the transistor width and capacitance of the different artificial ML neurons.

ML neuron		***Base***	***Slow***	***Fast***
Transistor width (m)	MPNa	600 n	200 n	400 n
	MNK	1.83 µ	2 µ	1.2 µ
	MN1	120 n	1.2 µ	600 n
	MN2	120 n	200 n	120 n
	MN3	650 n	200 n	–
	MP1	400 n	740 n	300 n
	MP2	580 n	740 n	360 n
	MP3	120 n	740 n	–
Capacitance (F)	C_m	50 f	30.33 f	4 f
	C_K	100 f	80.73 f	8 f
Maximal spiking frequency (Hz)		25 k	1.5 k *	270 k *
Energy efficiency (J/spike)		73.3 f	< 100 f *	4 f
Power consumption (W)		94 pW	–	100 pW

* Estimated in simulation with $V_{DD} = 300$ mV (else, from [45] at $V_{DD} = 200$ mV).

Digital Buffer

Required to control synapses operation, a digital buffer (two inverters in cascade) is added to each neuron which conforms the generated spikes to square shape signals $\overline{V_{out}}$ and V_{out} (Fig. 2.9). Outputs of the inverters are defined as

$$\overline{V_{out}} = \frac{V_{DD}}{2}\left(1 - \tanh\left(\frac{1}{2}\left(\frac{2V_m - V_{DD}}{0.6\eta V_T}\right)\right)\right) \tag{2.34}$$

$$V_{out} = \frac{V_{DD}}{2}\left(1 - \tanh\left(\frac{1}{2}\left(\frac{2\overline{V_{out}} - V_{DD}}{0.6\eta V_T}\right)\right)\right). \tag{2.35}$$

The digital buffer, fed by the spikes output by the neuron, provides an isolation of the neuron's membrane potential while establishing distinct input/output paths.

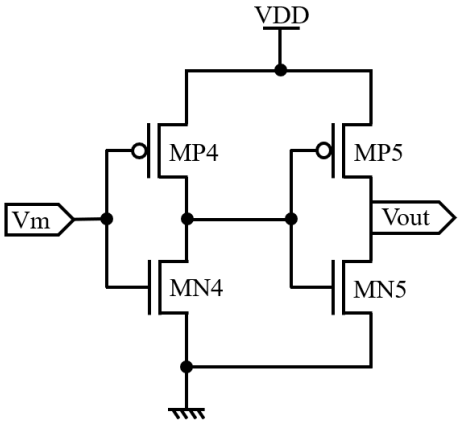


Fig. 2.9 Digital buffer circuit for spikes digitalization in output of an ML. The input voltage is the membrane potential V_m.

2.3.2 Fixed Weight Artificial Synapses

In opposition to plastic weight synapses, the fixed synapses have a constant weight in time once set, so it does not change with neuron activity. Plastic synapses have a connectivity whose weight changes according to defined rules on correlation between pre- and post-synaptic spikes. This principle originates from studies by neurologist Hebb [46]. With further observations by Bi and Poo, the spike-timing-dependent plasticity (STDP) emerged as biological rule [47] often used in neuromorphic computing for biomimetic unsupervised learning. Nowadays, learning schemes based on back-propagation of gradients for supervised or reinforcement learning issued from conventional artificial neural networks are favored for better performance. In digital systems, spikes are approximated for the computation of gradients according to a ground truth or reward/penalty signal. The computation capabilities of neuromorphic processors is not easily reproducible in analog technologies. Bioinspired learning schemes are thus preferred, like STDP, or training is performed offline then weights are hardcoded (set by voltage dividers in CMOS technology for example) for inference only by the embedded device.

For online learning, SRAMs can be used for update and retention of the weights in CMOS processes. Actually, IEMN team developed binary and ternary synapses (depending on the SRAM number of bits) based on STDP. Because these artificial plastic synapses are not used in this thesis (their integration in a SNN is still under investigation), no further detail will be provided; instead, we focus on the synapses with a fixed weight [48].

The artificial synapses are simply composed of two transistors, as depicted in Fig. 2.10. One of the two transistors controls the synapse operation (that is, active or inactive), the other transistor reflecting the weight of the synapse. These synapses, when active, propagate excitatory or inhibitory currents to post-neurons membrane capacitance. Because of the low value of V_{DD}, synapses energy consumption is neglected, with respect to neurons' consumption. The excitatory synapse operation is controlled by $\overline{V_{out}}$ while the inhibitory one is controlled by V_{out}.

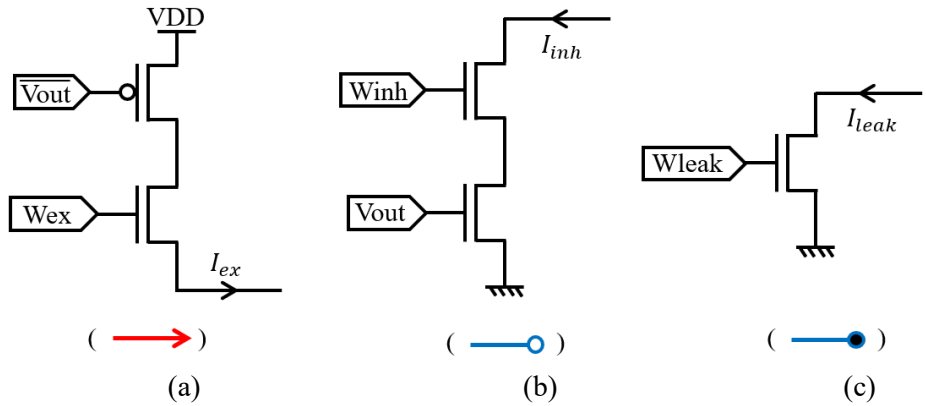


Fig. 2.10 Circuits of the artificial (a) excitatory and (b) inhibitory synapses, and (c) leak potential. Weights are W_{ex}, W_{ihn}, and W_{leak} for excitatory, inhibitory, and leak respectively. Here, V_{out} and $\overline{V_{out}}$ refer to the outputs of the neurons' digital buffer. Depolarizing or hyperpolarizing currents are then fed to a neuron's membrane V_m. In parenthesis is the representation of the different synapses in circuit diagrams throughout the thesis.

The output currents I_{ex} of excitatory synapses and I_{inh} of inhibitory synapses are defined as follows,

$$I_{ex} = I_{0_{ex}} w \exp\left(\frac{V_{out\_E\_pre} - \overline{V_{out\_E\_pre}}}{\eta V_T}\right)\left(1 - \exp\left(-\frac{V_{out\_E\_pre} - V_{m\_post}}{V_T}\right)\right) \times \left(1 + \text{abs}\left(\frac{V_{m\_post} - V_{out\_E\_pre}}{V_a}\right)\right) \tag{2.36}$$

$$I_{inh} = I_{0_{inh}} w \exp\left(\frac{V_{out\_E\_pre}}{\eta V_T}\right)\left(1 - \exp\left(\frac{-V_{m\_post}}{V_T}\right)\right)\left(1 + \text{abs}\left(\frac{V_{m\_post}}{V_a}\right)\right) \tag{2.37}$$

where $I_{0_{ex}}$ and $I_{0_{inh}}$ are bias currents. *Pre* and *Post* suffixes refer to parameters belonging to pre and post-synaptic neurons, such that V_{m_post} is the post-neuron's membrane potential, and $V_{out\_E\_pre}$ the temporally expanded output of the pre-neuron's digital buffer. The weight of the synapse is created by a transistor that controls the current according to the gate voltage level. In the equation of the resulting currents, it is simply and accurately translated to a unitless variable $w \in [0\,;1]$.

Often combined with these synapses, a so-called expander circuit can be associated (Fig. 2.11). The architecture of this circuit primarily involves a low-pass-like R-C filter and a double inverter, which together create a temporally expanded version of the input spikes. After transformation into an input current by the synapse, it allows a more lasting temporal excitation or inhibition. The duration is tunable by adjusting the weight voltage W_E, thus modifying the discharge time of the capacitance C_E. As illustrated in Fig. 2.11 with the corresponding circuit symbols, the outputs of the expander circuit may be connected to an excitatory or inhibitory synapse.

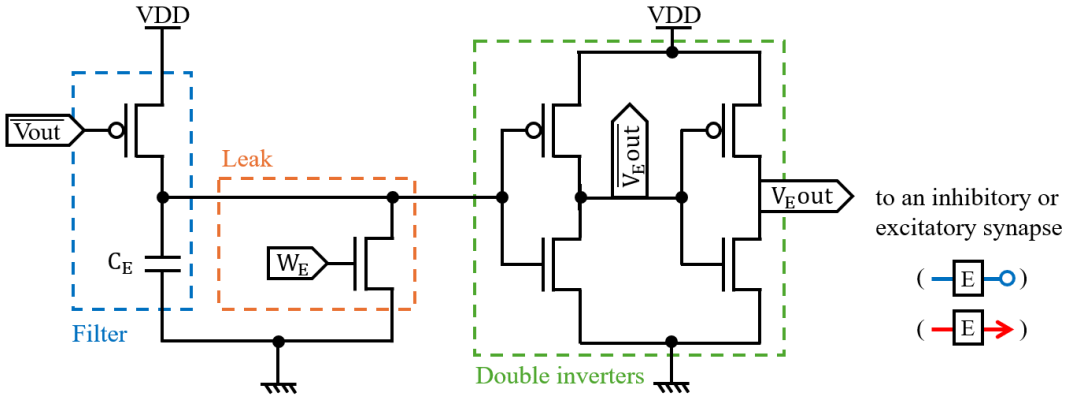


Fig. 2.11 Circuit of the expander. The outputs of the expander are connected to a synapse in the same manner as the neurons using $\overline{V_{E\ out}}$ and $V_{E\ out}$. Expander are represented as squared E.

2.4 Conclusion

From neurologists' observations of cortex activity and structure, rules and formalisms were extracted for the creation of neuromorphic technologies able to mimic neuronal dynamics.

Digital processors represent the majority of the mature neuromorphic hardware, with address-event representation of spikes in asynchronous architectures, but emergent technologies are also concurrently studied to propose more computationally and energy efficient devices. A focus is made in this thesis on the subthreshold mode of operation of MOSFETs, enabling an asynchronous and analog processing, for it is at the core of the team's research.

A resulting neuromorphic toolbox was presented including Morris-Lecar neurons and fixed weights synapses among smaller elements for adequate processing. It provides reliable neuromorphic computing units with a high level of technological maturation up to the evaluation of a prototype on the field. The ML artificial neurons generate fast or slow spiking responses with ULP power consumptions for supply voltages below 400 mV. By targeting simple neuronal architectures implementing these subthreshold neuromorphic elements, sub-nanowatt consumptions are reachable.

Central in this work, this subthreshold neuromorphic technology's potential is evaluated for biodiversity monitoring, and in particular sound source localization, for which simple neuronal circuits are targeted. Closed forms equations for artificial neurons and synapses were provided. The latter were useful to be implemented in a Brian2 environment for simulation purpose of these circuits, as described in the next chapters.

3
Neuromorphic Sound Source Localization

In the context of non-intrusive monitoring, and more specifically of position tracking, acoustic signals are judicious inputs when studying rather noisy subjects. Hearing provides non-negligeable cues for scene analysis, sometimes picking up on events that vision fails to detect or identify. Birds and insects hidden in dense vegetation, or great cetaceans swimming afar in cloudy waters, are such examples. Most of the time unseen to the eye but largely present in the soundscape of the studied environment.

Localization of living beings (or objects) using the acoustic signals they emit is more commonly referred to as sound source localization (SSL). This task of pinpointing the origin of a sound within an environment plays a major role in biodiversity monitoring. The ability to accurately determine the location of an acoustic source can significantly enhance situational awareness and tracking capabilities.

Traditional SSL methods often involve complex signal-processing techniques and extensive computational resources, which can pose challenges in low-power consumption applications. SSL is well studied with the use of mathematical methods, or conventional machine learning, and more specifically deep learning where convolutional networks are prevalent, often combined with recurrent layers [49], [50], [51]. Overall, traditional positioning or direction of arrival (DOA) estimation methods do not take inspiration from biological system but rely on powerful algorithms for increased performances in complex soundscapes. In contrast, neuromorphic computing presents a rather novel approach to address these challenges, where bioinspiration is mostly derived from the auditive system.

In this chapter, the fundamental mechanisms of auditory localization are described, with reference to the mammalian system, and an overview of the underlying neuronal processing is provided. Then, the neuromorphic SSL systems are reviewed, categorized by their focus on interaural time and/or level differences that are key binaural cues for localization. Overall bioinspiration level, methods, and performances of the reviewed systems are discussed, and in conclusion, the direction of this thesis is defined in view of the literature.

3.1 Biological Acoustic Sound Localization

The perception of sounds begins with the intricate working of the ear. In the mammalian auditory system [52], [53], the ear is divided into the outer ear (or pinna), the middle ear, and the inner ear. The pinnae and ear canal transform and direct the sound to the eardrum, which translates sound wave pressure into vibrations, while the middle and inner ear provide the necessary processing to send various types of information to the brain.

3.1.1 Binaural and Monaural Cues

The bilateral arrangement of sound collectors is a common trait of vertebrates, and some variations involving the use of other vibration sensors exist in insects, arachnids, or marine animals. The use of paired sound collectors is essential for sound localization and general perception of a soundscape [54]. The difference in waveforms received at the two collectors of a binaural pair results in major interaural cues [55]. Among them, the most used are described below.

Interaural Time Difference (ITD)

Having a binaural pair of sound collectors induces a difference in time of arrival when a sound is perceived, and greatly contributes to the localization of its source. It is maximal when the source is located at either side of the binaural pair, on the interaural axis (frontal plane), and null perpendicular to this axis (sagittal plane), so in front or back equidistant to the two collectors. The ITD perceived differs depending on the interaural distance, or baseline (distance between two sound collectors), which creates a maximum ITD. The smaller the baseline, the higher the frequency at which the ITD is non-ambiguous. In fact, when locating a source from a continuous sound, the ITD is determined by the interaural phase difference thanks to the phase-locked response of the auditory nerve fibers. The ITD can be identified at the onset of sounds knowing the first wavefront, but lateralization becomes ambiguous in continuous sounds when the period of the acoustic waves is smaller than the maximum ITD. In humans, this maximum is around 700 µs, leading to a limitation of the frequency at which the ITD is resolved at 1.5 kHz to 1.6 kHz.

Interaural Level Difference (ILD)

Jointly, the difference in sound magnitudes reliably informs on the sound source location above 3 kHz. At lower frequencies, ILDs are less impacted by head interferences or attenuation from propagation, and become barely noticeable. Magnitude differences can be expressed as a difference in decibel levels (that is, ILD), or directly as a difference in intensity, referred to as interaural intensity difference (IID), which are equivalent to a ratio and a subtraction of sound amplitudes, respectively. ILDs, or IIDs, are particularly reinforced by the pinnae's complex shape and the frequency segmentation by the cochlea, creating monaural spectral cues that play a key role in locating sounds in 3 dimensions (3-D). ITD and ILD cues are illustrated in Fig. 3.1a.

Spectral Notches

The asymmetric shape of the pinna, relative to all anatomical axes, creates multiple reflections of the direct sound. They interfere constructively or destructively in the ear canal at specific frequencies, resulting in spectral signatures that are location-dependent, as shown in Fig. 3.1b. Spectral notches are sharp drops in spectral gain due to destructive interferences. In particular, when observing the Head-Related Transfer Function (HRTF) of a human subject, a monotonical variation of the spectral content and notches between approximately 5 kHz and 9 kHz allows for the discrimination of sound elevation [56]. Spectral cues created by the pinnae have been artificially reproduced since they greatly contribute to estimating 3-D dimensions positions. In applications for robots, it is not unusual to mount microphones in artificial human heads or with artificial pinnae. Not only the pinnae, but also the shape of the head and the torso, contribute to altering the sounds in a location-dependent manner.

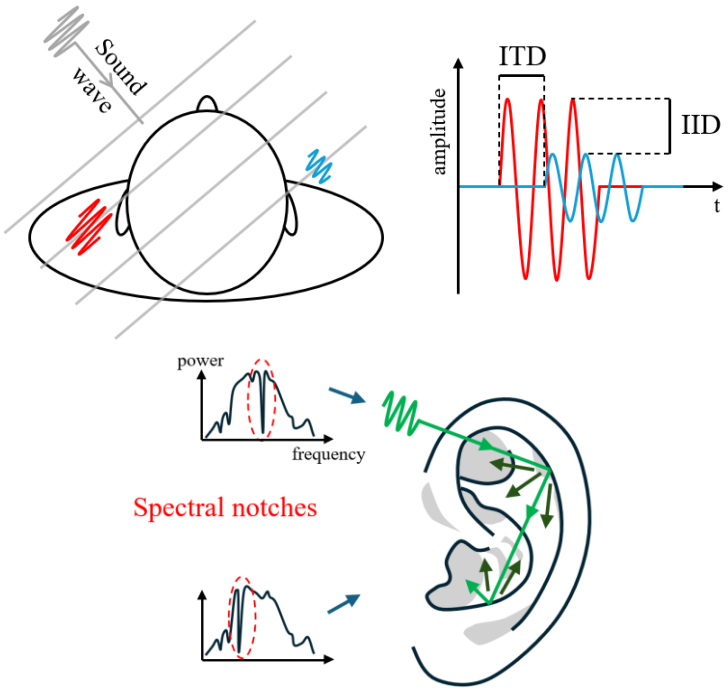


Fig. 3.1 Binaural sound localization cues. (a) ITD and IID (or ILD) cues are the difference in the time delay and amplitudes of received sounds shown in red and blue between the two ears, respectively. (b) Spectral notches are created by elevation-dependent interferences induced by the shape of the pinna, creating characteristic drops in received power in the spectrum, circled in red.

3.1.2 Cochlea

Although all information is processed using the same representation in spikes, different coding schemes are required to translate in meaningful information the various modalities available to the nervous system. Cellular and neuronal structures specific to these modalities enable neurons to process any stimulus. Just like photoreceptors are necessary to vision, and a tongue's chemical-sensitive cells are necessary to taste, neuron populations evolved to

accurately translate, transmit, and enhance information issued from sensors or other parts of the cortex. The cochlea is one of them, an organ part of the inner ear that plays a crucial role in enabling our ability to localize and interpret sounds in our environment by converting sound waves into nerve signals [53].

Composed of a canal in which pressure waves travel, it operates like a cascading filter bank with a band-pass response due to its spiral shape. The varying width and stiffness of the membrane within the canal causes its different regions to resonate at different frequencies, with a spatial arrangement following a tonotopic and logarithmic organization. Hair cells along the membrane move in synchrony with the traveling waves, creating a mechanoelectrical transduction from movement to spikes by opening/closing ionic gates. In other words, the cochlea performs a high-resolution frequency segmentation where each spectral component is encoded into spikes for further processing by the brain. This organ is not only essential to recognize sounds, but also to exploit spectral notches for sound localization.

Several cochlea models were described with mathematical formalism, such as Lyon's model [57] or Zilany's model [58], and neuromorphic artificial cochleae were introduced, such as AEREAR(2) [59], [60] and NAS [61], [62]. A reproduction of the processing in the cochlea is commonly performed following Lyon's model, which can be simplified as a bank of cascading band-pass filters (or parallel like in [63]), half-wave rectification, and suited spike encoding, as schematized in Fig. 3.2.

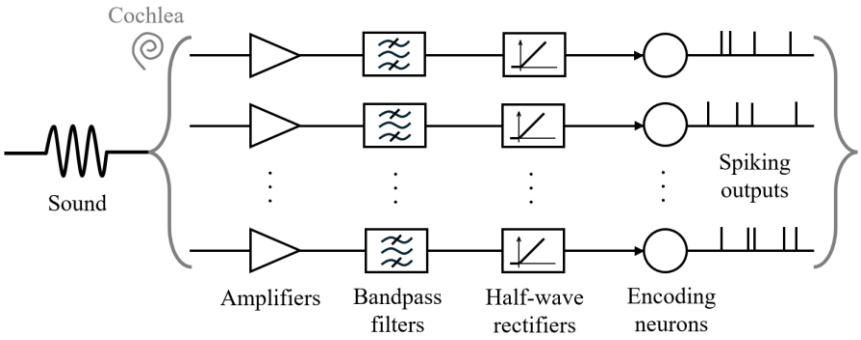


Fig. 3.2 Simplified processing performed by artificial cochleae. Input sounds are amplified to a suitable amplitude, filtered, and half-wave rectified to be encoded by neurons into spikes.

3.1.3 Neuronal Processing

While the cochlea provides the encoding into spikes of the sounds, several neuron populations are responsible for the extraction and integration of the interaural cues for final assessment of the source's position. Although still under investigation, a detailed organization of these populations in the mammalian auditory system has been refined over the years and many bioinspired works are based on this.

The main contributors for the extraction of sound localization cues are located in the superior olivary complex neuron population. It can be decomposed into the medial superior olive (MSO) and the lateral superior olive (LSO), which both contribute to identifying the direction of arrival of sound sources at different frequency ranges [53].

The MSO extracts ITDs by means of delay-lines and coincidence detections. A simplified model of the MSO was introduced by Jeffress [64], which is represented in Fig. 3.3. Delay-lines propagating spikes issued from the two ears are connected to a population of neurons, such that only one neuron activates for a given ITD. They are considered coincidence detectors because a spike is generated only when they receive simultaneous excitation from both ears. Each coincidence detection neuron is tuned for a specific ITD.

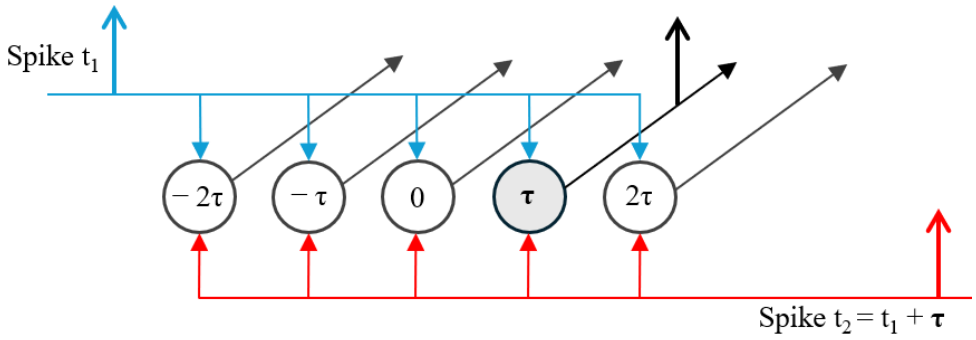


Fig. 3.3 Jeffress model of delay-lines and coincidence detection neurons. Only the neuron with the corresponding time difference τ generates a spike through temporal summation.

The LSO provides a spiking rate reflecting the IID or ILD by balancing excitation and inhibition between the two ears.

Finally, outputs from the superior olivary complex are integrated into a neuron population called inferior colliculus (IC) for sound localization with other neuronal information, like the response of neurons particularly sensitive to spectral notches or feedback from the auditory cortex.

3.2 Neuromorphic Sound Source Localization Systems

The insights gained from the structure and function of the auditory system serve as a foundation for examining neuromorphic implementations of SSL. In this section, a compact review of SSL neuromorphic spiking or pulsed (impulse-like) systems is made, from the earliest to the last published works and categorized by the main binaural cues used for the position or sound DOA estimation, namely ITD and/or ILD. Here, the most relevant features are extracted from the neuromorphic SSL solutions that highlight the trends, state-of-the-art performances, and current problematics encountered in the literature.

3.2.1 ITD Only

Being a prominent cue in binaural hearing, ITD provides precise localization cue in the horizontal plane with low-frequency sounds, making it robust in various listening environments.

With a biomimetic approach, Lazzaro and Mead [65] were the first to build a neuromorphic pulse network in the context of SSL in 1989, although no correspondence between output ITDs

and position was studied. The authors built a silicon model of the time-coding pathway of the owl based on observations in [66] reproducing the Jeffress model. Direct correspondence to an azimuth was studied in [67] with the use of a IC-inspired layer self-inhibited in a winner-takes-all (WTA) fashion [68] for ITD selection (Fig. 3.4).

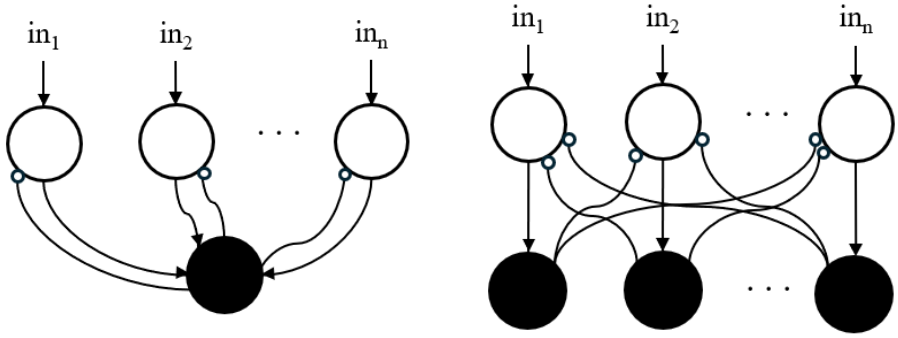


Fig. 3.4 Architecture of WTA layers. (a) Using a global inhibitory neuron, or (b) using a layer of inhibitory neurons. It is vastly used to limit spiking activity or for decision making by creating single neuron activations. Excitatory and inhibitory connections are schematized as triangular and white-filled dot arrows, respectively.

Then, Horiuchi [69] explored the use of [65]'s ITD extractor in a very-large-scale integration (VLSI) circuit based on threshold zero-crossing for application of neuromorphic technologies to robotics. The output pulses of the ITD vector extractor were combined with eye position units to create a retinotopic-auditory output vector.

In 2000, Schauer et al. [63] then further studied the ITD extractor in [65] with slight modifications for stronger WTA selection and added visual support to assess its practical application in robotics with speech processing.

Following the description made in [70] of nuclei in the auditory system of rabbits, Voutsas and Adamy [71] designed a single delay line coincidence detection model, contrasting with the Jeffress model of two-delay lines. It takes inspiration from the asymmetrical contributions of ipsilateral and contralateral sides (excitatory and inhibitory respectively) to the MSO from cochlea outputs.

Kugler et al. [72] completed the field-programmable gate array (FPGA) implementation introduced in [73], which performed simultaneous binaural SSL and classification of six sound sources with SNNs. A piecewise linear approximation implemented the hair cells processing and spike generation. The final DOA in the SSL module was estimated from the average firing rate of output neurons.

In 2010, Glackin et al. [74] presented an SNN of the MSO from observations in the cat auditory cortex, and studied it using ear canal recordings of a cat [75]. The bioplausible sound locator was proposed as a neuroscience study with supervised STDP learning. Before ITD extraction is the particular use of a bushy cell neuron layer, which converted the phase-locked burst coding in output of the cochlea to single spikes using refractory periods. The output layer is in fact the ITD extractor's neurons for which the delayed excitations from the bushy cells are weighted by the bioinspired learning rule.

The same year, Chan et al. [76] also reported a neuromorphic ITD-based binaural SSL system for azimuth estimation mounted on a wheeled robot. The AEREAR cochlea processed incoming sounds recorded by microphones embedded in a spherical head. A WTA SNN was enhanced with an additional soft-WTA layer, weakening inhibition to obtain more than one winner, and was then evaluated in simulation with supervised learning.

Unlike previous works that used a Jeffress model, Finger and Liu [77] proposed a spike-based algorithm that computed a running histogram of weighted ITDs in output of the AEREAR2 cochlea. The weights corresponded to the interspike interval preceding two paired spikes in order to highlight onsets in continuous localization.

A modified version of the biomimetic SNN incorporating bushy cells in [74] was evaluated by Wall et al. [78] for application to mobile robotics. It was separated into two symmetric SNNs to process the two lateral spaces of a ±60° angular range, trained with supervised STDP. No cochlea was used; instead, sounds were encoded into spikes using the Ben's spiker algorithm [79] from dummy head recordings.

In 2012, Chan et al. [80] also further studied their SSL system [77] in a reverberant environment and with visual feedback from a transient vision sensor for supervised online learning. After playing noise, an azimuth estimation was made, to which the robot turned. Then, flashing light emitting diodes at the sound source gave the true position with which the localization error was computed.

Focusing on hardware implementation, Park et al. [81] developed an ITD extractor in a VLSI system using AEREAR cochleae for spike encoding connected to a multiplexing neuron and IC layers. No SSL task was performed although the extractor was introduced in the context of low power-consuming hearing aids.

Until now, neuromorphic SSL systems were designed for binaural hearing and mostly using a biomimetic approach. Working on the assumption that with more sensors better accuracy can be reached, Faraji et al. [82] evaluated with 2 to 8 microphones a WTA network implemented on FPGA. Acoustic signals were encoded into spikes using a voltage comparator with adaptive threshold commanded by the spike count.

Among neuromorphic SSL systems, Beck et al. [83] proposed an 8-microphone azimuth estimator taking interesting inspiration from sand scorpions' processing of vibrations. Extended to acoustic signals, it takes advantage of the numerous sensors mounted in a circular array.

In 2018, Encke and Hemmert [84] reproduced in simulation the auditory system of gerbils for a detailed study of the influences of neuron populations and ITD sensitivity to just noticeable differences in timing across the spectrum (below 1 kHz). The SNN was able to extract small ITDs from speech by linear decoding from firing rates. A comparison of neuron populations' firing rate, called opponent-coding mechanism, was investigated using also a conventional artificial neural network (ANN) that predicted the ITD.

Without using delay-lines, Luke and McAlpine [85] performed a lateralization task (DOA is simplified in left-right estimation) in noisy conditions by directly feeding cochleae outputs from a binaural pair of microphones to an SNN.

Schoepe et al. [86] proposed a sound tracking system able to avoid obstacles thanks to visual feedback. The NAS cochlea and AER DVS128 retina chip were used for spike encoding and

SpiNNaker for neuromorphic processing. An integration subnetwork combined auditory inputs with optical information. Spike events from the optical flow provided information about close range obstacles, whereas the SSL subnetwork led the robot to the sound source. The path was then chosen through concurrent excitation and inhibition in a WTA fashion so only one direction could be considered.

In 2020, a photonic SNN for azimuth detection was implemented by Song et al. [87] for the first time, demonstrating feasibility. It consisted of photonic neurons based on excitable vertical-cavity surface-emitting lasers with embedded saturable absorbers that showed similar dynamics to LIF neurons. By analyzing the 2-D map of spiking responses in output, the side from which sounds originated could be identified.

Excellent localization accuracies were reported in Pan et al. [88], who introduced the multi-tone phase coding encoder, inspired by the Jeffress model of coincidence detection for ITD extraction. Two SNNs which were not biomimetic and closer to conventional deep learning architectures, a recurrent SNN, and a convolutional SNN were investigated for processing the temporal cues. The recurrent network performed best in all scenarios and in challenging conditions (real-world data).

In the context of bioplausible SSL systems, Zhong et al. [89] focused on the development of memristor-based oscillation neurons whose spiking activity depended on the ITD of input sounds. A simulated SNN processed the rate-based output of the oscillation neurons for azimuth estimation.

In 2023, Chen et al. [90] demonstrated the advantages of hybrid coding schemes throughout the layers of SNNs for improvements in accuracy. Different combinations of direct, rate, phase, burst, and time-to-first-spike coding were studied in convolutional SNNs separately for pattern recognition and SSL. By combining the encoder in [88] at the input layer, burst coding in the hidden layer after ITD extraction, and time-to-first-spike coding in the output layer for efficient decision making, this study reported greater accuracies than their simulation of the recurrent SNN in [88] which previously held state-of-the-art SSL precision.

Finally, Schoepe et al. [62] presented in 2024 a full hardware SSL neuromorphic system on FPGA, relying on a robotic head-tilting movement for sound source tracking. This study used their previous adaptation of time delay extraction units introduced in [91], inspired by the motion detection in vision, to SSL [92]. Artificial pinnae and NAS cochleae were mounted on a pan-tilt unit, and ITDs were extracted by time delay extraction units. No SNN processed the ITDs; instead, a motor was directly controlled by the extractor spike train outputs to orientate the robotic head toward the source.

Table 3.1 summarizes in chronological order the reviewed works with ITD-only based SSL neuromorphic systems. The numbers of neurons reported in this table (and the following tables 3.2 and 3.3) are retrieved from the literature or computed from SNN dimensions. The reported performances in these tables do not take into account multisource scenarios. Unless specified, the reported average performance corresponds to the angle precision.

Table 3.1 ITD-only neuromorphic SSL systems.

Ref.	# Mics	# Neurons	Learning	SNN Implementation	Angular Resolution / Accuracy ±Tolerance	Test Environment	Sound Source	Average Performances
[65]	2	10,540	–	VLSI	–	–	–	–
[67] [1]	2	–	–	VLSI	–	Quiet	Broadband sounds	–
[63] [1]	2	2308	–	Simulation	~2.8°	Quiet Reverberant	Speech, Pink noise	–
[71]	2	25,856	Evolutionary algorithm	Simulation	30°	Quiet	Pure tones	59% ACC
					90°		SAM signals	90% ACC
[72]	2	210	–	FPGA	30°	Quiet	FM noise	98.5% ACC
							Alarm bell	61.4% ACC
[74]	2	1029	STDP (S)	FPGA	±5°	Quiet	Pure tones	78.64% ACC
					±10°			91.82% ACC
[76]	2	–	Gradient descent (S)	Simulation	3°	Quiet	Pure tones	6.05° RMSE
							Noise	4.1° RMSE
[77]	2	1024	–	Simulation	~2.6°	Noisy Reverberant	Speech	0.17° MAE
[78]	2	10	STDP (S)	Simulation	20°	Quiet	Pure tones	3.4° MAE
[80] [1]	2	~3000	Gradient descent (S)	Simulation	3°	Reverberant	Pink noise	5° RMSE
							White noise	4.4° RMSE
[81]	2	51,752	–	VLSI	0.9°	Quiet	Narrow band sound pulses	–
[82]	2	>890	State machine	FPGA	0.32°	Quiet	Speech	3.49° MAE
						Noisy		5.57° MAE
	4					Quiet		0.99° MAE
						Noisy		1.18° MAE
[83] [2]	8	>10	–	Simulation	–	Quiet	Pure tones	4.05° MAE ±3.01° SD
[84]	2	–	–	Simulation	–	–	Speech	–
[85]	2	855	Surrogate gradients (S)	Simulation	90°	Noisy	FM noise	87% ACC
[86] [1,2]	2	1122	–	FPGA	–	Quiet	FM sound	89% correlation
[87]	2	4	–	Photonic	–	Quiet	Pulses	–
[88]	4	1981	Surrogate gradients (S)	Simulation	1°	Quiet	Speech, Noise	1.02° MAE
						Low noise		1.07° MAE
						High noise		10.75° MAE
[89]	2	>36	Backpropagation (S)	Simulation	15°	Quiet	Pulses	96% ACC
[90]	4	4261	Backpropagation (S)	Simulation	2.5°	Noisy	Speech, Broadband, sounds	0.6° MAE 95.61% ACC
[62] [1,2]	2	326	–	FPGA	5°	Quiet	Pure tones,	1.92° MAE
							Speech	5.5° MAE

ACC–Accuracy; MAE–Mean Absolute Error; RMSE–Root Mean Square Error; SD–Standard Deviation; S–Supervised; FM–Frequency Modulated; STDP–Spike Timing-Dependent Plasticity.

[1] Use vision. [2] Use body movement.

Around 2015, the focus of research shifted from being centered around biomimetic systems, that would bear resemblance to the first work in this field [65], to investigating lighter bioinspiration with the aim of improving localization accuracy. More than two microphones [82], [83], [88], [90] were used, architectures closer to ANN were studied [88], [90], and inspiration for the extraction of time differences was taken from vision [62]. Satisfying results can be observed, especially in [90], which suggests that adapting the encoding for spike generation across the layers of an SNN enhances the processing and brings out better performances. No attempt was made to estimate the distance of the sound source.

3.2.2 ILD Only

Similarly to studies conducted on ITD alone, ILD or IID have been the focus of several works, complementing ITD for comprehensive localization across the frequency spectrum.

A bioplausible SNN emulation of the LSO for IID-only sound source localization was studied by Wall et al. [93] in 2012. The SNN is composed of an inhibitory neuron population node, an LSO layer, and a receptive field layer fully connected with the output neurons. In the same manner as the previous work [78], two symmetric SNNs individually process the two hemispheres of the angular range.

A software implementation of the mammalian auditory pathways was presented by Feng and Dou [94] using only IF neurons for processing ILD cues. Here, a one-shot learning algorithm set the synaptic weights, and the output neurons in the IC layer with the highest spike rate correspond to the estimated azimuth with 10° angular resolution. Their model is completed to process both ILD and ITD in [95], with the claim that this was more faithful to the physiology organization than previous works.

In 2018, Escudero et al. [96] implemented on FPGA a rate model of the LSO for IID-based SSL with the NAS cochlea. The azimuth was estimated as a direct correspondence to the output spiking rate, and with good robustness to noise. A “spike hold and fire” block is specifically designed here to reproduce the extraction of the IID in the LSO with spike rates. A motor was controlled by the output firing rates to perform tracking of sound sources with head rotation.

A neuromorphic implementation of the LSO on TrueNorth was reported by Oess et al. [97]. Binaural inputs, recorded with a dummy head and pinnae in a sound-attenuated room, were converted to spectrograms (algorithmically) to encode temporal and spectral intensities into spike trains. The resulting spike rates were normalized and processed by an SNN. An additional 19 readout neurons for the ±90° angular range, external to the FPGA implementation, were tuned to a certain ILD value by receiving a weighted sum of the SNN’s last layer.

With the increase in digital neuromorphic accelerators, Schmid et al. [98] recently conducted investigations on a generic procedure to map a rate-based SNN to different neuromorphic FPGA hardware. A binaural SSL model inspired by the LSO was taken as a reference, originating from [97]. A three-step mapping was proposed and investigated in two neuromorphic processors. While the TrueNorth implementation was thoroughly tuned for performance, involving additional pre- and post-processing steps, on the SpiNNaker only the LSO mechanism was studied for evaluation of the generic procedure.

Table 3.2 ILD-only neuromorphic SSL systems.

Ref.	# Mics	# Neurons	Learning	SNN Implementation	Angular Resolution	Test Environment	Sound Sources	Average Performances
[93]	2	–	ReSuMe (S)	Simulation	±10°	Quiet	Pure tones	91.39% ACC
[94]	2	12,500	One-shot (S)	Simulation	10°	Quiet	Pure tones, White noise	90.56% ACC
[96]	2	>128	–	FPGA	15°	Quiet	Pure tones	3.41° MAE
						Noisy		5.63° MAE
[97]	2	1043	–	FPGA	10°	Noisy	Natural sounds	62.5% ACC
[98]	2	1024	–	FPGA (TrueNorth)	10°	Quiet	Natural sounds	18° MAE
				FPGA (SpiNNaker)				29° MAE

ACC–Accuracy; MAE–Mean Absolute Error; S–Supervised.

Table 3.2 summarizes in chronological order the reviewed works with ILD-only-based neuromorphic systems. Relying only on ILD (or IID) does not lead to results as encouraging as systems using ITD. The method for extracting intensity or level differences was biomimetic in all the works, based on the LSO, with the use of HRTF data [93], [95] or dummy heads [96], [97], [98]. Some variations in the tuning, implementation, or test sounds resulted in different performances that, in fact, lead to the conclusion that the ILD cue is not sufficient to estimate the azimuth of a sound source.

3.2.3 ITD and ILD

Finally, both ITD and ILD were used to fully take advantage of the localization information carried in incoming sound waves.

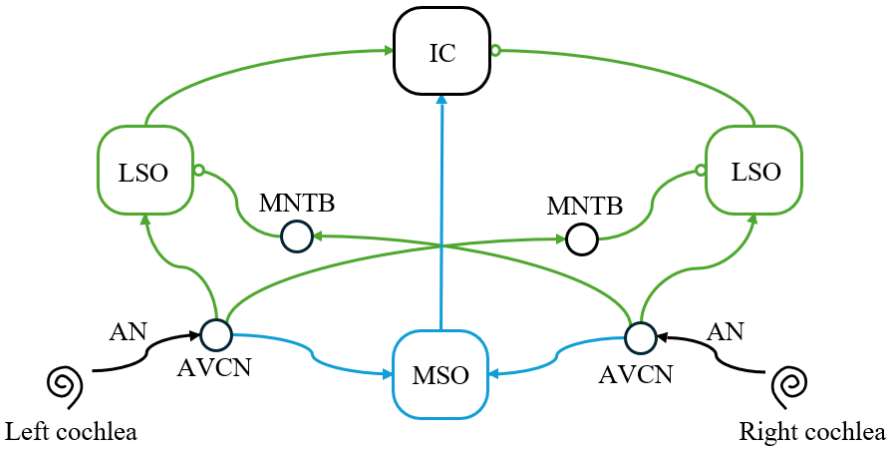


Fig. 3.5 Diagram of the mammalian auditory pathway with ITD- and ILD-specific pathways highlighted in blue and green, respectively. The populations of the neurons are represented by circles or rounded-edge rectangles. AN–auditory nerve; AVCN–anteroventral cochlear nucleus; MNTB–medial nucleus of the trapezoid body.

Unlike previously mentioned works, Liu et al. [60] used both ITD and ILD to estimate the azimuth. In a biomimetic approach, an MSO and an LSO were modeled in layers with different delayed connections to filtered sounds, and both finally integrated into an IC output layer, as shown in Fig. 3.5. All the layers were matrices encoding in all frequency channels the ITD,

ILD, and azimuth estimation, using corresponding neurons. No competition was added to the SNN, such that the authors also tested the SSL system with two simultaneous speakers. It was later further studied by Dávila-Chacón et al. [99] with a Nao robot.

The binaural spiking SSL system presented in 2010 by Goodman and Brette [100] also used ITD and ILD cues. The authors of this work used the synchrony patterns induced by location-dependent filtering for angle estimation. They also introduced a filtering of cochlea spikes by successive feed-forward neuron layers, before identifying the azimuth and elevation from overlapping neuron assemblies of coincidence detection.

In 2022, Gao et al. [101] implemented in a memristor array an analog SNN for binaural SSL processing both ITD and ILD, and supporting in-situ training. Pulses were sent in the memristor array to modify the weights according to a multi-threshold update scheme where the number of applied pulses depends on several threshold values on the weight change. Each neuron output corresponded to a vector, such that the final angle estimated was the combination of all, as depicted in Fig. 3.6.

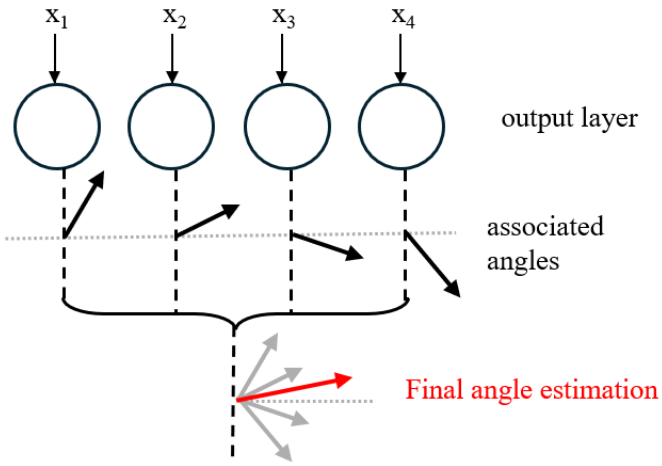


Fig. 3.6 Final output direction vector from summation of individual neuron output vectors.

Focused on neuromorphic computation, Xu et al. [102] studied in 2023 their patterning process for organic electrochemical synaptic transistor array fabrication with an SSL function simulated in a cross-grid array. Synaptic transistors of these arrays show long-term memory effects thanks to conductance modulation when suppression or excitation pulses are applied. Binaural signals were preprocessed by Fourier transform to generate 60 characteristic values, then fed to the feed-forward fully connected neural network. Similarly to [101], the output layer was a weighted combination of all vectors.

The same year, Li et al. [103] proposed a liquid state machine (LSM) [104], generically schematized in Fig. 3.7, relying on both ITD and ILD by directly providing filtered binaural sounds to the reservoir. The readout layer was then connected to an ANN classifier for azimuth estimation whose hyperparameters were chosen through Bayesian optimization and trained using soft labeling.

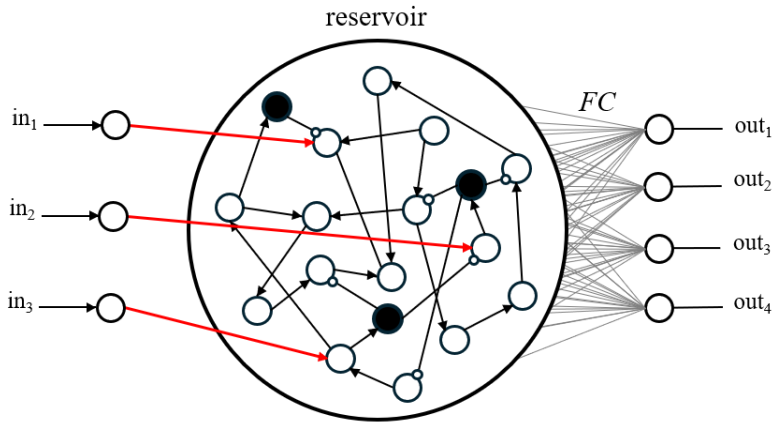


Fig. 3.7 Architecture of LSM reservoir networks, with input and output layers of 3 and 4 neurons, respectively. The reservoir contains randomly connected excitatory and inhibitory neurons to produce complex spiking patterns which can be further enhanced by adding delay-lines or connecting populations with different dynamics. Neurons chosen as inhibitory are shown as black disks. FC–Fully-Connected.

Another study that used LSM was presented by Roozbehi et al. [105] in 2024, who introduced a dynamic rescaling of the reservoir's size. Based on the small-world connection technique, the generation of new neurons allows the network to increase its computation capacity. Neurons of the LSM had spatial coordinates that represented the real space in which a sound is recorded at a distance lower than the baseline. The location of the source was estimated by the coordinates of the neuron with the highest membrane potential. The network then grew depending on the localization error to give a more precise estimation.

Table 3.3 ITD/ILD neuromorphic SSL systems.

Ref.	# Mics	# Neurons	Learning	SNN Implementation	Angular Resolution/ Accuracy ±Tolerance	Test Environment	Sound Sources	Average Performances
[60]	2	–	–	Simulation	30°	Low noise	White noise, Speech	80% ACC
[100]	2	106	–	Simulation	15°	Quiet	White noise, Speech, Instruments	2°–7° MAE (azimuth) 7°–20° MAE (elevation)
[99]	2	–	–	Simulation	15°	Quiet	White noise	2.5° MAE
							Speech	27° MAE
[101]	2	67	Backpropagation (S)	Memristor array	40°	Quiet	–	12.5° RMSE
		261		Simulation				5.7°
[102]	2	–	Backpropagation (S)	OESTs array	40°	Quiet	–	0.11 normalized MAE
[103]	2	1516	Bayesian optimization	Simulation	10°	Quiet	Speech	86.33% ACC
[105]	2	100	Modified STDP (S)	Simulation	–	Noisy	Natural sounds	69.8% ACC 3.4° MAE 0.38 m MAE [1]

ACC–Accuracy; MAE–Mean Absolute Error; S–Supervised; OESTs–Organic Electrochemical Synaptic Transistors; RMSE–Root Mean Square Error.

[1] Distance estimation average performance.

Table 3.3 resumes in chronological order the reviewed works with ITD/ILD-based neuromorphic systems. Starting with similar biomimetic systems around 2010, new approaches were then explored from 2022 onwards in order to successfully integrate both cues. Emerging technologies are also being investigated [101], [102], more in phase with the context of low-power consumption.

Surprisingly, little research was carried out using both ITD and ILD cues, but the resulting performances speak for themselves. We have yet to see a full neuromorphic system using both cues with localization performances surpassing the best precision obtained with ITD only [90]. With added information, the task should be easier, unlike what is now observed. Localizing with both cues would likely benefit from propositions of novel ILD extractors or enhanced processing of the standard LSO's output.

The question of range estimation (distance of sound sources) is not studied in any works but [105] (including the previous sections 3.2.1 and 3.2.2), which gave spatial coordinates but constrained to within the baseline of the microphone array. HRTF data and/or dummy heads used in all works but [105] provide spectral dependance to the sound source's location that might make the distance-related cues less relevant. Yet apart from [105], sound sources are always placed at a fixed distance, although IIDs vary with the distance to the microphones, which casts doubts on the results' interpretation of the commonly used LSO.

3.3 Discussion

The trends, features, and limits of existing SSL systems are identified and discussed in this section on which the conclusion builds to determine the most suitable SSL model for the subthreshold neuromorphic technology.

3.3.1 Methods and Precision Performances

Neuromorphic systems with spike representation in SSL are mostly driven by research on the biological neuronal pathways in auditory systems. Most works have reproduced the mammalian binaural auditory system. The creation of bioplausible models of SSL is usually not with the aim of enhancing performances but rather of introducing new possibilities for spike processing. Beyond the commonly reproduced MSO, LSO, and IC, other minor neuron populations have been simulated to refine the understanding of binaural cues and subtle differences related to HRTF data [84].

The popular bioinspired WTA mechanism for single output activation is rather common in neuromorphic computing where the SSL task is performed as a classification. Apart from [82], [83], [89], [105], all the works used an artificial cochlea or a filter bank in their preprocessing usually for overall integration in an SNN. However, being biomimetic often does not benefit localization precision, as was demonstrated in [88], [90]. The work carried out on SSL tasks revealed the advantages of exploring bioinspiration with conventional processing or architectures used in conventional artificial neural networks. At this moment, reproducing the brain structure and neuronal pathways does not enable the outstanding performances seen in

mammals without providing an equivalent computational capability. The number of neurons and synapses is difficult to replicate, and simply playing with this parameter is bound to have its limitations. Goodman and Brette [100] used 106 neurons for a minimum of 2° mean absolute error (MAE) in azimuth, whereas Pan et al. [88] obtained ~1° MAE with ~2 k neurons, and Chen et al. [90] reported less than 1° MAE with ~5 k neurons. In fact, the use of network architectures and supervised learning methods from conventional deep learning adapted to spike encoding allows systems to reach higher levels of precision [90].

As neuromorphic technology and its capabilities are being studied, different architectures and processing techniques naturally appear, changing the focus of research in this field. Over the years, test conditions and system topologies have become more complex and sometimes greatly differed. Some works focused on pure tones and/or common noise distributions, while others used speech and/or natural sounds for performance evaluation. Multisource scenarios are shortly investigated with two and three simultaneous sound sources in [60], [63], [71] and [94], [95] respectively. Source differentiation mostly relies on a frequency segmentation in preprocessing, and is favorized by multilabel classification output layers of SNNs-based systems, when multiple output neurons can be activated simultaneously.

An increase in precision and angular resolution of the proposed SSL solutions is observed, although mainly in ITD-only systems. The angular resolution of works using ILD remained at 10° and with precisions not yet competing with ITD-only systems. Additionally, systems that mostly relied on IID or ILD [60], [93], [95], [96], [97], [98], [99], [100], [101], [102], [103], [105] did not evaluate the evolution of the precision with the distance. Models inspired by or mimicking the LSO for the extraction of intensity-related cues are typically used, and only a one study [105] has considered varying distances in the evaluation of their system's localization performances. Further study is required to determine if precision changes with distance or if spectral cues from HRTF data provide enough information, but the former is more likely.

Similar to how mammals process their environment, localization can be assisted by or combined with multimodal information such as video [62], [67], [69], [80], [86]. Furthermore, mobile mount can enhance the systems' perception by incorporating movements, like in [62], [83], [86]. Using data from several sensors of different modalities would certainly lead to more efficient systems, with increased robustness in real-world applications by having supplementary information like vision and hearing provide.

Although the Jeffress model is the most common extraction method in ITD-based systems, other methods were employed such as direct processing in LSM [103], [105] or memristor array [101], by taking advantage of oscillatory neuron dynamics [89], or through algorithms [77], [80] and image processing [102]. Moreover, bioinspiration can also be transmodal, that is, the adaptation of the biological process to data issued from other sensors, like visual motion [62], [86] or sand scorpion vibration detection [83]. Especially, the use of a time delay extractor initially introduced in computer vision as an alternative to the Jeffress model to extract ITDs suggests that transmodality has a great potential in bioinspired approaches.

3.3.2 Hardware and Energy Efficiency

Enabled by the surge of neuromorphic processors, FPGA implementations of SNNs are growing in all fields, and this can be seen in SSL as well. The majority of hardware implementations were made with VLSI circuits in the early studies of neuromorphic SSL, which then shifted to FPGA and more specifically with TrueNorth and SpiNNaker accelerators. In digital systems, and also in simulation, authors prioritize simple IF [65], [72], [95], [103] or LIF [62], [74], [76], [78], [81], [85], [86], [94], [97], [98], [100], [103], [105] neuron models for their lighter computational cost. Few works used custom or more complex IF models [63], [74], [82], [100] and only [84] used the biomimetic HH model while memristor-based oscillatory neurons were also investigated in [89].

Moreover, memristive arrays have recently been developed and tested on a SSL task [101], [102]. These technologies will certainly continue to be investigated in SSL since extensive research is ongoing to create extremely energy-efficient memory units for synaptic weight retention with application to any SNN.

Neuromorphic solutions aim to improve the energy efficiency of localization tasks, especially among deep learning methods, by emulating the neural mechanisms of biological auditory systems while having sufficient precision to be used in real-world applications. Progress in real-world experiments remains to be achieved, but mostly, a higher focus on the actual power consumption of the proposed solutions must be made. Table 3.4 reports the power consumption and energy efficiency of the hardware implementations in the reviewed works (chronologic order).

Table 3.4 Power consumption and energy efficiency of neuromorphic SSL systems.

Ref.	Binaural cue(s)	Hardware	Total Power Consumption	Energy Efficiency
[96]	ILD	FPGA	58 mW	-
[101]	ITD, ILD	Memristor array	0.306 µJ	-
[89]	ITD	Memristor-based	-	0.9 nJ/spike
[102]	ITD, ILD	OESTs array	-	2.03 fJ/synaptic event
[62]	ITD	FPGA	972 mW	-

OESTs–Organic Electrochemical Synaptic Transistors.

The asynchronous event processing and sparsity of spike encoding allow for a reduction in power consumption, which should lead to implementations being more energy-efficient than conventional algorithmic of ANN-based methods. However, among the 33 reviewed works, only 5 recent works ([96] in 2018, [62], [89], [101], [102] between 2022 and 2023) reported energy-related metrics.

In fact, little information is available on the energy efficiency of most reviewed papers unless the hardware is the focus of the work or if it uses the full capacity of well-characterized digital processors. Especially for SNNs, which claim to offer higher energy efficiency, it would be relevant to look at the precision performances against the power consumption or computational cost for a more accurate comparison of the different approaches. The combination of these metrics is an indicator of the system's complexity, as balancing high performance and low-

power consumption remains a challenge. Here, some low and ultra-low power consumptions are reported, but superiority in power consumption is not always showcased in SNNs: unless workloads, processing, and learning methods adapted to the temporality of spikes and aligned with neuromorphic architectures are used, overall performances fall behind those of ANNs according to recent studies [106], [107]. Digital processing has the advantage of being easier to implement, but lacks compatibility with spike representation, leading to additional computational resources being spent to emulate SNNs. Therefore, unless a power consumption is given, interpretation of the results may be uncertain and unsatisfying.

Systems using emerging materials and technologies [89], [101], [102] reported energy-related performances but appeared to provide more of a proof of concept than a possible device implementation, with extensive characterization of the novel components. Further maturation would be required, also regarding variability, stability, and scalability, to consider these technologies in real-world conditions. They are tested in the context of sound localization within laboratories, and cannot ensure the same functioning outside laboratory conditions or over multiple instances of the same device yet, unlike FPGA boards which are more flexible and stable. Yet, emerging technologies hold the key to the ULP consumption, which digital hardware has difficulty reaching, by providing greater compatibility with spike computing in asynchronous and analog processing, among other characteristics.

3.4 Conclusion

Research in neuromorphic SSL was driven, first and foremost, by the motivation to validate in simulation neurological processes. Then as the spike toolbox expanded and new computational capability was unlocked, higher localization precision was obtained. Methods found in the literature are centered around the main binaural cues, ITD and ILD, typically extracted after spectral segmentation. A review of the literature, published in [Au1], showed that further research is required for the successful integration of both cues to compete with current solutions in ITD-only systems. The increasing diversity of models and approaches seems to have grown in particular in the past few years. The limit of this growth has yet to be reached, and constant progress can be observed in the design of efficient neuromorphic hardware, SNN architectures, and learning methods, which should bring SNNs closer to achieving overall higher performances in embedded applications.

Ultimately, bioinspiration holds the promise of low power consumption systems with sufficient precision for most applications. Most recent reviewed works reveal the motivation to replicate the precision performances observed in the living by limiting the biomimetism and adopting a lighter bioinspired approach. Also in terms of energy cost reduction, advances in emerging technologies and procuration of digital neuromorphic processors opens the path to application-oriented studies in hardware implementations. Nevertheless, evaluation of energy efficiency is yet to be sufficiently assessed considering its significance to the neuromorphic community. Therefore, this thesis’s work is centered on achieving ULP consumption with the mature subthreshold technology.

Unraveling the different approaches in the literature further bring to light better precision in ITD-only systems. Contrasting with common biomimetic Jeffress-based solutions of delay lines, a recent study (segmented in several works [62], [92]) of a transmodal time delay extractor revealed a successful use of a simple motion detection model in SSL. With its limited number of neurons independent of the angular resolution and simple processing, this model of coincidence detector is chosen to be studied as a precomputing layer with the IEMN team's neuromorphic toolbox for DOA estimation on which next chapter 4 focuses.

Besides, multisource scenarios are lightly studied in neuromorphic SSL and generally handled using frequency segmentation for differentiation of sounds with non-overlapping spectrums. Having a separate recognition joint to an SSL system would certainly provide additional discriminant information on multiple sounds even in the same frequency band. With that motivation in mind, this thesis evaluates a simple recognition mechanism based on characteristic temporal features of sounds. The focus is given to the implementation in the subthreshold neuromorphic technology of an ULP detector presented in chapter 5 for which perspective applications like enhanced SSL are reviewed in chapter 6.

4
Coincidence Detection for Sound Source Localization

Having analyzed the existing works in neuromorphic SSL by pointing out the trends and limits, a methodology adapted to the subthreshold CMOS neuromorphic technology is chosen. Unlike most reviewed works, the interest is given to the precomputing stage of SSL systems, namely at the extraction of binaural cues. Since ITD provides the best angular resolution and localization accuracy according to the literature review in chapter 3, an ITD extractor is investigated suitable for IEMN team's technology and sub-microwatt power consumption.

In the neuromorphic field, bioinspired ITD-based sound source localization systems were primarily focused on the modeling of the MSO by Jeffress as a well-established biomimetic solution for ITD extraction. Few works proposed bioinspired alternatives that did not involve delay lines, among which the most promising and sparse model was a transmodal time delay detector based on the Hassenstein-Reichardt detector (HRD).

Unlike the Jeffress model, the HRD can provide a constant number of neurons for encoding a wide range of temporal delays without the use of delay lines, a function complex to implement in an ULP analog system. Although the HRD model has already been intuited as a potential estimator of ITDs [91], and investigated in sound tracking with a rotating robotic head [62], [92], it was without considering the impact of ILDs and power consumption of the system. In this chapter, the HRD is adapted to be compatible with subthreshold neuromorphic technology for continuous and real-time-oriented processing. By incorporating slight modifications and preprocessing, ITDs are extracted with a sparse HRD-based model. Its localization performances are evaluated using simplified multilateration techniques with the motivation to infer its potential as a bioinspired precomputing layer in a sound DOA estimator.

4.1 HRD-Based Time Delay Estimator

While the MSO is often modelled after the Jeffress model of delay lines for coincidence detection, it implies a multitude of neurons each tuned to a specific delay (or spanning). The number of neurons has a direct influence on the angular resolution as the number determines the temporal resolution. Besides, it requires the implementation of temporal delays in neuronal processing which is not as easily performed in ULP analog hardware than digital computing.

However, coincidence detection can also be found in the visual system which was modelled with a sparse neuronal architecture by B. Hassenstein and W. Reichardt. They developed a simple motion detection model by observing neuronal responses of a beetle (*Chlorophanus viridis*, commonly called Green Weevil) to visual stimuli [108], [109] known as the "correlation-type motion detector" or HRD. It computes the direction of motion by correlating in time the changes in luminance across two neighboring photoreceptor units using simple processing and few neurons. The HRD yields a response in accordance with the speed of motion as an estimate of a time delay, and this can be translated to the acoustic domain to estimate the ITD of a binaural pair.

4.1.1 HRD Model

In [109], B. Hassenstein and W. Reichardt explores through a system-theoretical and experimental approach the neuronal processing involved in the characteristic behavior of Green Weevils –displayed as well by other insects– in response to visual movements. Stuck to a wooden stick in order to hold it in place, a beetle carried with its own strength straws chips arranged in looped paths with Y-shape intersections for observation of its movements [108]. The insect was subjected to rotating striped patterns, such that its turning tendency during optokinetic reaction was analyzed, determined by the ratio of right and left turns.

The neuronal structure was reverse-engineered by correlating controlled stimulus input with observed behavior, ending with an abstraction of the system in simple and complex models. The authors begin with a simple time-interval model to detect motion, then duplicated asymmetrically to distinguish the direction of movement (right-to-left and conversely) by encoding the order of stimuli. To account for contrast polarity (light against dark changes), they introduce separate processing channels for *ON* / *OFF* signals, allowing the system to avoid false motion signals. Ultimately, they combine these elements into a unified model and multi-pathway network that can robustly explain how insects perceive motion. A schematization of the basic time-interval model and its improved two outputs structure of motion direction detection are available in [109] (figures indicated as Abb.3 and Abb.4e). We will briefly refer to the latter as the HRD, which has two inputs and two outputs.

The key operations performed by the HRD are multiplication and low pass filtering of pulses. The exponentially decaying trace of one input pulse is sampled by the other at a certain voltage that depends on the on the time delay. At this point, only the output channel of the preferred direction generates an activity. With few neurons and a bioplausible modelling of neuronal processing, the HRD is a perfect candidate for implementation as an ULP neuromorphic sound

localization tool. Nevertheless, some modifications are required to translate this model to acoustic signals.

In the context of event-based vision, [91] proposed a spiking implementation of the time-interval model in an asynchronous neuromorphic circuit. Complementing the direction detection, an exponentially decaying current resulting from the base model's output through adaptative non-linear synapse efficacy scaling encodes into spikes motion velocity. Corresponding to the time delay between the two inputs, this information is held into the burst's interspike intervals. Later, [62], [92] introduced [91]'s spiking model of time delay extraction in the context of sound localization and head movement steering.

Two limitations were identified in the context of acoustic signals processing. (i) When providing the raw waveforms, the time to make the input neurons spike depends on the amplitude of the signal. Because the time difference is extracted by these input neurons, it implies the HRD encodes both ITD and ILD in the same output. (ii) Encoding the time-delay by filtering the output amplitude, certainly encodes the information but in temporal interspike intervals. Retrieving the duration of these intervals requires further processing or adapted coding scheme in the connected circuit that will process the spike bursts. In that manner, appropriate pre- and postprocessing are introduced which slightly modify the HRD model and are suitable for the subthreshold CMOS neuromorphic technology.

4.1.2 Proposed Adaption of the HRD Model

The proposed model of coincidence detection based on the HRD is shown in Fig. 4.1. It estimates the ITD of a binaural pair and is composed of 2 inputs and 3 outputs. Unlike [62], [92] that uses two outputs separating positive and negative time-delays (with reference to one channel), one of this adaptation outputs leads to the estimation in spike count of the ITD while the others inform from which side the sound originates. Spike count encoding was chosen as a simple scheme over rate or temporal coding since ULP counters were developed by IEMN's team and a direct reading of the output provides the estimation.

Description of the Signal Processing

Acoustic signals captured by microphones, possessing both ITD and ILD, are passed as voltages to the model. In order to process only the time delay, the ILD is removed as much as possible by applying to the input signals an abruptly high amplification and saturation at the maximal voltage V_{DD} (supply voltage). A threshold V_{sat} ensures that meaningful waves are retrieved after saturation by being the minimum voltage at which signal amplitudes trigger the detection neurons. V_{sat} is a crucial parameter chosen above the noise floor and thus depends on the signal-to-noise ratio (SNR). Before applying V_{sat}, only positive amplitudes are kept by applying half-wave rectification such that the input voltage remains within 0 V and V_{DD}. After saturation, the resulting square shaped voltages are passed to the coincidence detector.

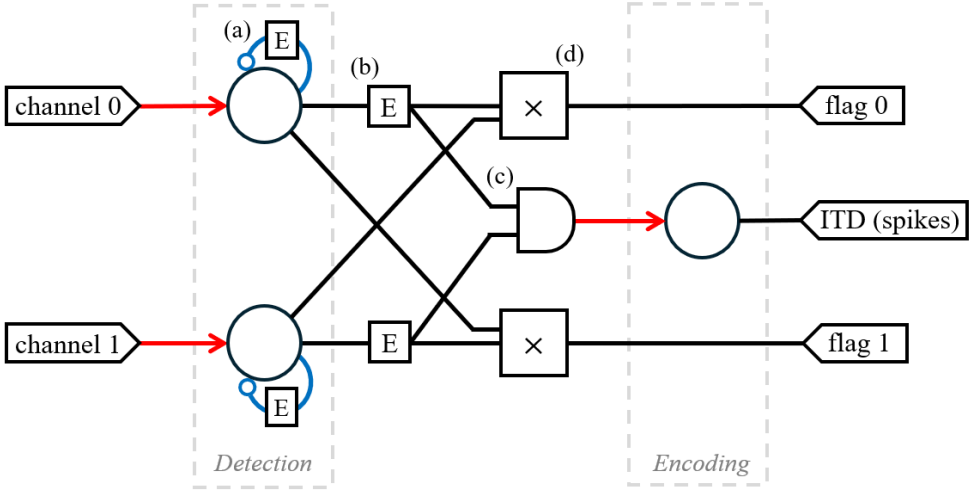


Fig. 4.1 HRD-based coincidence detector model. (a) Self-inhibition is performed with an expander for refractoriness. (b) Spikes of the detection neurons are expanded temporarily and an AND operation is performed. (d) Cross-multiplication determines the channel from which the sound is received first.

Each of the two input channels are linked to a detection neuron that will fire one spike ideally at the first up-front of the resulting square signals. The goal of the coincidence detector is to estimate the ITD at the onset of a sound. Consequently, the refractory period of the detection neurons, created by self-inhibition (Fig. 4.1a), is set to be longer than the coincidence window. At this point, the time difference between the spikes of the two channels corresponds to the ITD of the binaural pair, but further processing is needed to encode this duration into spikes.

Focusing on one detection neuron, the spike produced is expanded in time (Fig. 4.1b) using an expander, generating a square shaped signal whose width corresponds to the maximal ITD, denoted ITD_{max}, enabled by the microphone array's baseline (that is the distance between two microphones of a binaural pair) plus a chosen margin. This margin allows the discrimination between a situation with no signal and a maximal ITD.

An AND operation with a threshold ($V_{DD}/2$ for example) is applied on the expanded spikes of the two channels (Fig. 4.1c), which outputs a square signal with a width equal to $ITD_{max} - abs(\tau) + m$, with τ the ITD and m a margin. It is then encoded into spikes by one output neuron for spike count encoding of the ITD. As a consequence, it is expected to fire the maximal number of spikes when the source is located in front (on the perpendicular bisector of the segment formed by the two microphones), and a minimal number of spikes when the source is to either far side of the binaural pair. A linear correspondence gives an estimation of the ITD according to the spike count generated in output such that, for a spike count N, an approximative discrete correspondence $f(N)$ to ITD is

$$f(N) = ITD_{max}\left(1 - \frac{N - M}{N_{max} - M}\right), \tag{4.1}$$

where N_{max} is the maximal spike count obtained for $ITD = 0s$, and M the spike count expected for $\tau = ITD_{max}$.

Finally, the left or right attribution of the ITD estimation is performed by cross-multiplication of the detection spike of one channel with the expanded spike of the other (Fig. 4.1d). The

channel for which the spike is considered ahead in time outputs a spike while the other does not. It thus informs on the sign of the ITD estimated, or in other words from which side the sound is heard.

Software Implementation

The model is simulated in Python using the library Brian2 with which the neuron dynamics are emulated. The processing steps can be visualized in Fig. 4.2 that shows the intermediary and final results for an ideal square signal directly given in input of the detection neurons. In the simulation, digital and analog mathematical operators are ideal. For real acoustic signals, amplitudes above V_{sat} are abruptly amplified and saturated for extraction of the time delay, and amplitudes below V_{sat} are set to 0. They are then passed as input voltages between 0 and V_{DD} (300 mV) to the detection neurons. They generate a spike at the first square wavefront in each channel. The spikes are expanded in time, on which an AND operation is applied at 150 mV. The resulting square signal is encoded into a spike train, corresponding to the ITD estimation, by a neuron supplied with a higher V_{DD} (400 mV) to reach a spiking rate of 1 MHz in simulation. In Fig. 4.2d,e, the voltage membrane is saturated to 300 mV. A zoom on the spike train is plotted where horizontal line symbols are spike events. The side from which the sound originates is determined using cross-multiplication by a spike in the corresponding microphone's channel (signal ahead of time).

All neurons are emulated ML *Fast* artificial neurons for fast processing and output spike trains. Neuronal parameters are tuned without considering potential jitter in the emulated electrical components. The supply voltage V_{DD} at detection is set to 300 mV, while the encoding neurons are supplied with 400 mV in order to reach a spiking frequency of 1 MHz for an excitation of 300 mV. In fact, this scenario is ideal but near the bound of the MOSFETs subthreshold operation mode's supply voltage range. A lower V_{DD} could be chosen at encoding that would still allow a very high spiking frequency while complying with the limits of the subthreshold regime and suitable for a correct operation of the ML *Fast* neurons (so, 300 mV $\leq V_{DD} <$ 400 mV). For evaluation of best accuracies however, the encoding V_{DD} is set to maximum.

Knowing the spiking behavior of the emulated neurons, the minimal distance between microphones can be identified to deduce ITD_{max} of the system. The choice of baseline is constrained by the targeted binaural angular resolution α, the sampling rate f_s, the maximal spiking frequency f_{spike} of the encoding neuron, and can be mathematically expressed as

$$\frac{\Gamma}{2\alpha} \times \frac{v_{son}}{min[f_{spike}\ ,\ f_s]} \leq d \tag{4.2}$$

with d the baseline, Γ the angular range, v_{son} the sound velocity. For the simulation, the sampling rate must be considered because of the analog to digital conversion. In a completely analog system, only f_{spike} would be taken into account. For f_{spike} of 1 MHz, f_s that can reach 192 kHz, Γ of 180° and α of 1°, the baseline must be at least of 15.94 cm corresponding to $ITD_{max} =$ 469 µs. We choose a baseline of 17 cm close to the human head's width, and corresponding to $ITD_{max} =$ 500 µs.

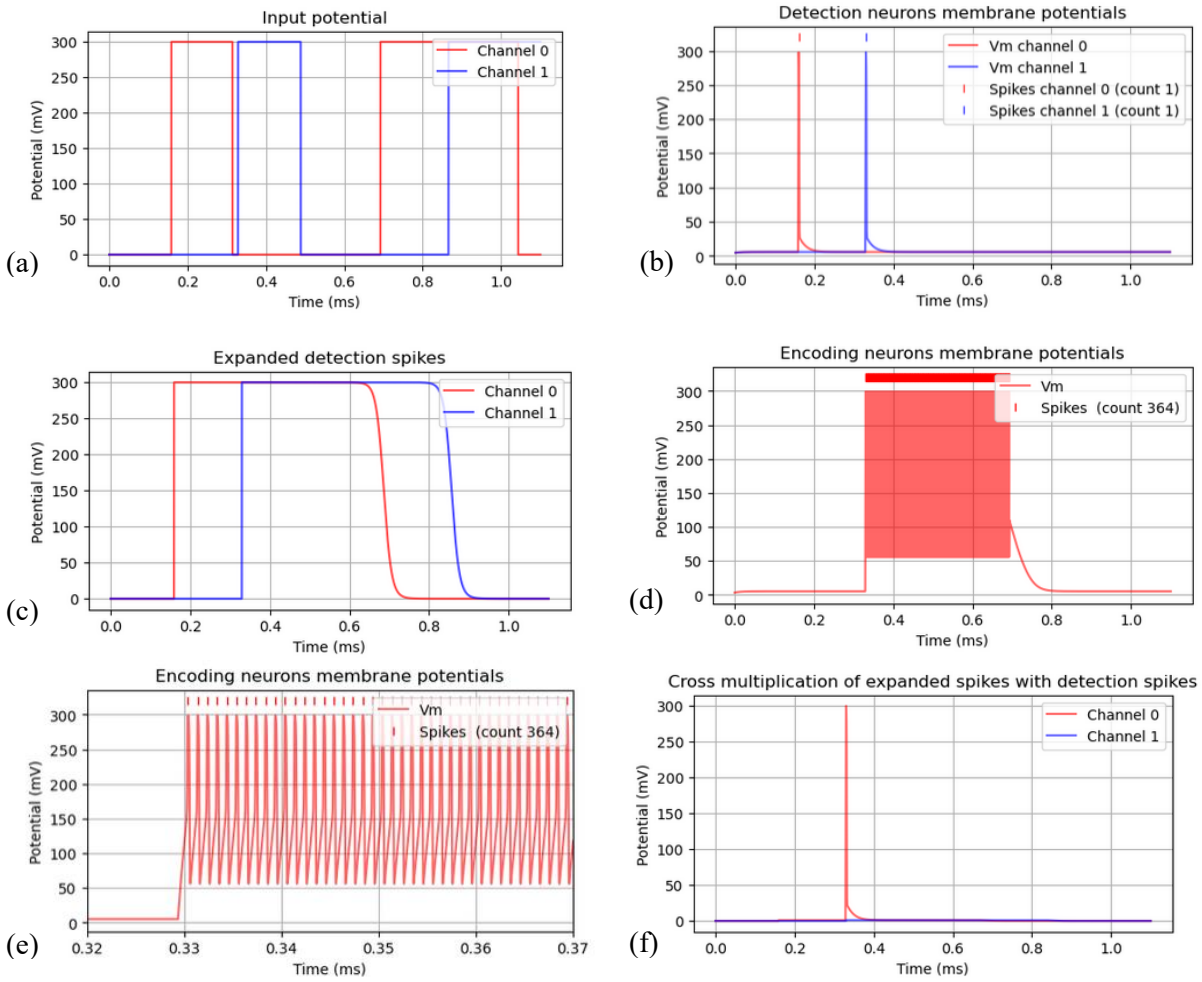


Fig. 4.2 End to end acoustic signal processing by the coincidence detector model in simulation for a binaural pair.

Concerning the range of spike counts in output, the expanders are tuned so the response to a maximal ITD provides M spikes in simulation. This value is arbitrarily chosen so the model always outputs spikes when sound is detected, and still allows the observation of the estimated positions of detections with larger time delays resulting from poor matching of wavefronts. The expander's time constant is set to 757 µs in order to obtain a square width of about 540 µs at 150 mV ($V_{DD}/2$), the value of the AND operation threshold voltage. It thus encompasses ITD_{max} of 500 µs and arbitrarily chosen margin of around 35 µs for the simulations. At V_{DD} = 400 mV in simulation, an excitation of the ML *Fast* with amplitude 300 mV lasting 35 µs and 500 µs results in 34 and 500 spikes. Using (4.2), the known ITD_{max} leads to a mean angular resolution α of approximatively 1° with f_s = 192 kHz.

The signal processing and essential parameters now set up, the simulated coincidence detector is able to compute ITD estimations of input signals recorded by a microphone array with binaural pairs of baseline 17 cm. In order to make the extractor comparable to state-of-the-art systems, spatial location estimations must be obtained from extracted ITDs for assessment of localization performances. Because tools at the precomputing level are

researched, a computational inexpensive method is chosen among standard solutions that do not involve neural networks.

4.1.3 Simplified Multilateration from ITD-Pairs for Position Estimation

A most straightforward way to evaluate the accuracy of the coincidence detector for ITD extraction is to use a mathematical resolution of the source position. The multilateration technique employed in this work is based on hyperbolic intersection which provides a well-defined analytical resolution [110]. Most papers using this method considered a very general case of microphone placement, but with an rectangular placement of the microphones (or orthogonal), a factorization in the initial set of equations is possible. Then follows a simplification of the overall analytical resolution of the source position. It is of particular interest to this work since it reduces the computational complexity and, therefore, the computational cost if the multilateration technique was to be implemented together with this adaptation of the HRD model in an embedded application. After reviewing the literature on these multilateration techniques, the simplification was never seen before.

Using multilateration based on hyperbolic intersection, the location of sound sources is estimated for arrays of 3 and 4 microphones in a rectangular configuration. The equations are detailed for this particular microphone array configuration that enables a facilitation of the analytical resolution in 2-D. Resolution in 3-D space is available in Appendix A.1. Noise is not taken into account in these equations to observe the full impact of the detector's estimation error on localization performances.

Binaural Microphone Pair

We consider foremost one binaural pair of microphones (Fig. 4.3a). Arrays with more than two microphones being in fact multiple binaural pairs combined, it provides the basis to solve the equations with more microphones. Two microphones, E_1 at $(x_0 + c, y_0)$ and E_2 at $(x_0 - c, y_0)$, compose the array with its baseline centered around (x_0, y_0). The source location is supposed at unknown 2-D coordinates (x, y). Having an ITD between the pair resulting from the sound source position from the array, the following relation can be written,

$$d_1^2 = (x - x_0 + c)^2 + (y - y_0)^2 \tag{4.2}$$

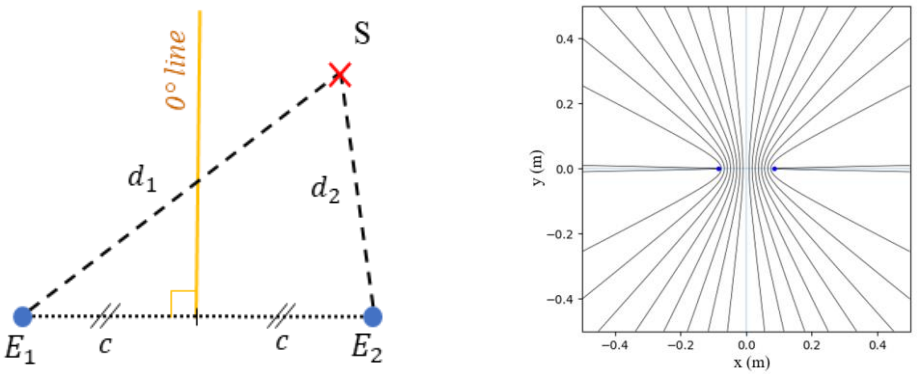
$$d_2^2 = (x - x_0 - c)^2 + (y - y_0)^2 \tag{4.3}$$

where distances of E_1, E_2 from the source S are d_1, d_2.

The possible coordinates solution of this equation is a hyperbola defined by

$$\frac{(x - x_0)^2}{a^2} - \frac{(y - y_0)^2}{b^2} = 1 \tag{4.4}$$

where (x_0, y_0) is the center of the baseline, $a = \frac{d_1 - d_2}{2} = \frac{\tau\, v_{son}}{2}$, τ the ITD, v_{son} the speed of sound, and $b = \sqrt{a^2 - c^2}$ the semi-axes of the hyperbola. Fig. 4.3b shows the hyperbolas resulting from different ITDs for a baseline of 17 cm. Confusion between right and left can easily be solved knowing the sign of the ITD, but confusion between front and back would require additional cues with isotropic microphones (omni-directional reception).



(a) (b)

Fig. 4.3 Position estimation with a binaural pair on the 2-D plane. (a) Binaural array. The microphones E_1 and E_2 form a binaural pair centered on (x_0, y_0) with baseline $2c$. S is the sound source. The yellow line indicates the 0° angle considered for each array. (b) Solution set for a 17 cm baseline and ten ITDs varying linearly between 1 µs and 499 µs. Microphones are marked by blue dots.

Using only the ITD, no exact position can be determined from a single binaural pair. With multiple synchronized binaural pairs recording the same sound source, however, the source position is the intersection of all hyperbolas. By manipulating the relations deduced from a binaural pair, we find the coordinates solution of the resulting algebraic system with 3 and 4 microphones. Using an arbitrary array increases the calculation cost, which can be minimized by the arrangement proposed below.

3-Microphone Array

In the 3-microphone rectangular configuration shown in Fig. 4.4, the reference microphone is marked as M, and the microphones to the right and top of the reference, E_1 and E_2 respectively. The two binaural pairs (M, E_1) and (M, E_2) have respective baselines D_1 and D_2. Distance of M from the source S is d_M. The estimated ITDs divided by the speed of sound of binaural pairs (M, E_1), (M, E_2) are δ_1, δ_2. We note that $\delta' = \delta_1 - \delta_2$.

The microphone array needs to be in a normalized configuration to reveal simplifications, in other words microphone M is at coordinates $(0, 0)$, E_1 at $(0, D_1)$, and E_2 at $(D_2, 0)$, the two baselines being perpendicular. Knowing the position of the microphones leads to the expressions of δ_1^2, δ_2^2, and δ_A^2,

$$\begin{aligned} d_M^2 &= x^2 + y^2 \\ d_1^2 &= (x - D_1)^2 + y^2 \\ d_2^2 &= x^2 + (y - D_2)^2 \end{aligned} \tag{4.5}$$

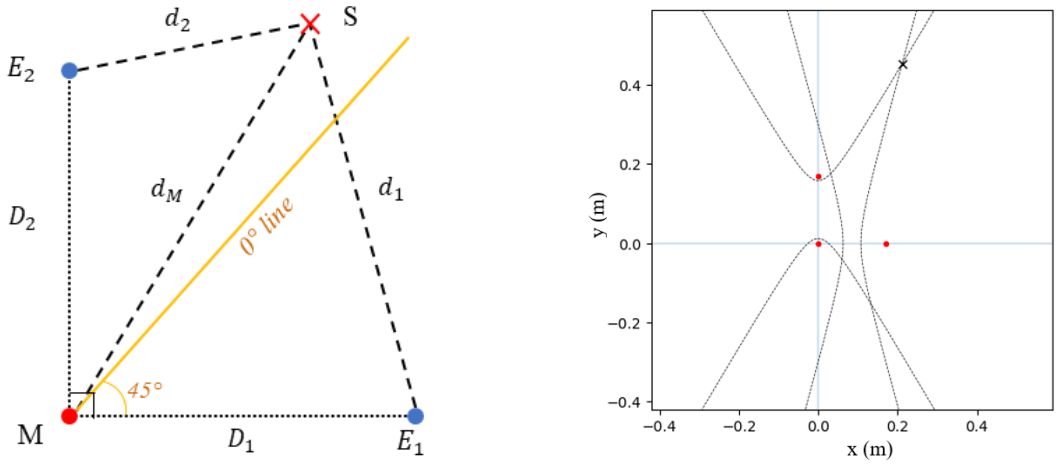
with which we can write,

$$D_1(D_1 - 2x) = \delta_1(d_1 + d_M) \tag{4.6}$$

$$D_2(D_2 - 2y) = \delta_2(d_2 + d_M) \tag{4.7}$$

By writing $\delta_2(4.6) - \delta_1(4.7)$, a relation between the source coordinates, the baselines, and the ITDs is obtained,

$$\delta_2 D_1(D_1 - 2x) - \delta_1 D_2(D_2 - 2y) = \delta_2\delta_1(\delta_1 - \delta_2) = \delta_1\delta_2\delta_M \tag{4.8}$$



(a) (b)

Fig. 4.4 Position estimation with a 3-microphone rectangular array on the 2-D plane. (a) A 3-microphone array in a rectangular configuration. Microphone M is the reference for azimuth and distance evaluation, (M, E_1) and (M, E_2) are binaural pairs with baselines D_1 and D_2. (b) Estimated position for $D_1 = D_2 = 17$ cm with ITDs 128 μs and 420 μs, corresponding to azimuth 20° and distance 50 cm from M. Resulting position and source are marked by a cross and microphones by blue dots. Hyperbolas are plotted in solid black curves.

that can be reorganized to express x as a function of y and the remainder,

$$x = \frac{\delta_1 D_2}{\delta_2 D_1} y - \frac{\delta_1 D_2^2}{2\delta_2 D_1} - \frac{\delta_1 \delta'}{2D_1} + \frac{D_1}{2} \tag{4.9}$$

The ordinate y is found by using the hyperbolas of the two binaural pair (M, E_1) and (M, E_2),

$$\frac{\left(x - \frac{D_1}{2}\right)^2}{a_1^2} - \frac{y^2}{b_1^2} = 1 \tag{4.10}$$

$$\frac{\left(y - \frac{D_2}{2}\right)^2}{a_2^2} - \frac{x^2}{b_2^2} = 1\,, \tag{4.11}$$

leading to the characteristic expression of a 2nd order polynomial equation by multiplying (4.10) and (4.11) by $-\frac{1}{b_2^2}$ and $\frac{1}{a_1^2}$ respectively, and removing the term x^2 by subtracting the two resulting equations, such that it may be written $Ay^2 + By + C = 0$, with

$$\begin{aligned} A &= \frac{1}{b_1^2 b_2^2} - \frac{1}{a_1^2 a_2^2} \\ B &= \frac{\delta_1 D_2}{\delta_2 a_1^2 b_2^2} + \frac{D_2}{a_1^2 a_2^2} \\ C &= \frac{D_1^2}{4a_1^2 b_2^2} - \frac{D_2^2}{4a_1^2 a_2^2} - \frac{\delta_1 D_2^2}{2\delta_2 a_1^2 b_2^2} - \frac{\delta_1 \delta'}{2a_1^2 b_2^2} + \frac{1}{a_1^2} + \frac{1}{b_2^2} \end{aligned} \tag{4.12}$$

where $a_i = \frac{\tau_i v_{son}}{2}$ and $b_i^2 = a_i^2 - \left(\frac{D_i}{2}\right)^2$ the semi-axes of the hyperbola, with τ_i the time delay of the binaural pair with baseline D_i. For a positive discriminant, this equation gives two solutions. The correct ordinate y is selected knowing the sign of the time delays.

Conversion to azimuthal angle and distance is then performed considering the 0° line,

$$d_M = \sqrt{x^2 + y^2} \tag{4.13}$$

$$\theta = \theta_{max} - \text{atan2}\left(\frac{y}{x}\right) \tag{4.14}$$

with d_M the distance from the reference microphone to the source, θ the angle from the 0° azimuth of the configuration, and θ_{max} the angle at the maximum ITD equal to 45° for the rectangular configurations.

Final solutions are found by solving a 2nd order polynomial equation, instead of a quadratic in x and y as it is the case with arbitrary arrays. It computes the exact position of the source while reducing computational complexity. An example is depicted in Fig. 4.4b for an arbitrary ITD pair and baselines of 17 cm.

The simplification of the multilateration technique is scalable to the 3-D space for which the resolution is detailed in Appendix A.1.

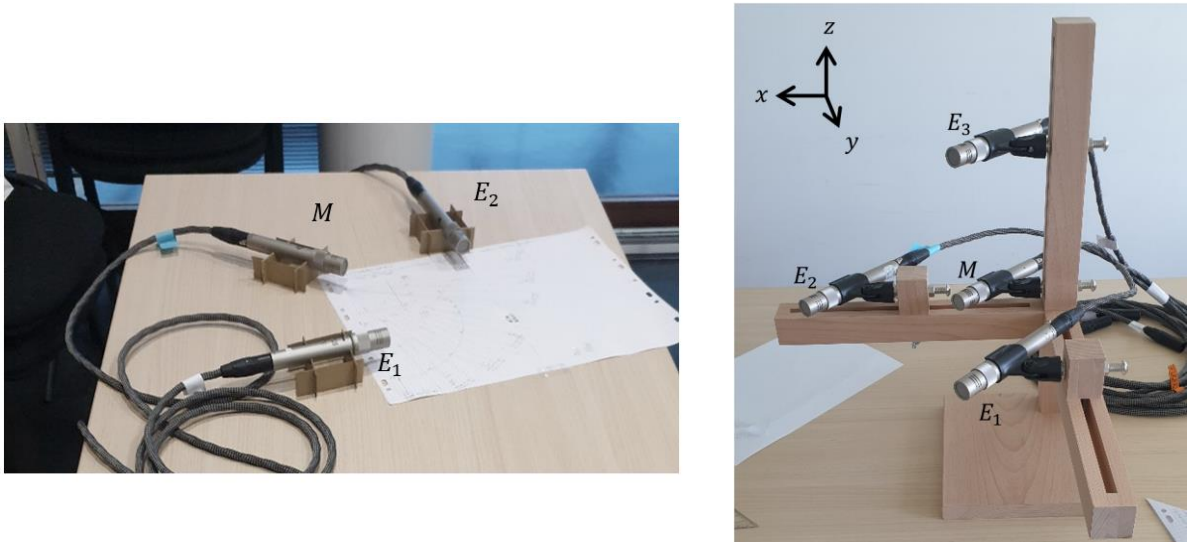
4.1.4 Indoor Recording of Sound Signals

So as to control the environment, sound types, source distances, and more importantly the microphone array configuration, no localization dataset outside of the recordings created was used.

Sounds in the present work are recorded with Shure KSM141 microphones in omnidirectional mode, with a Solid State Logic SSL12 mixing table at sampling frequency 192 kHz, and encoded in float-32 WAV files. No artificial head, torso, or pinna is used. The baseline of 17 cm is the same for all pairs of microphones. The sounds used for the evaluation of the coincidence detection model are either digitally generated or pre-recorded to be played multiple times.

A small loudspeaker with output size of 8 mm × 2 mm is placed at different positions within the angular range of the rectangular microphone arrays. Recordings are high-pass filtered at 100 Hz digitally to remove the very low frequency ambient noise, increasing all average SNR above 20 dBu. Then, actual recording is performed in a quiet meeting room with low ambient noise. Most noise are footsteps, doors opening/closing, voices, rain, and road traffic, attenuated by the room and the building. In addition, ambient noise is produced from the room's and recording equipment. This chapter does not cover multisource scenarios, thus only one sound of interest is played at a time.

For early 2-D recordings, the loudspeaker and microphones are placed on a table with an elevation of 3~4 cm as depicted on Fig. 4.5a. To produce 3-D recordings, a 4-microphone wood stand in a rectangular configuration was designed and handmade. Not available at the start of the PhD, it could not be used in all 2-D recording sets. Fig. 4.5b shows a picture of the 4-microphone stand where rails allow the microphone clips to be moved along the axis in order to adjust the baselines. Orientation of the microphones are adjustable as well in most direction thanks to the clips.



(a) (b)

Fig. 4.5 Microphone setup for recording. (a) Placement of the microphones on a table for 2-D recording of sounds emitted on the same plane (elevation is 0°). (b) Wood stand for recording in a 4-microphone rectangular configuration. In this picture, all baselines equals to 17 cm. Wooden blocks allow to orient the microphone clips in the same direction for facilitated setup of the array. The array is formed by the reception area of the microphones (access to membranes is at the front and sides of the tip), shifted in translation from the stand's orthogonal base.

Signal preprocessing is performed with Scipy (v1.11.4) for acoustic signal filtering, and standard libraries for algorithmic and mathematical operations. ML *Fast* neurons are slow to emulate in simulation with Brian2 as they require a simulation time step of 200 ns. To speed up the processing of all individual sounds in the recording set, short signal segments from the audio files are extracted using a framing threshold manually set to only detect the sounds whose maximum amplitude is distinct from the noise floor (Fig. 4.6a). At low SNR (noisy environment or distant sound source), the framing threshold cannot take into account all sounds of interest without adding noise to the simulation batch, and thus making some noise detections a part of the results. Consequently, not all individual sounds of the recording sets are detected in these scenarios. Especially, when the sound is drowned in the noise, sound and noise sources are detected. At high SNR, the framing threshold is equal to V_{sat}.

The complete process is automatic for fast sequential launching of simulation batches. Signal segments are sized according to fixed parameters, then the detection threshold V_{sat} is applied on each segment for amplification and saturation (Fig. 4.6b). Simulation are further reduced by determining a neuronal simulation window centered on the onset detection. Furthermore, if it is identified that detection spikes of each channel will result in an impossible ITD considering ITD_{max} given by the array baseline, simulations for the studied frame are passed.

The different processing steps of the simulation are resumed on the block diagram depicted in Fig. 4.7.

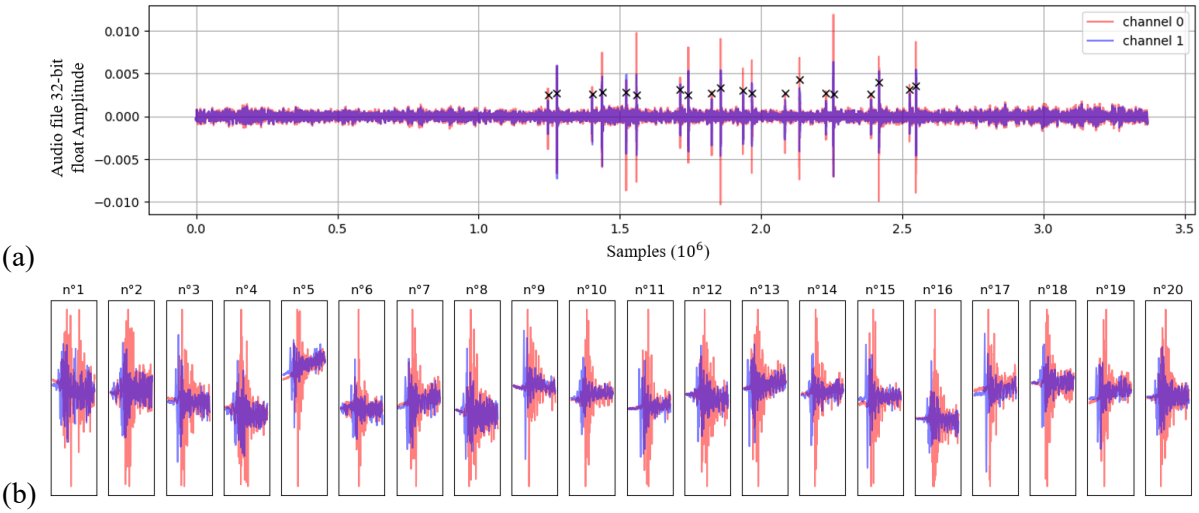


Fig. 4.6 Signal processing performed upstream of the coincidence detector. (a) Initial audio file example containing impulsive target sounds (finger snaps). The sounds of interest are automatically detected (crosses) for faster processing using an arbitrary framing threshold. (b) Each individual sound can then be extracted from a low-pass filtered version of the audio to be fed to the coincidence detector. In (b), signals are visualized without filtering.

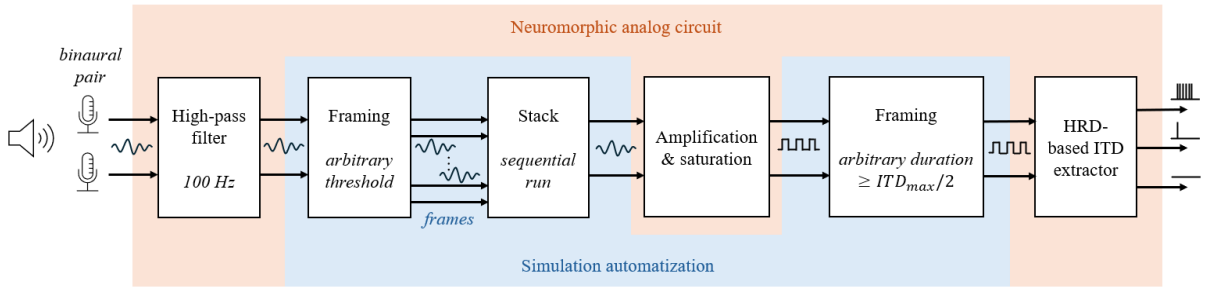


Fig. 4.7 Block diagram of the HRD-based coincidence detector. The intermediary state of the input acoustic signal is shown between the processing steps. The steps specific to the automatization of the simulation runs are in blue, while the steps of the actual neuromorphic circuit are in orange.

After the simulation of the coincidence detector for all identified frames, a csv file gathering all key information is generated for computation of the spatial position using the simplified multilateration technique and analysis of localization performances.

4.2 Localization Results

After observing the correct operation of the coincidence detector in simulation with artificial square inputs, it is studied with different sound types for assessment of its main strengths and weaknesses. For analytical purposes, the evaluation of the coincidence detector is divided into broadband impulsive sounds and tonal sounds, the latter being more sustained and narrowband in nature.

4.2.1 Impulsive Sounds

In natural soundscapes, impulsive broadband acoustic events play a distinctive role in biodiversity monitoring. Such sounds often arise from interactions between animals and their physical environment, for example, sudden movements or impacts, like the snapping of twigs. They are characterized by short duration, broadband spectral content, and a temporal profile that contrasts sharply with the continuous background noise of the habitat. These characteristics can make them valuable indicators of animal presence or activity, even when the species is not talkative and cannot be directly observed. Furthermore, it is noteworthy that some species, such as cetaceans and bat, have evolved to rely predominantly on impulsive broadband emissions (echolocation clicks) for navigation and foraging.

With this focus in mind, natural impulsive sounds are produced and recorded with the microphone array to infer more detailed localization performances. They will be concisely referred to as *clicks* or *click-like* sounds.

Recording Sets

Several sets of 2-D click recordings are referred to in this chapter whose characteristics are summarized in Table 4.1. Unless specified in parentheses, the angular ranges reported concern all distances. The angular resolution refers to the angular step between each position within the angular range. Azimuthal ranges were chosen considering the available space without rotating the microphone array; it ensures exact recording conditions between all recorded positions.

Table 4.1 Characteristics of the click-like sound recording sets.

Set	**A**	**B**	**C**
Number of microphones	3	3	3
Click type	Generated	Pre-recorded sounds	Pre-recorded sounds
Number of clicks per position	5~8	60~64	60~64
Distances* (m)	0.24, 1, 2, 3	1, 2, 3	10
Azimuthal range* (°)	±45° (24 cm) -15°→7° (3 m) -15°→-10°	-13°→14° (3 m) ±15°	-35°→38°
Angular resolution (°)	5°, 1°	5°	–

* From reference microphone M and 0° line of array.

Pre-recorded sounds were produced from hand claps, object's mechanism activations, and objects colliding. Their amplitude were then normalized to even out as much as possible the loudness of each sound. Besides pre-recorded sounds, one input sound was generated by performing a weighted sum of 3 sinusoids according to the following definition,

$$s(t) = 0.5\sin(2\pi f_0 t)\, d(t, t_0) + 0.3\sin(2\pi f_1 t)\, d(t, t_1) + 0.28\sin(2\pi f_2 t)\, d(t, t_2)\,, \tag{4.21}$$

with $s(t)$ the resulting sound, a decay $d(t, t_i) = 2^{\frac{-10t}{t_i}}$, $f_i \in \{800, 700, 600\}$ in Hz, $t_i \in \{0.1, 0.8, 0.01\}$ in seconds, and t the time vector. This sound provides a clean and simpler test sound.

A common characteristic of all recorded click-like sounds is that they are brief pulses with short rise time followed by a quick decay. An example before recording (audio file provided to the loudspeaker) of each click source type listed above is shown in Fig. 4.8.

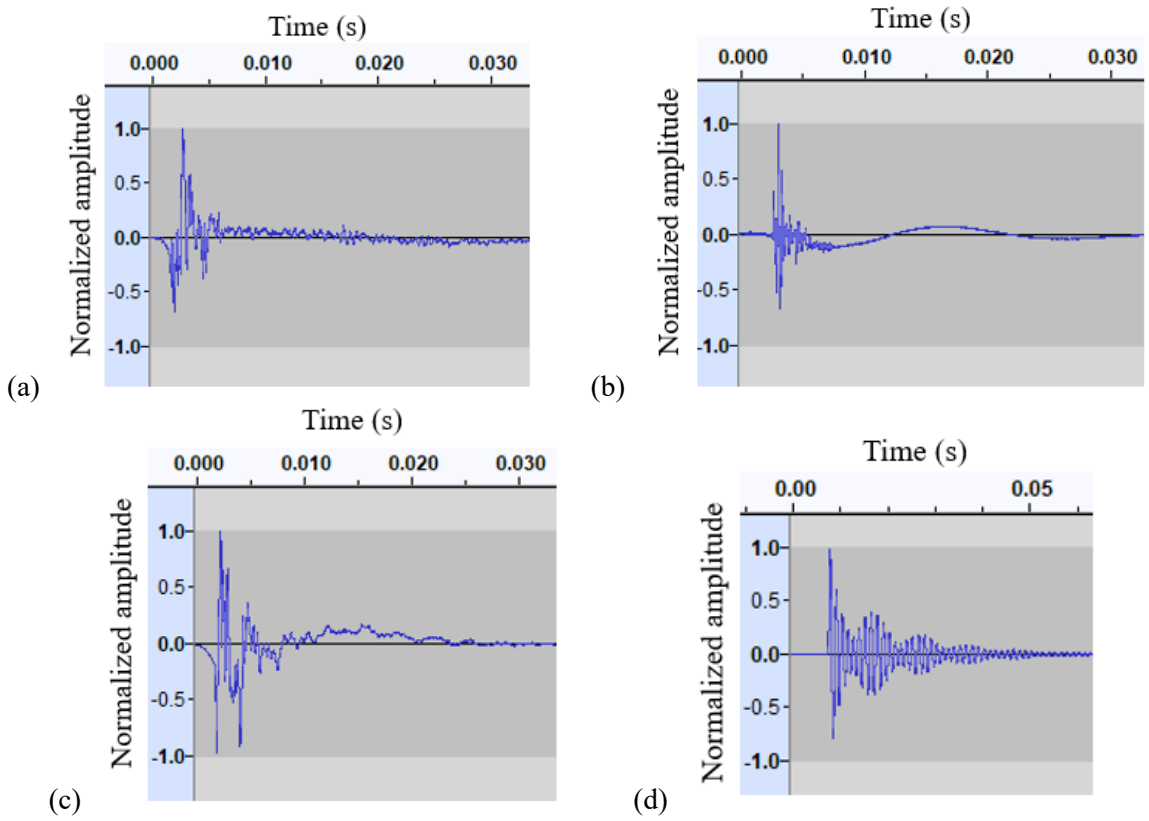


Fig. 4.8 Examples waveform of pre-recorded click-like natural and artificial sounds. Source types are (a) hand claps, (b) object's mechanism activations, (c) objects colliding. Additionally, (d) generated weighted sum of 3 sinusoids is played in loops at each position as a more controlled and low frequency waveform.

In the recording sets B and C, each click is different per position, whereas clicks in set A are the same generated click played in loop. In the case of set C, the loudspeaker was placed at arbitrary angles for evaluation at farther distances. For all sets A to C, the 3-microphone setup depicted in Fig. 4.5a was used to record the clicks.

ITD Estimation

The ITD is estimated for each recording set and both binaural pairs composing the rectangular 3-microphone array. The threshold V_{sat} is fixed at different values to observe its impact on the ITD estimation. Depending on the SNR, the ILD cannot be completely ignored such that the choice of V_{sat} may lead to poorer results due to a wrong matching of wavefronts at the detection neurons. Fig. 4.9 provides an example of such incorrect matching of wavefronts.

In the output of the model, the estimations in spike counts are compared to the ideal spike count expected for the given position using (4.1). A notable difference in precision can be observed depending on the chosen V_{sat}, where the best choice for a same distance is different

from the value minimizing the error of all recordings. The sound signals are rescaled between 0 V and V_{DD} (300 mV) after the strong amplification and saturation step, but the detection threshold V_{sat} is applied before digitally as a float value on the wav file written with 32-bit floating points.

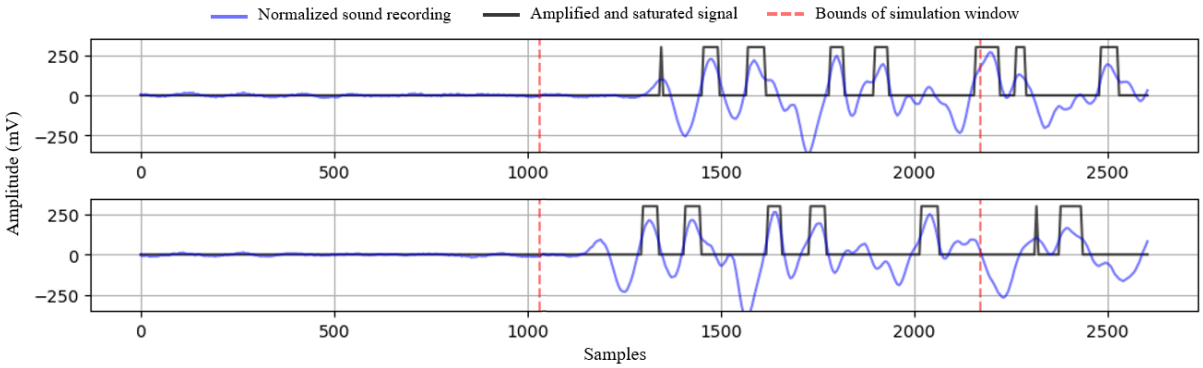


Fig. 4.9 Incorrect matching of wavefronts creates an ITD estimation error. The first front in the second channel is slightly lower than V_{sat} from ILDs, such that the square signal (input to the coincidence detector) has incorrect matching.

The V_{sat} values used throughout the chapter are reported in dBu in Table 4.2. A few arbitrary values of V_{sat} were tried, that are above noise level and as low as possible. In order to better understand these values, the proportion of the values with the average noise amplitude and with the highest signal maximum amplitude are reported. The maximum amplitude is chosen after high-pass filtering at the closest distance 24 cm and at 0°, where both channels' amplitudes are maximized. Amplitudes of the recorded sounds are rather low in the audio files; despite being relatively loud to the human ear, they are small in front of the maximal recordable amplitude.

Table 4.2 Values of the Detection Threshold V_{sat}.

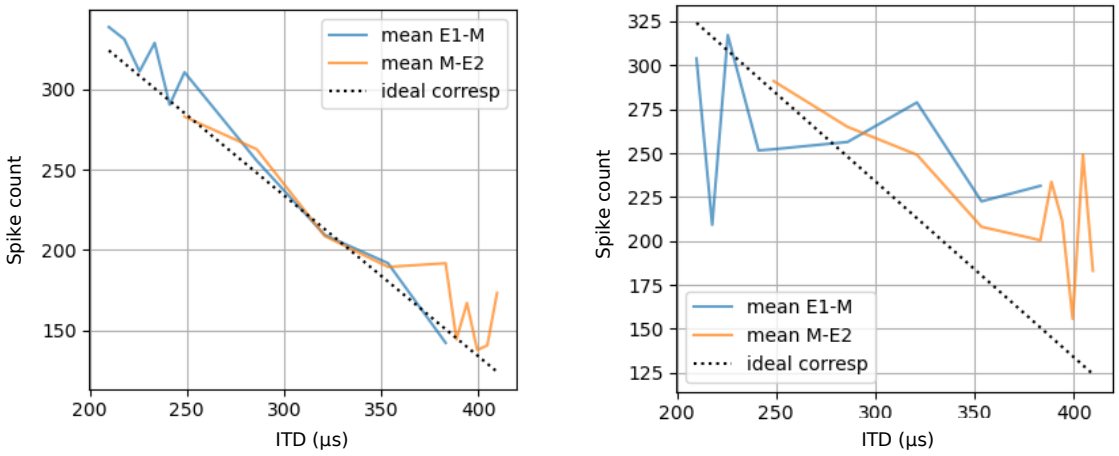
Name	Value[1] (dBu)	Proportion of the Average Noise Amplitude[2] (dBu)	Proportion of the Highest Signal Amplitude[3] (dBu)
V_{sat1}	-65.54	7.88	-40.76
V_{sat2}	-62.62	10.81	-37.79
V_{sat3}	-58.69	14.74	-33.70
V_{sat4}	-53.5	19.93	-28.54
V_{sat5}	-47.48	25.95	-22.67

[1] From the float-32 value of the wav file.

[2] Average value -71.22 dBu of all recording sets.

[3] Highest value -22.62 dBu across all recording audio files and binaural pair that both channels reach simultaneously.

Fig. 4.10 gives an example of the comparison between the ITD estimated with the theoretical ITD for two different values of V_{sat} on recording set A at 1m. Already, one can observe that the choice of V_{sat} has a great impact on the results. This impact on DOA estimation is further examined and quantified in the next section.



(a) (b)

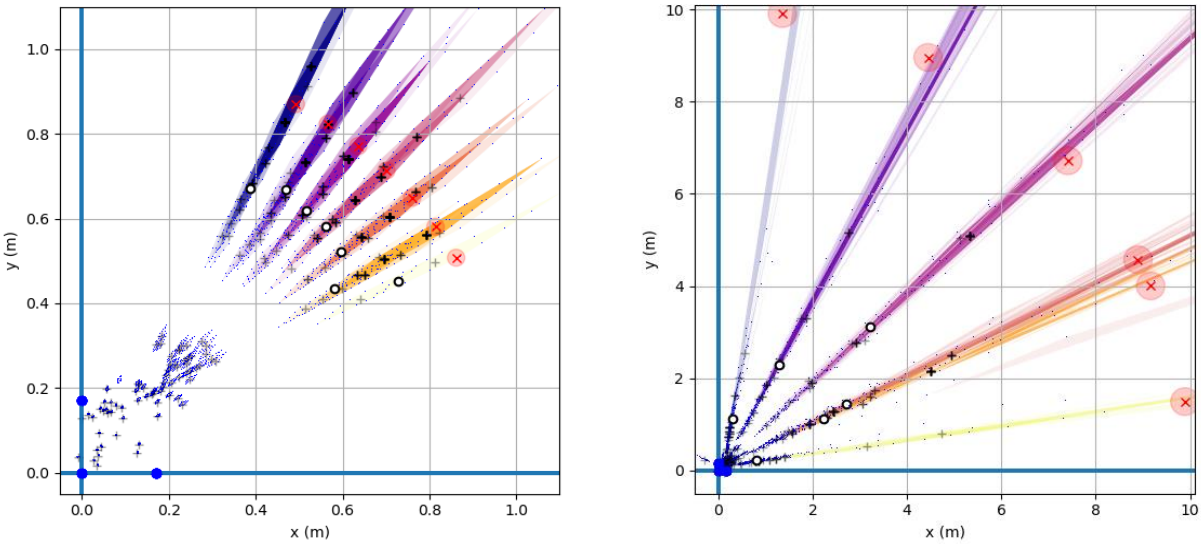
Fig. 4.10 Output of the coincidence detector model for set A at 1 m with thresholds (a) V_{sat3}, and (b) V_{sat5}.

Spatial Coordinates from ITD Estimations

To evaluate the localization performance of the model, the position estimations are computed from the ITD measured by solving (4.11) and (4.14). Polar coordinates, namely the azimuth and distance, are then retrieved from the 2-D coordinates.

In practice, different error sources impact the precision of the final estimation. Firstly, acoustic sources are manually placed in space. An error approximated to 1° (higher for set C) is created due to the imperfect projection of the angle from the reference microphone for the source placement. This error is represented in Fig. 4.11 by a red disk. Secondly and according to (4.2), discretization by spike encoding of the measured time delays limits the angular resolution to approximatively 1° if the limitation imposed by the sampling frequency is considered, as previously explained in section 4.1.2. Besides, this error can easily be rounded to 1° since the microphones can hardly be considered as points in space due to their volume.

Expressed in microseconds, these errors are considered visually by creating a confidence interval around the ITD estimation for both binaural pairs of the rectangular array, resulting in an area of possible coordinates on the 2-D plane with a diamond shape due to the ITD incertitude taking the form of hyperbola. Clearly visible on Fig. 4.11, a small change in azimuth (or ITD) can lead to a great change in distance estimation, a phenomenon known as a high geometric dilution of precision (GDOP). Fig. 4.11a shows the position estimation for all detected sounds in the recording set A at distance 1 m (V_{sat5}). Already at this distance from the reference microphone, the estimated distance may vary of about 50 cm at 0° and 150 cm at ±30° when including errors from discretization, so the error may reach around 50% and 150% of the real distance respectively. The high GDOP is very limiting for estimating 2-D spatial coordinates of a sound source. However, the configuration of the pairs allows the DOA to be correctly determined. Moreover, for sound sources placed at distances far superior to the microphone array's baseline, the hyperbolas intersect at a very narrow angle. Error zones plotted in Fig. 4.11b for a sound source placed at 10 m from the array show more than 100 m of possible distances.



(a) (b)

Fig. 4.11 Example plot of all estimated positions with error zone and centroid for (a) set B at 1 m with threshold V_{sat5}, and (b) set C at 10 m with threshold V_{sat2}. Error zones for estimation of the same truth angle have different colors, placement error zones are red disk around the truth position marked by a red cross (2 cm radius at 1 m, and 30 cm at 10 m). Estimated positions are marked by black transparent crosses, and estimated positions are averaged per angle by a centroid plotted as black circles with white fill. The blue disks represent the microphones. Small blue dots are points delimiting each estimated error zones.

With a suited V_{sat}, as the examples shown in Fig. 4.11, one can observe that the distance is better estimated for angles close to 0° and the accuracy deteriorates when approaching ±45°. Error zones indicate that the estimation is within the theoretical errors considered. While the set A at 1 m with suitable V_{sat3} has 66% of its estimation and placement error zone overlapping, or the set B at 1 m with V_{sat5} has 65% of overlaps, unsuitable thresholds like set A at 24 cm with V_{sat5} or V_{sat1} have no overlapping of its error zones with the real positions of the source.

Overall DOA estimation performances are evaluated using the MAE of the azimuth angle whose distribution for different detection thresholds are represented by boxplots in Fig. 4.12a (default parameters of the boxplot function from the Python library Matplotlib are used). Outliers are not shown for legibility as many have extreme values due to poor detections. The computation of the overall MAEs does not take into account absences of solution to the multilateration equations (results are undefined, *NaN* values as in "Not a Number").

Better results are observed for V_{sat3} and above, with mean and higher whisker respectively below 5° and 7° MAE. Fig. 4.12b shows azimuth MAE distributions per distance with most tuned V_{sat} for recording sets A and B. With a 24 cm distance from the reference microphone, the lower whisker does not reach 0° unlike at other distances. Besides, V_{sat3} is chosen as the most collectively suited V_{sat} value that maximizes the performances when considering all distances. Corresponding azimuth MAE distributions are plotted in Fig. 4.12c.

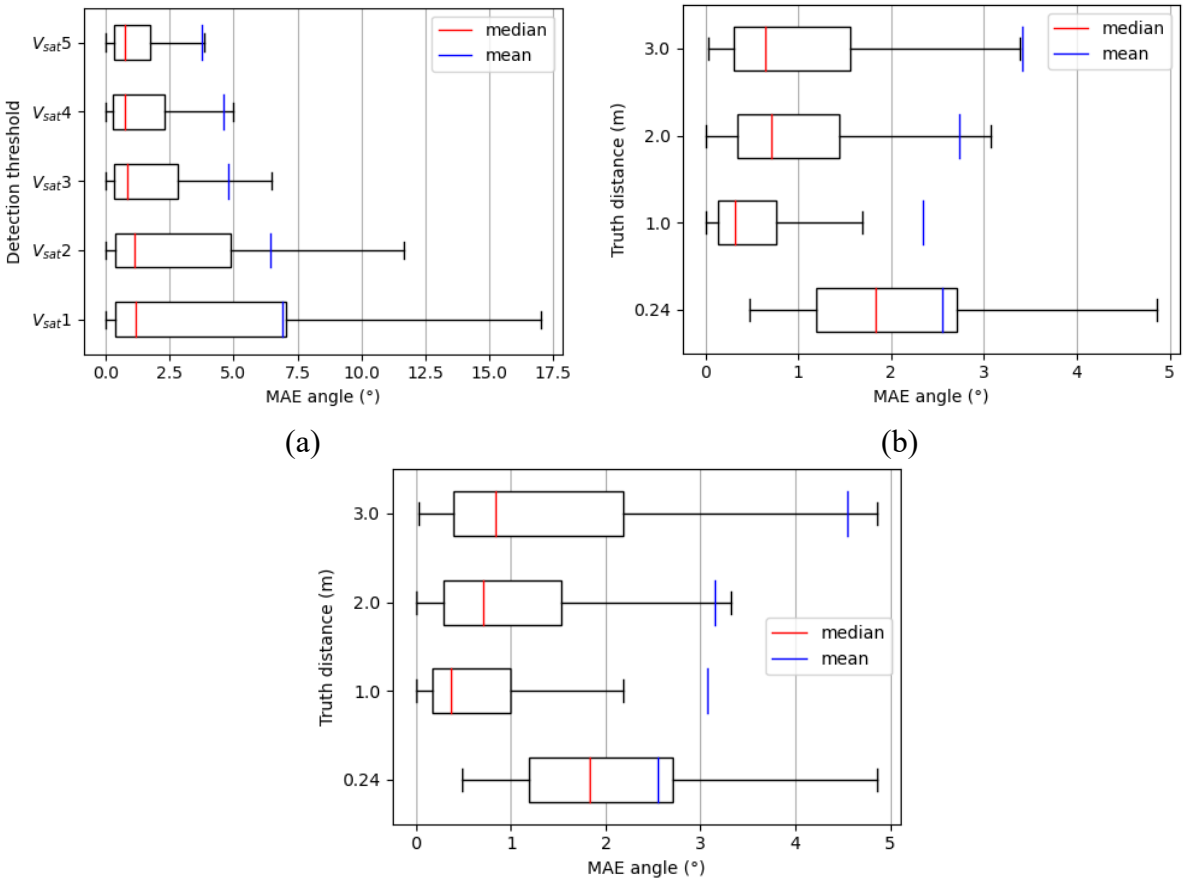


Fig. 4.12 MAE angle boxplots. MAEs are considering (a) all detections shown per V_{sat}, (b) best individually suited V_{sat} per distance (in ascending distance V_{sat3}, V_{sat5}, V_{sat5}, V_{sat2}), and (c) best collective suited threshold V_{sat3} for all distances except 10 m for which the distribution's box and outer whisker reach 15° and 30° respectively. Outer whiskers are represented by vertical black line symbols such that the higher whisker has the value $1.5(Q3 - Q1) + Q3$, and the lower whisker value $-1.5(Q3 - Q1) + Q1$), with Qi the quartiles. Red and blue vertical lines are the median and mean of the distributions.

Complementary to Fig. 4.12, Table 4.3 reports performance metrics for the most suited thresholds individually and collectively per distance for DOA estimation with different angular tolerances. Highest performances per reported metric are highlighted in bold font. The most suited V_{sat} are chosen from best MAE performances. The azimuthal accuracy for a tolerance ϵ is computed according to

$$ACC\ (\pm\epsilon) = \frac{1}{N_d}\sum_{i=0}^{N_d}[|x_i - \hat{x}_i| < \epsilon]\ , \tag{4.22}$$

with N_d the total number of detections, $|x_i - \hat{x}_i|$ the absolute error where x_i is the truth angle and $\hat{x}_i$ the estimation, and $[|x_i - \hat{x}_i| < \epsilon]$ is a Boolean equal to 1 if $|x_i - \hat{x}_i| < \epsilon$ and 0 otherwise.

Table 4.3 Performances of most individually and collectively suited V_{sat} for DOA estimation per distance.

Distance (m)	**0.24**	**1**	**2**	**3**	**10**
Best Individual V_{sat}	V_{sat3}	V_{sat4}	V_{sat5}	V_{sat2}	V_{sat2}
Detection with Solution (%)	73.55	**97.90**	96.60	96.32	71.03
Accuracy ±1° (%)	11.57	**79.87**	56.14	62.34	7.95
Accuracy ±2.5° (%)	52.89	**85.53**	82.42	75.76	23.33
Accuracy ±5° (%)	71.07	**86.16**	86.01	82.9	37.44
Accuracy ±10° (%)	72.73	88.68	**89.79**	87.01	50.51
Best Collective V_{sat}			V_{sat3}		
Detection with Solution (%)	73.55	88.96	93.96	**94.41**	57.42
Accuracy ±1° (%)	11.57	**67.29**	55.85	52.61	3.57
Accuracy ±2.5° (%)	52.89	74.58	**75.85**	71.17	16.48
Accuracy ±5° (%)	71.07	76.88	**80.00**	74.23	32.97
Accuracy ±10° (%)	72.73	81.25	**84.53**	82.7	39.84

Best performances with individually tuned V_{sat} are obtained at 1 m, reaching at low tolerances 79.87% (±1°) and 85.53% (±2.5°). At 2 m, accuracies do not fall behind for a tolerances 2.5° and 5° and is 1% better for tolerance 10°. The highest proportion of the detections with estimated ITD pair solving the aforementioned multilateration equations is also obtained at 1 m. For distances between 1 m and 3 m, accuracies are all above 87%. At 10 m with individually tuned V_{sat2}, the accuracy reaches 50.51% (±10°) for a proportion of correct fronts detected at both binaural pairs of 69.4%. Using the common value V_{sat3}, the best accuracy drops to 75.85 (±2.5°), 80% (±5°) and 84.53% (±10°) for a 2 m distance. Nonetheless, accuracies for distances 1 m to 3 m are all above 71.1% (±2.5°), 74.2% (±5°) and 81.2% (±10°).

Depending on the detection threshold V_{sat} and as a consequence of the ILD, a relative proportion of the detections matches the incorrect wavefronts between the two microphones of a binaural pair, leading to significant errors in ITD and position estimation. For example, with suitable thresholds V_{sat3} for set A at 1 m, 85% of correct wavefronts are detected, but with V_{sat1} and V_{sat5} for set A at 24 cm, only 38% and 23% are.

Noisy Environment

In order to introduce the limits of the adapted model in a noisy context, a study is performed to quantify the loss in accuracy of the DOA estimation for six SNR levels between 3 dB and 20 dB. Two recordings from the set B are selected, at 0° and 15°, and 32 clicks of similar amplitude in total are kept for the simulation. A white noise is added at different amplitude levels by software. In this study, V_{sat} is fixed according to the noise level according to $V_{sat} = \bar{n} + m$, with V_{sat} the detection threshold in dBu, $\bar{n}$ the mean noise level in dBu, and m = 23 dBu. In fact, without adaptation of V_{sat}, no relevant ITD estimation can be obtained when V_{sat} reach below $\bar{n}$. For the following results, the margin is set based on $\bar{n}$, observed for all SNR.

A strong decrease in accuracy from 10 dB SNR to 3 dB SNR can be observed in Fig. 4.13, which shows the evolution of the DOA estimations. For lower SNRs, fewer detections lead to a solution, as a result of a high detection threshold compared to the signal, and the influence of ILD at 15° of azimuth. Among the detection with solutions, the accuracy remains higher than 79.4% for tolerances ≥ 2.5° and for SNR of ≥ 10 dB. The accuracy falls to 15% (±2.5°) and 28.12% (±2.5°) at 3 dB and 5 dB SNR respectively. Nevertheless, except for an SNR of 5 dB, at least 60% of the detections with solutions are correctly estimated with a tolerance of 1°.

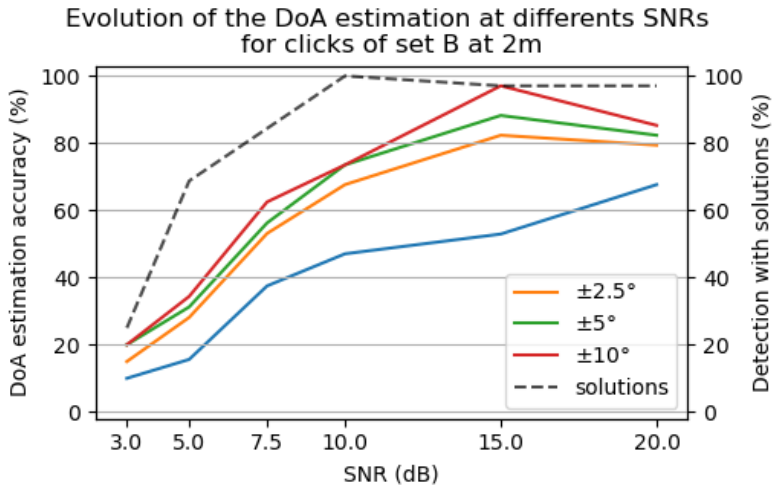


Fig. 4.13 Evolution of the DOA estimation accuracy for SNR between 3 dB and 20 dB. The sound from the set B are played at 2 m from the reference microphone. Accuracies for tolerances 1°, 2.5°, 5°, and 10° are plotted as well as the proportion of detections with a solution.

Before further discussing the results obtained for impulsive sounds, localization of tonal sounds with the model is investigated.

4.2.2 Tonal Sounds

Tonal signals dominate most studies in SSL studies because of their prevalence and strong taxonomic relevance. They are typically continuous or modulated in pitch, often forming structured sequences such as songs, chips, or calls. Acoustic patterns tend to be highly distinctive, allowing reliable identification of the species or individuals and making tonal sounds a strong focus in monitoring.

Consequently, the HRD-based coincidence detector is first tested with simple pure tones as common practice, then with natural tonal sounds. However, several tests with bird and field cricket sounds reveals the incapacity of the model to extract consistent ITD estimations from natural tonal sounds. The causes of this limitation are therefore examined and described in this section.

Recording Sets

Since a higher interest is given to the study with natural sounds, a small recording set of pure tones is produced. Pure tones with frequencies 1 kHz, 2 kHz, 4 kHz, and 8 kHz are generated

digitally and modulated by a square window. Each pure tone is played once at each position during 1 s.

In the case of natural tonal sounds, the recording set used to examine the coincidence detector is not a mapping of the angular space but contains a series of recordings at 0°, ±45°, and random azimuths. Because incoherent localization results were systematically observed, it was deemed unnecessary to complete a precise mapping in additional recording sets. Both recording sets' characteristics are reported in Table 4.4.

The 4-microphone stand, available at the time of the tonal sound study, was used to produce the 2-D recording sets D and E by using only the microphones M, E_1, and E_2. The loudspeaker was placed at elevation 0° to remain in the 2-D plane of the microphones used.

Table 4.4 Characteristics of the tonal sounds recording sets.

Set	**D**	**E**
Number of microphones	3	3
Sound type	Generated pure tones	Pre-recorded bird and cricket calls and songs
Distances* (m)	1	1, 2
Azimuthal range* (°)	±45°	±45°
Angular resolution (°)	15°	–

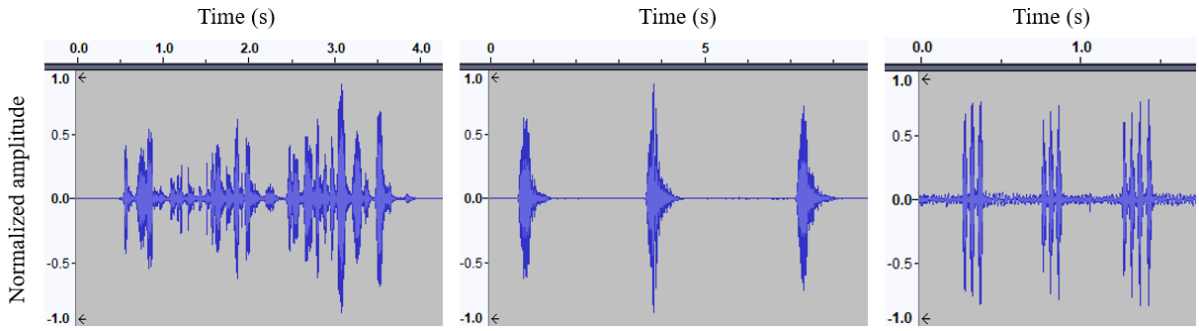
* From reference microphone M and 0° line of array.

Bird songs and calls, as well as field crickets calling songs, were used for building the recording set E. Audio recordings of real birds and field crickets in their environment were retrieved from the online dataset Xeno-Canto [111] as pre-recorded sounds to be played at different positions within the microphone array's angular range (including the azimuths ±45° and 0°). Arbitrarily chosen to obtain different temporal and spectral characteristics, bird songs and calls in the following recordings are used: XC85360[1], XC106781[2], XC189068[3], XC322175[3], XC352372[3], XC391137[3], XC444029[3], XC566055[3], XC664557[3], and XC933620[3]. For field cricket calls, the following recordings are used: XC821599[2], XC853470[4], XC867042[4], XC967416[4], and XC967429[4]. The extracted sequences from the original audio file have duration ranging between 2 s and 14 s for birds, and between 9 s and 24 s for crickets.

An example of each sound type present in set E is shown in the Fig. 4.14, where the waveforms correspond to the audio files played by the loudspeaker; original audio files are denoised in software to isolate sounds of interest. While field crickets calling songs are narrow band and short, bird calls differ more across species and span over a wider spectral band. Bird songs occupy an even larger spectrum, usually composed of complex successions of chirps and calls in specie-specific patterns. Overall, the extracted sequences of bird vocalization have

[1] CC BY-NC-SA 3.0
[2] CC BY-NC-ND 4.0
[3] CC BY-NC-SA 4.0
[4] CC BY-SA 4.0

spectrum spanning between 200 Hz and 7 kHz. The spectrums of field cricket calls span between 4.2 kHz and 5.8 kHz. For both, harmonics can reach up to 19 kHz.



(a) (b) (c)

Fig. 4.14 Examples' waveform of pre-recorded tonal natural sounds. Sound types are (a) bird songs, (b) bird calls, and (c) field crickets calling songs, illustrated by reproduced subsequences of the audio files XC566055, XC189068, and XC821599 respectively. CC BY-NC-SA 4.0 and BY-NC-ND 4.0. Visualization on Audacity.

ITD and DOA Estimation of Pure Tones

The most suitable detection threshold V_{sat1} is used for the whole recording set D. The DOA estimation accuracy is relatively similar to the previous results obtained for impulsive sounds, with an azimuth MAE of 0.97° and ITD MAE of about 43 µs overall. The DOA estimation accuracy is 42.86% (±1°), and it reaches 57.14% (±2.5°) and 60.71% (±5° and ±10°) which correspond to all detection with the solutions. The best precision is obtained for the source placed in front at 0° with an accuracy of 100% (±1°), while there is no solution at the extremes ±45° of the angular range. The same observation is made for distance for which the overall MAE is 36.9 cm, with 14.0 cm MAE at 0°. Positions are estimated nearer to the microphone array when the source is placed closer to the extreme azimuths.

Pure tones are rather artificial compared to the wide range of acoustic sounds found in real-world soundscapes. Although localization results are similar to the reported performances of click inputs, the limitation of the coincidence detector appears when providing natural tonal sounds in input.

Impossible ITD Estimation of Complex Tonal Sounds

Against the initial expectations, the HRD-based coincidence detector is not able to provide coherent ITD estimations on the tonal sounds recorded as a result of the detection method. All detections are incorrect, either from detecting noise instead of the sound's wave, or from extracting ITDs higher than ITD_{max}. In the end, no meaningful DOA estimation is obtained as a large majority of the estimated ITD pairs lead to random locations or to the impossible resolution of the multilateration equations.

Using the most suited detection threshold V_{sat1} for set E, 85.23% of the detections lead to ITDs larger than ITD_{max} equal to 500 µs. If ITDs up to 10 ms are enabled, it would result in a mean ITD estimation of 1.87 ms, with a median of 656 µs and 2.71 ms of standard deviation. Only 4.21% of all detections are solution of the multilateration equations, such that an

accuracies of 1.58% for tolerances up to ±10° are obtained in azimuth when taking into account all detections. Putting aside detections without solution, an MAE of 18.39° is obtained. These results are insufficient to consider reliable the estimation of ITDs on tonal sounds.

To better understand the extent of the problem, let us suppose a fine filtering in pre-processing without considering the hardware. A simulation with a narrow band-pass Butterworth filter of order 10 is run to isolate a meaningful part of the target sound's spectrum from the noise. The cricket calling song in XC967429 is taken as example, where the low and high cutting frequencies of the filter are set to 5 kHz and 5.1 kHz respectively for a band of 100 Hz. In the end, even with V_{sat} finely tuned, extracted ITDs are inconsistent with their theorical value although an ideal and selective filter was applied upstream of the coincidence detector. About 97.4% of the detection results in ITDs larger than ITD_{max}, with a mean ITD estimation of 3.89 ms, a median of 3.59 ms and 1.95 ms of standard deviation when allowing ITDs up to 10 ms. No detection is solution of the multilateration techniques.

Identifying the Limitation

The poor localization results observed are essentially caused by two factors creating limitations related in fact to the chosen detection method based on a threshold.

Firstly, the microphones have a location and frequency-dependent radiation pattern affecting the signals shape at the onset. Indeed, depending on the frequency and location of the acoustic source, the reception in decibel varies more a less and differently across the spectrum. When using pure tones, the impact of the microphone's radiation pattern does not pose a problem since only the amplitude of the pure tone is affected. With complex sound spectrums, however, disparate modifications of the frequency components' amplitude lead to the alteration of the signal's overall envelope in the time domain. Enhanced by the inputs' rich and narrowband spectrum, onsets are altered such that it is impossible to determine the ITD with a simple detection threshold. Furthermore, even to the eye using visual cues on the channels' waveform, the ITD cannot be correctly estimated. Computation of standard mathematical ITD extraction methods are then required, like generalized cross-correlation, to provide an estimation.

Secondly, low frequency modulation (envelope of the signal) throughout the duration of tonal sounds is naturally also present at onsets. Unlike the sharp onsets of click-like sounds providing distinct first fronts, the tonal sounds have long onsets that are incompatible with the simple detection method employed. Slight variations of the waveform amplitude, resulting either from ILDs or the microphones' disparate radiation pattern, can cause great variations in time difference at slowly rising onsets. In order to highlight the influence of the onset modulation, an additional recording set F, detailed in Table 4.5, is produced using the same pure tones as set A and white noises signals.

White noises follow an uniform distribution around 0 (spectral activity between 1.5 and 15 kHz). Each example (tone or noise) is played 7 times at the same position during 1 s with a Hann window more or less pronounced for modulation of the envelope. The window is applied at the onset and offset of the sounds with increasing total duration in {5, 10, 100, 200, 500, 1000} in ms. An example of a digital generated modulated pure tone is provided in Fig. 4.15.

Table 4.5 Characteristics of the modulated simple sounds recording set.

Set	**F**
Number of microphones	3
Sound type	Generated pure tones and white noise
Number of sound per position	45
Distances* (m)	1
Azimuthal range* (°)	±45°
Angular resolution (°)	15°

* From reference microphone M and 0° line of array.

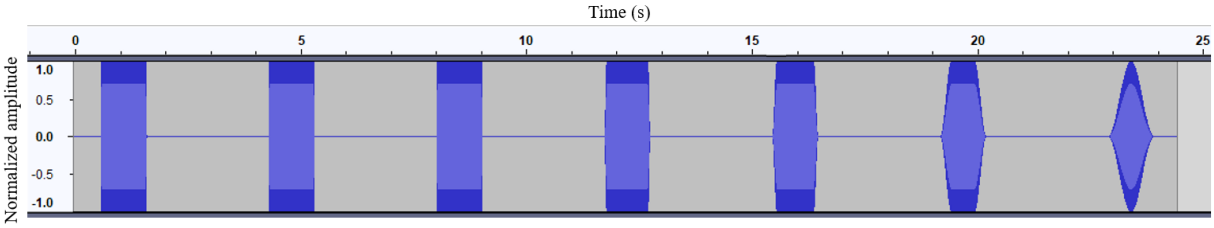


Fig. 4.15 Application of a Hann window with increasing duration on an 1 s long 1 kHz pure tone. At the far left, no window is applied, then a Hann window is applied with 5 ms to 1 s of total rising and falling duration, encompassing the whole sound at the far right. The waveform is the generated audio file before recording. Visualization on Audacity.

Table 4.6 reports the detection and localization performances per onset duration. The ITD extraction results gradually deteriorate, with best performances reached when the onset of the pure tone or white noise is the sharpest. DOA estimations are still obtained unlike with set E since the sounds are pure tones or broadband, and thus less affected by the microphones' radiation pattern. Even longer onsets are tested than those observed in set E, which provides position estimations but with poor precision.

Table 4.6 Detection and DOA average performances on set F per onset duration.

Onset duration (ms)	**0**	**2.5**	**5**	**50**	**100**	**250**	**500**
MAE ITD (µs)	**71.56**	220.09	242.65	312.15	325.7	299.2	294.68
Detection with solution (%)	**56.67**	26.67	50.0	60.0	46.67	46.67	50.0
Accuracy ±1° (%)	**33.33**	10.0	3.33	6.67	6.67	10.0	3.33
Accuracy ±2.5° (%)	**46.67**	13.33	10.0	13.33	13.33	13.33	6.67
Accuracy ±5° (%)	**53.33**	13.33	16.67	13.33	13.33	13.33	6.67
Accuracy ±10° (%)	**56.67**	20.0	30.0	16.67	16.67	13.33	13.33

With click-like sounds, these factors are minimal thanks to their natural sharp onsets and broadband spectrum. As a result, the coincidence detector's localization potential is limited to clicks, or at least sounds with similar onset sharpness.

4.2.3 Further Validation of the Model's Potential

Although tonal sounds did not lead to any results, localization performances were successfully evaluated with click-like sounds, providing a ground for further validation of the model potential. Until now, only 2-D results were reported, yet the simplification of the multilateration equations extends to 3-D. A click recording set is used to identify the elevation estimation accuracy of the model with a 4-microphone array.

On the basis of the 3-D results detailed thereafter, localization performances with click-like sounds are also assessed with band-pass filtering of the input signals. In the literature, a filter bank or artificial cochlea is systematically used to extract ITDs for various frequency components. Despite the model not being suitable for tonal sounds, it is highly relevant in SSL; a study is required in the case of clicks whose sharp onset is strongly related to their rich spectrum.

Validation of DOA Estimation in 3-D Space

In the same manner than estimating in 2-D space, the coincidence detector extracts ITDs of the three binaural pairs from the 4-microphone array in a rectangular configuration. Table 4.7 describes the 3-D recording set G, which was produced using the 4-microphone stand depicted in Fig 4.5b. Elevations were chosen according to the loudspeaker stand height range, maximum at about 1.7 m. Since the tonal natural sounds did not provide any localization results, the pre-recorded click-like sounds introduced in section 4.2.1 are used for validation of DOA estimation in 3-D space.

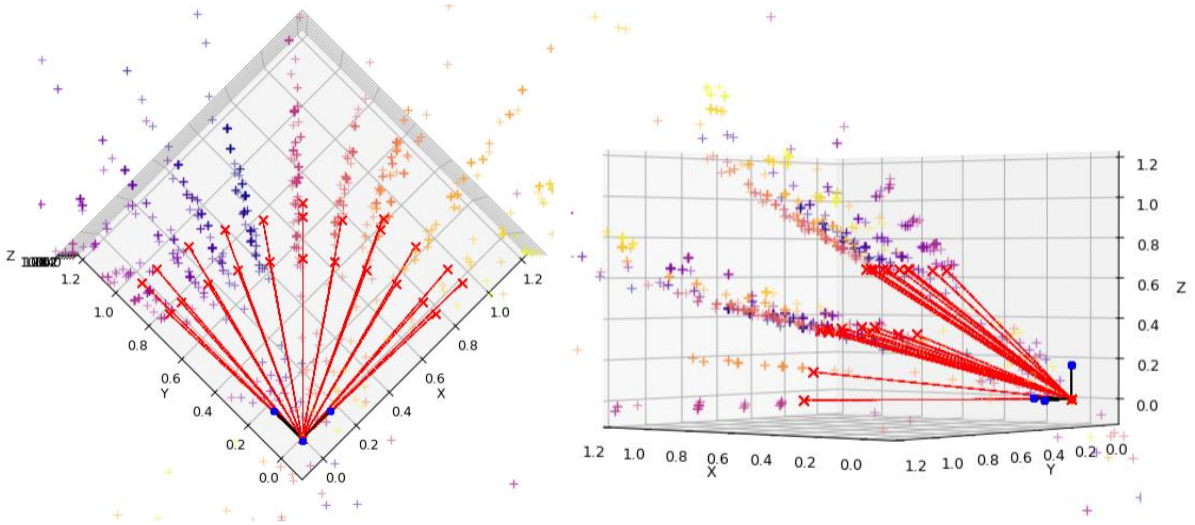
Table 4.7 Characteristics of the 3-D recording set.

Set	**G**
Number of microphones	4
Sound type	Pre-recorded click-like sounds
Number of sounds per position	60~64
Distances* (m)	1
Azimuthal range* (°)	±45°
Elevation range* (°)	0°→40°
Angular resolution (°)	10°, 5° (azimuth) 20° (elevation)

* From reference microphone M and 0° line of array.

As expected, position estimation made in the 3-D space give similar results to 2-D estimations in azimuth and distance. We observe that elevations can also be obtained with great precision. Fig. 4.16 shows a scatter plot of the estimations made on the detected clicks from the recording set G (all are at a distance of 1 m from microphone M, most elevations are 20° or 40°). Here, the estimations are performed with adapted detection threshold V_{sat3}, such that the maximum amount of detections are obtained. Overall accuracies in azimuth and elevation for set G reached 68.44% (±5°) and 73.09% (±5°) respectively. At a high tolerance accuracies become 73.65% (±10°) and 74.72% (±10°) in azimuth and elevation respectively. The MAE is

11.91° and 4.53° in azimuth and elevation, such that by taking only detection with a solution, 89.84% and 91.02% of the respective non-null estimations fall below the 5° tolerance.



(a) (b)

Fig. 4.16 Point scatter of positions in 3-D space from set D (1 m) with V_{sat3}. Position estimations give correct precision in (a) azimuth and (b) elevation. For better legibility, error zones are not plotted but would have a diamond-like shape with variable thickness as a consequence of the hyperboloids' shape. Estimations are marked by crosses whose color indicates its truth position. Dashed red lines connected to truth positions highlights elevations and DOA estimations since error zones are not plotted.

Noticeable gaps in azimuth can be seen for several positions on Fig. 4.16a. It appears to be the result of an incorrect placement of the source for a given truth angle. Estimations that are well aligned together and not with the truth position reinforce the hypothesis. Estimations scattered outside the angular range of the microphone array are the consequence of incorrect detections or random noise detection (imperfect automatic framing). Hence, the performances with low tolerance could be increased by offsetting the placement error.

Band-Pass Filtering

For known spectral characteristics of the target sounds, adding a selective band-pass filtering instead of a single high-pass at 100 Hz would improve the detection condition. In the environment, noise sources are often thermal noise, largely spread over the spectrum, or parasite sounds emitted in the vicinity of the microphone array. Band-pass filtering the input signal contributes to isolating the target sound. For this short study, the recordings from set G are kept at high SNR (no added noise) in order to observe the impact of band-pass filtering on the system.

In the case of clicks, the spectral content of these impulsive sounds spans from 0 Hz to 10 kHz (faintly until 20 kHz). Using a Butterworth band-pass filter of order 2, bands 500 Hz, 1 kHz, and 5 kHz are considered for narrow and wide spectral segmentation. For each band, center frequencies every 1 kHz from 1 kHz (3 kHz for bandwidth 5 kHz) to 10 kHz are tested. Since the spectral content and resulting temporal amplitudes varies according to the bandwidth

and central frequency of the filter, different V_{sat} values suitable individually for each audio file and not reported in Table 4.2 are used.

Table 4.8 reports the DOA estimation accuracies on set D for each bandwidth averaged over all central frequencies, as well as non-filtered results for comparison. Better performances than the 3-D DOA accuracies previously reported are naturally observed since V_{sat} is finely tuned to each audio file.

Table 4.8 Detection and DOA average performances with band-pass filtering.

Bandwidth (Hz)	**500**	**1000**	**5000**	**(No filtering) ***
Detection with solution (%)	73.75	74.74	74.8	**80.3**
Azimuth accuracy ±1° (%)	**16.17**	10.31	10.34	11.04
Azimuth accuracy ±2.5° (%)	**40.16**	27.62	28.09	31.24
Azimuth accuracy ±5° (%)	60.27	57.62	59.9	**68.44**
Azimuth accuracy ±10° (%)	62.54	63.95	65.57	**73.65**
Elevation accuracy ±1° (%)	38.46	38.89	41.54	**46.86**
Elevation accuracy ±2.5° (%)	57.59	57.42	60.12	**67.82**
Elevation accuracy ±5° (%)	61.47	62.91	65.1	**73.09**
Elevation accuracy ±10° (%)	63.88	65.34	67.13	**74.72**

* No band-pass filtering is done, only the prefiltering at 100 Hz.

Best accuracies are obtained without band-pass filtering. In fact, DOA performances increase with larger bandwidth. Looking at individual central frequencies for each bandwidth, best results in both azimuth and elevation are obtained at 3 kHz and 5 kHz for bandwidths 1 kHz and 5 kHz. Overall best result is reached with bandwidth 1 kHz centered on 3 kHz with accuracies around 67.68% (±5°) and 75.87% (±10°) for azimuth, and 75.55% (±5°) and 76.52% (±10°) for elevation with 84% of detections with a solution. Therefore, considering a frequency band adapted to the target sound does increase DOA accuracies.

Without focusing on the most suitable center frequency, overall drops in accuracy with band-pass filtering in a quiet environment are explained by the nature of clicks. Impulsive sounds have sharp onsets suitable for first fronts detection. On the other hand, slowly rising amplitudes created by the filtering contribute to the generation of more ITD estimation errors resulting from ILDs or loud in-band noise. In any case, accuracy drops can be viewed as a compromise in a noisy context where noise is not broadband.

4.3 Discussion

The experimentations in simulation with real-world acoustic recordings provided great insights on the capabilities of the proposed adaptation of the HRD-based detector, where both satisfying and incoherent localization performances were obtained. Results enlightening on the current potential of such ITD extractor are discussed in this section.

4.3.1 DOA Estimation Performances Analysis

The architecture of proposed model allows high angular resolution with few neurons and simple processing. The required number of neurons to estimate the ITD does not bound the resolution of the time delay estimations compared to a Jeffress-based model. Instead fast firing neurons enables the detection and encoding of ITD as low as 1 µs (or 3.7 µs for an encoding V_{DD} set to 300 mV and a spiking frequency of 270 kHz) if the sound collectors allow it.

The model is able to give good DOA estimations for the majority of the click detections with suitable parameters between 1 m and 3 m of distance. Although few detections reach below the 1° MAE, an accuracy of at least 71.1% (±2.5°) can be observed. At 24 cm from the reference microphone, the source is very close to the array creating strong ILD around ±45°, and at 10 m, the click-like sound received is more altered by ambient noise, resulting in lower accuracies.

The azimuths estimated using multilateration on the 3-D click recording sets provide similar precision as 2-D estimations. Because the 3-D position is computed simply using an additional coincidence detector for estimation of a third ITD, this great precision was expected. Aside from a placement error evaluated between 1° and 4° at certain positions in set G, DOA estimations show consistency across the detected clicks played at a same position. In the end, it demonstrates that the system can easily be extended to the 3-D space.

A small study was conducted with white noise in different SNR scenarios introducing the current limits of the model in a noisy context. The loss in accuracy is essentially the result of incorrect matching of wave fronts that originates from the presence of ILD, as it has already been detailed. Besides, adding noise deforms the waves of the signal, triggering a detection slightly before or after the wave front of the clean sound signal. Above 3 dBu SNR, the sound is drowned in the noise and no relevant detection can be performed without using a frequency segmentation upstream to increase the SNR in the frequency band of interest and to identify the signal.

To obtain localization performances, a straightforward multilateration method was used without integrator of error factors to improve the estimations; the goal is to evaluate the ITD extractor with all its limitations. As such, a flaw of the hyperbolic intersection resolution is its low error tolerance, being very basic and without consideration of the noise. Estimating the ITD of one pair with a difference too large compared to the truth value can leave the equation system with no solution. In 3-D, four channels are required, increasing the chance that one channel has an incorrect detection. Besides, it is impossible to infer which ITD is incorrect once this system is deployed for a localization application. Nevertheless, when ITD values are missing, a position on the resulting 2-D plan can still be computed with two ITDs estimations. For only one ITD available, DOA can be estimated from the binaural equations (see section 4.1.3) with front-back confusion. In these conditions, the model can still provide localization information with the multilateration technique despite increasing the incertitude.

4.3.2 Detection Threshold: Strong Dependence and Limitations

The ITD errors that the multilateration technique is not able to handle are in fact, for the most part, caused by the thresholding method employed to detect first fronts. Measuring the ITD at

the onset of a sound allows to discard the effect of small reverberations and if repeating reverberated sounds are sufficiently spaced in time. The choice of V_{sat} and the form of the studied sound become important for correct measurements. In fact, the preprocessing is currently not able to sufficiently reduce the impact of noise or completely remove the ILD since the presence of background noise prevents the choice of a low V_{sat}. This creates a strong dependence of the model to the sound and noise levels, highlighted by the dependency to V_{sat}, because of the impact of the waveform to a fixed threshold. This value needs to be chosen as the most suitable for a wide range of distances and according to the SNR of the sound's environment. Hence, not all distances can be considered when working simultaneously with small and great distances. Choosing V_{sat3} for example, would be the best compromise when working in near-field, but results in fewer detections and lower precision at 10 m.

One of the most challenging problems is the noise added to the raw waveform (or reverberation traces when close temporally), which induces incorrect matching of fronts, thus leading to poor estimations. Alongside noise, ILDs (even small) create differences in amplitudes, and together, they affects the meeting point between V_{sat} and the wavefronts such that incorrect fronts are detected as a pair.

This issue is accentuated with tonal sounds and band-pass filtering. The tonal recording set E did not produce any consistent localization results because of the microphone radiation pattern and long onset, as it has already been described in section 4.2.2. With a basic threshold, a small change in amplitude creates a large temporal difference for slowly rising amplitudes. This behavior is similar to the changes in distance for a small azimuth change occurring from the GDOP in the hyperbolic intersection resolution. The incompatibility of the proposed system to tonal sounds, which represent a large proportion of biodiversity acoustic activity, is a major limitation that suggests the use of another method for wavefront detection.

Yet, the overall observation of the proposed ITD extractor's incapacity to handle soft onsets in tonal sounds does not contradict biological observations. In biology, inner hair cells continuously encodes into spikes the positive waves of sound in a phase coding scheme. Naturally, extracted ITDs are impacted by the period of the waves. Above a certain frequency, ITDs between halfwaves are not consistent with the true position of the acoustic source. Then, other location-dependent cues like ILDs and spectral notches prevail. The coincidence detector could be further adapted to perform an ITD extraction between all wavefronts using zero-cross detection as a phase-locked scheme. Following a biomimicking approach, band-pass filtering would become highly relevant, and the proposed system, compatible with tonal sounds but only for frequency components below 1 kHz for $ITD_{max} = 500$ µs; above half the input signal's half period, the determination of the binaural channel in phase advance relative to the other is ambiguous. Nevertheless, even with the limitations imposed by the thresholding method, the DOA estimation is not constrained to low frequencies in the current system.

4.3.3 Confrontation with the Literature

The proposed model is adapted from the HRD and similar to [62], [92], but allows a more linear correspondence between the spike count in output and the time delay. The AND operation

on the expanded detection spikes holds the ITD information only in the excitation duration of the encoding neuron. Whereas in [62], [92], the spike is expanded with low-pass filtering after cross-multiplication, so the time delay relies upon the amplitude of the excitation, leading to a non-linear correspondence, although small. Because spiking rate of neurons essentially depends on the input stimuli amplitude, having a constant amplitude for encoding the ITD increases estimation precision. Moreover, the presence of the ILD impacts the estimation of the ITD using the HRD model. For a same ITD, a change in amplitude of the signal at the detection neuron leads to a different spike timing. Consequently, when estimating the ITD between two channels, an increasing error can be observed with increasing ILD. A modification of the HRD model was thus required, namely the addition of an amplification and saturation.

In this end, [62]'s model was proposed in the context of a closed-loop system for tracking of sound sources with head rotation, such that the spike encoding was suitable for managing the acceleration and speed of the robot's movements, and studies frequencies below 1 kHz. This thesis's model, however, performs one-shot estimations studied with the highest precision and lightest processing allowed by the system. In this adaptation of the HRD and unlike the majority of the literature, the time delay estimation was performed directly on the raw waveform without the use of a cochlea or filter bank. Being compatible with sharp onsets essentially (mostly clicks), the use of filtering was not beneficial (see section 4.2.3).

The model does not achieve the high accuracy in DOA estimation reported in recent works that use MTPC [88], [90], and it has been evaluated within a more limited angular range and sound types, but the results are encouraging considering the model's low complexity. With a 3-microphone array, accuracies superior than 74% (±5°) and 82% (±5°) are obtained for distances 1 m to 3 m with a fixed collectively and individually suitable detection threshold respectively. Using 4 microphones but also a quieter environment and at 1 m only, the azimuth estimation accuracy with the collectively suited V_{sat} decreases to 68.44% (±5°), and 73.09% (±5°) of accuracy is obtained for estimation of elevations.

Table 4.9 provides a summary of the most relevant SSL systems based on the TDOA only for comparison. Works with angular resolution higher than 20° and without SSL results are omitted. Highest accuracies, MAEs below 1°, and lowest number of neurons are highlighted in bold font. For the proposed model, the reported performances do not include the results at 10 m, and the MAE performance is the mean of V_{sat3} overall results.

Unlike the neuromorphic SSL literature's commonly studied sound types, the proposed SSL system provides localization results only for impulsive sounds. The proposed model does not showcase the best SSL performances, but it is without defining an ULP limit on the computational cost of the systems.

4.3.4 ULP Consumption Potential

Considering the HRD-based model and simplified equations for DOA estimation, this simulated work is a good candidate for a low-power consumption hardware implementation. Because it does not involve delay lines, it would facilitate an implementation in a complete analog system.

Table 4.9 Comparison with neuromorphic DOA estimation systems based on ITD only.

Ref.	# Mics	# Neurons	ITD Extraction Model	Learning	Angular Resolution / Accuracy ±Tolerance	Test Environment	Sound Source	Distance (m)	Average Performances
[74]	2	1029	Jeffress	STDP (S)	±5°	Quiet	Pure tones	1	78.64% ACC
					±10°				91.82% ACC
[76]	2	4682	Jeffress	Gradient descent (S)	3°	Quiet	Pure tones	2.6	6.05° RMSE
							Noise		4.1° RMSE
[77]	2	1024	Algorithm	–	~2.6°	Noisy Reverberant	Speech	–	**0.17° MAE**
[78]	2	10	Jeffress	STDP (S)	20°	Quiet	Pure tones	1.4	3.4° MAE
[80] [1]	2	~3000	Jeffress	Gradient descent (S)	3°	Reverberant	Pink noise	2.5	5° RMSE
							White noise		4.4° RMSE
[82]	2	>890	Jeffress	State machine	0.32°	Quiet	Speech	5	3.49° MAE
						Noisy			5.57° MAE
	4					Quiet			**0.99° MAE**
						Noisy			1.18° MAE
[83] [2]	8	16	Scorpion-inspired	–	–	Quiet	Pure tones	1	4.05° MAE ±3.01° SD
[88]	4	1981	Jeffress-based	Surrogate gradients (S)	1°	Quiet	Speech, Noise	1, 1.5	1.02° MAE
						Low noise			1.07° MAE
						High noise			10.75° MAE
[90]	4	4261	Jeffress-based	Backpropagation (S)	±2.5°	Noisy	Speech, Broadband sounds	1.5	**0.6° MAE 95.61% ACC**
[62] [1,2]	2	326	HRD-based	–	5°	Quiet	Pure tones,	0.5	1.92° MAE
							Speech		5.5° MAE
This model [3]	3	**6**	HRD-based	–	~1°	Quiet	Clicks	0.24, 1, 2, 3	4.8° MAE
					±2.5°			1, 2, 3	73.86% ACC
					±5°				77.04% ACC
					±10°				82.83% ACC
					±5°, ±10°	Quiet	Pure tones	1	60.71% ACC
					±5°, ±10°	Noisy (SNR 10 dB)	Clicks	2	73.53% ACC
					±5°	Noisy (SNR 5 dB)	Clicks	2	31.25% ACC
	4	**9**			±5°	Quiet	Clicks	1	(az) 68.44% ACC (el) 73.09% ACC

ACC–Accuracy; MAE–Mean Absolute Error; RMSE–Root Mean Square Error; SD–Standard Deviation; S–Supervised; STDP–Spike Timing-Dependent Plasticity; (az)–Azimuth; (el)–Elevation.
[1] Use vision. [2] Use body movement. [3] Results for best collective suited threshold V_{sat3}.

The HRD-based model is designed to be compatible with the subthreshold neuromorphic technology. Based on consumptions reported in [45] and simulation in LTspice XVII of the ML *Fast* neurons, as well as the knowledge on the different (pre-)processing components, the power consumption of the proposed coincidence detector can be roughly estimated between 1 nW and 10 nW. The lower bound of the power consumption estimation is computed by considering only the neural network, and the upper bound, including amplifiers. High-pass and eventual band-

pass filtering is performed using passive filters which can be implemented at a negligeable cost in practice.

Currently localization performances are evaluated for an encoding V_{DD} at 400 mV and ideal fast spiking frequency of 1 MHz. Yet, by decreasing the supply voltage, and also having a single supply power source (so V_{DD} common for all neurons), the power consumption can be reduced. Provided in Appendix A.2, the evaluation of the DOA performances in 3-D (click recording set D) with $V_{DD} = 300$ mV shows equivalent DOA estimation results for a temporal resolution of 3.7 µs. Although the ITD extraction precision decreases, localization performances are not affected. The spiking frequency remains higher than the sampling frequency of the recording, therefore the additional error is not visible.

Challenges other than the ones identified in simulation arise in hardware implementations, especially in subthreshold CMOS designs. The circuit is subject to jitter from thermal noise, dispersion effects in the transistors, and variability of and in the electrical components. The final ITD estimation in spike count will certainly be impacted from variations of the expanders duration in output and in the spiking rate of the encoding neuron. However, considering the errors already resulting from incorrect front matching, it is possible that this additional error would compensate for the incorrect detections at times, besides accentuating the estimation error. In fact, it has been observed with the reduction of V_{DD} at encoding; the loss in precision has no effect on the DOA performances.

Additionally, the thresholding method in analog implementation is provided by the abrupt amplification upstream of the detection neurons that only occurs above a threshold equal to V_{DD}/G determined by the amplifier's gain G and supply voltage V_{DD}. At the moment, is it not possible to adjust the threshold of the amplifier once integrated on chip, making the study of the DOA performances' dependance on V_{sat} highly relevant. A comparator would be required, but it involves a power consuming component unsuitable for the low supply voltage of subthreshold CMOS.

Among the works listed in Table 4.9, only [62] reports power consumptions, with their ITD extractor consuming 1.5 mW (scaled according to the number of extractors from the global consumption of 12 mW), a NAS cochlea at about 30 mW, and a SpiNNaker chip consuming between 255 mW and 930 mW according to [62]. Although they announce a future work on an analog CMOS hardware implementation of their ITD extractor, a large power consumption range 1.4 nW ~ 500 µW is estimated. Hence, the proposed model of this thesis has a greater potential for an ULP implementation in comparison. This is further verified by the demonstrator of another neuronal circuit implementing the subthreshold neuromorphic technology which introduced and characterized in the following chapter 5.

4.3.5 Processing Enhancements

At the cost of complexifying the processing and architecture of the whole ITD estimation system, several improvements may be investigated for increase of the performances or application scope.

The current detection method is not dependent on the frequency of the input signal (unlike phase coding systems), thus instead of a half-wave rectification, the full rectification and/or envelope of the signal could be used for a higher resolution of the ITD. Moreover, having a fixed V_{sat} is very limiting in a noisy context. With an increase of the background, the fixed threshold then always detects the noise. An adaptable V_{sat} as a function of the average signal on a given time window would certainly increase the detection and DOA performances. As previously mentioned, an adaptable amplification threshold is not possible in the target ULP analog hardware. Nevertheless, it does not prevent the possible development of this feature in the future for an ULP system. Hence, a short study is available in Appendix A.3 that shows an increase in DOA accuracy with the click 3-D recording set D.

Then, considering the strong dependence on V_{sat} and often incorrect matching of wavefronts, the SSL system could be overloaded with multiple HDR-based coincidence detectors with different V_{sat} values. In other words, by providing multiple ITD estimations for a same binaural pair but with different parameters values, correct ITD estimations have a higher chance of being spotted in a pool of correct and incorrect (following more or less a random distribution) estimations. A subsequent system, like an SNN, would then provide a smarter estimation.

Furthermore, it would be interesting to introduce an ILD extractor that would not directly provide a position estimation but more information on interaural differences. In fact, the radiation pattern of the microphones being different according to the sound's DOA and spectrum, it is not possible to extract meaningful ILDs for integration in the multilateration method. A known HRTF using a dummy head is required and/or an SNN to learn the location and spectral dependency of the microphone's reception. Instead, cues on the relative amplitude difference between two audio may be useful to adjust the final ITD estimate as a correspondence table, or again, using an SNN.

Finally, the present model was evaluated in monosource scenarios, that is only one sound of interest is produced at a time. Pushing even further the complexification of the proposed SSL system, a joint recognition mechanism taking the form of an SNN could be introduced to focus on target sounds. In multisource scenarios, being able to attribute ITD estimations to the sources is highly relevant in a tracking context.

4.4 Conclusion

Using a mathematical resolution involving two to three ITDs, we demonstrated that the overall proposed system could estimate the DOA of click-like sounds in 2-D and 3-D with a high angular resolution enabled by fast-firing neurons of our HRD-based coincidence detector. The system was investigated at very near and far distances, and observed it cannot be used as a distance estimator, a consequence of working with small array baselines and only ITDs. However, thanks to a high ITD resolution of the coincidence detector enabled by the ML *Fast* neuron, the DOA estimation of click-like sounds reached a great precision considering the system's complexity.

Leads on enhancements for increase of the detection and DOA performances reveals additional potential of this model. The threshold detection method is a clear weakness of the model, creating an incompatibility with soft onsets like it is often the case with tonal sounds, but no ULP alternative has yet been identified that is not constrained to low frequency. The evaluation of this DOA estimator was published in [Au2], with click-like sounds only.

The proposed DOA estimator is not fully neuromorphic and it is possible to replace the multilateration with an SNN, as it is often performed in the literature. Instead of following the same direction, however, this thesis takes interest in the less studied case of multisource localization. The HRD-based coincidence detector alone cannot handle a multisource scenario, thus a simple bioinspired recognition mechanism relying on characteristic temporal patterns is chosen to add another dimension to the proposed localization system.

5
Temporal Inter-Pulse Characteristics Detection

The ITD extractor introduced in the previous chapter, and DOA estimator if multilateration is included, has a great potential to provide (near field) localization information on sound sources and to be embedded on chip using the neuromorphic subthreshold technology for a consumption close to the nanowatt. Following these results and since there is rarely a single source in real soundscapes, it is logical that we explore ways to provide a smarter localization, especially if similar power consumptions are reachable.

Multisource scenarios, whether it is by considering several sources of interest or one target source among other sources viewed as interferences or noise, are frequent in application to the biodiversity. Even in the brain, we are naturally processing localization cues jointly with movement or visual information in a multimodal manner, and mostly with our ability to memorize and recognize spectro-temporal characteristics of sounds at a very high level of abstraction in the cortex. In fact, we understand sounds while localizing. Recognition of spectral components and the experience accumulated since birth provide knowledge on the source itself, on the distance, on the environment from the sound's known behavior, or even provide additional localization cues. Moreover, in a multisource context, selective attention mechanisms like the cocktail party effect on speech [112] add enhanced monitoring capabilities.

In this chapter, the interest is given to a recognition task that could be runed concurrently and jointly to the DOA estimator. Whether it is a classification or a detection task, sound recognition remains a subject of significant interest within the neuromorphic community with integration of learning and training methods derived from conventional artificial neural networks and neurological studies. To this day, extensive research is conducted on all kinds of audio signals be it acoustic scenes, animal or anthropogenic sounds, music, or, most commonly, speech [113]. Efforts are mostly concentrated on deep learning-based solutions where recurrent and/or convolutional architectures are predominant [114].

Neuromorphic recognition systems have been proposed whose architecture and preprocessing take into account more or less complex acoustic datasets with varying temporal patterns, spectrums, and environment effects [115], [116], [117], [118], [119], [120], [121], [122], [123], [124]. These solutions explore the precision of spiking neural networks (SNN) in comparison to conventional artificial neural networks, but few report energy costs or power consumptions. Efforts are concentrated on evaluating new or adapted conventional methods on neuromorphic architectures to extend reveal the computational potential of the paradigm.

Among the works describing energy-related metrics and by way of context, Martinelli and al [120] performed in 2020 voice activity detection with a recurrent SNN and provided an estimation of the energy consumed by the networks synaptic operation based on the neuromorphic processor TrueNorth [28], leading to a total power consumption ranging between 80 and 105 mW depending on the SNN's optimization (between 25.1 and 33 µW without considering the static consumption, the latter being preponderant). Also for voice detection, Dellaferrera and al [121] proposed an SNN whose lower bound power consumption of about 3.8 µW was estimated according to the specifications of the neuromorphic processor Loihi [29]. Wu and al [122] carried out a speech recognition using an SNN for which its relative energy cost was compared to that of a conventional network knowing the number of synaptic operations executed and their correspondence to the processor's actual digital operations. In the end, only a ratio was reported as an hardware implementation would be necessary to evaluate an energy consumption. Then in 2021, Bensimon and al [123] used digital neurons described as ULP with power consumptions below 10 nW to extract acoustic features and classify sounds, but did not report an power consumption estimation of their system or core processing although they announce a better energy efficiency than the TrueNorth processor. Finally in 2024, Yang and Chang [124] developed an accelerator for acoustic signature extraction and performed a speech recognition task with a recurrent SNN. A power consumption of 71.2 µW is reported in the least consuming configuration of the digital accelerator (up to 35.5 mW with highest clock frequency for lowest latency), enabled by in part by a 0.8 V supply voltage and a high optimization of the underlying processing in a 28 nm CMOS process.

However, these consumptions are not sufficient to establish the actual benefit of the spike representation. In fact, a similarity of these propositions is a digital processing mostly thought for a generic digital processor. Apart in [120], [124], only dynamic consumptions are put forward while static consumptions should not be disregarded. Indeed, neuromorphic processors have rather high static power consumptions; TrueNorth has typical workload of 65~70 mW, and Loihi's idle power consumption reached 30 mW according to a keyword spotting benchmark [125].

Moreover, the proposed SNNs have a large number of neurons for appropriate learning, with more than 100 neurons and 2000 weights or network parameters, up to about 200k parameters in [124]'s pruned SNN. The analog neuromorphic technology used here for all implementation and considerations, however, still require a rigorous study and evaluation in networks of plastic synapses developed by IEMN's team for the design of capable small sized SNNs. In fact, densely connected networks and/or high number of neurons implies a high number of synaptic weights. With limited silicon surfaces, the number of pads is also limited, being the largest

element on the chip. Yet, pads are mandatory to apply weight voltages, thus using fixed synapses limits the number of weights. Although weights could be quantized to reduce the number of pads, it impacts the learning capacity of the network. Since it still needs a substantial work, this lead was therefore not chosen.

Instead, the focus is given to a compact and very simple architecture able to discriminate sound sources by relying on their characteristic temporal features, or more specifically on characteristic rhythmic features. Precisely synchronized sound sequences, such as successive songs, cries, and bursts, or rhythmic noises, help identify the origin of a sound within biodiversity but not only [126]. In birds and singing insects in particular, the recognition of an individual of the same species relies largely on the spectro-temporal characteristics of calls and songs. After investigation, the recognition mechanism in female crickets at the neural level of the delay between sound bursts (that are, stridulations) emitted by male crickets was chosen, and modelled through a simple circuit by Schöneich and al [127] in 2015. This neural network features the advantage of being adaptable to any signal with a sparse architecture, in a way similar to the ITD extractor previously studied in the chapter 4.

The model from [127], which held our attention, was previously studied by Sandin and Nilsson [128] in 2020 where it was adapted for a hardware implementation in the neuromorphic mixed-signal processor DYNAP-SE [34]. This processor combines digital interfaces, routers, and memories with sub-threshold analog neurons. With the motivation of reaching ULP consumption and demonstrating the potential of this hardware, the authors reproduced the delay recognition mechanism using excitatory-inhibitory disynaptic elements, later investigated for the design of a spatiotemporal correlator in a spiking neural network [129]. Although the authors reported the primary energy consumption associated with the circuit's core operations, the additional power consumption arising from the use of mixed-signal hardware was not addressed (especially the processor static consumption, operating under 1.8 V).

As it is later analyzed in this chapter, the subthreshold neuromorphic technology has the potential to reduce by at least a thousand the energy consumption of [128] by implementing the adaptation of the neuronal circuit in a fully analog integrated circuit with a 300 mV power supply. The inter-pulse delay detector is described and evaluated with ideal and real-world sounds signals. Then, its biggest limitation is investigated, namely the manual tuning of the weights, and further processing is proposed suitable for automatizing the tuning.

5.1 Overview

Starting with a strongly bioinspired design, our circuit was later adapted to comply with constraints related to the subthreshold neuromorphic technology. Starting with the description of the biological mechanism and neuronal circuit, the proposed circuit is then described which was implemented in a demonstrator for a complete evaluation on-chip. It may be noted that only one demonstrator was produced during this thesis, and this circuit was chosen as a more suitable candidate than the ITD extractor. Indeed, it does not require post-processing for temporal feature detection while several ITD extractors and a multilateration is necessary to estimated DOAs. Besides, more certainty of a successful feature implementation was

determined, and the architecture closer to biology was more aligned with the bio-inspiration aimed in this thesis. Moreover, the existing implementation of a similar adaptation in similar mixed-signal neuromorphic processor [128] provided a clear comparative base for highlight of the subthreshold neuromorphic technology's ULP consumption.

5.1.1 Biological Temporal Delay Detection Model

With the objective to unravel the complete neuronal process by which some animals are able to selectively respond to temporal pulse patterns from their specie-specific acoustic signals, [127] studied female Mediterranean field crickets (*Gryllus bimaculatus*) that localize males by recognizing the characteristic pulse period of 30 to 40 ms in the succession of 3 to 5 pulses with frequencies ranging from 4.3 to 5.2 kHz. Unlike humans or mammals, field crickets possess hearing organs in their front legs. Tympanic membranes on the tibias and spiracles of the auditory trachea on the abdomen both receive the sound for encoding into neuronal impulses by tonotopically arranged auditory afferent neurons of the hearing organ. Laser measurements of the mechanical oscillations of a tympanic membrane showed a band-pass filter with best response at 5.3 kHz [130]. The frequency tuning of the afferent neurons does not follow a continuous distribution but instead three quarter of these neurons are tuned to the male calling song with the remainder tuned to the male courtship song and the echolocation ultrasonic signal of bats.

Among the several connections made to various neuronal populations, afferent neurons make synaptic contact to an ascending auditory interneuron denoted AN1, which deliver alone information of male calling songs to the female's brain thanks to its corresponding frequency tuning. In this manner, AN1 can be considered the entry neuron to the model described thereafter. Previous to [127]'s study, AN1 and three local auditory neurons, LN2 to LN4, were identified in the anterior protocerebrum of the brain in female field crickets to which [127] added the discovery of a non-spiking local neuron, denoted LN5, from intracellular recordings of spiking activity in response to controlled stimuli of AN1. In [127], the paper's fifth figure schematizes the biological model and the identified neuronal processing in the case of no detection and a recognized pattern. Extensive biological explanations about neurological and statistical details are available in [127] and [130]. The delay recognition process depicted in the paper's figure is summarized and qualitatively described for an ideal scenario.

AN1 receives the auditory stimuli from the afferent neurons and passes half of its spiking activity to LN2, such that LN2's first spike is delayed by one spike with AN1's onset spike. Forming an inhibitory connection with LN5, LN2's activity produces a hyperpolarization of LN5's membrane potential that induces a delayed depolarization (a rebound) reaching a maximum of 5 mV above the resting potential. Moreover, AN1 provides an excitation to LN3 of a topology such that the integrated energy is slowly accumulated or leaked. One spike train from AN1 thus raises LN3's membrane potential to its spiking threshold and to which the membrane potential remains close due to its slower leak compared to the other neurons. Simultaneously, LN3 receives the delayed depolarization from LN5 providing additional energy as a consequence, and when timed to the onset of a second auditory stimulus at AN1, it provides sufficient energy for LN3 to produce additional spikes.

Finally, LN3 having an excitatory connection with LN4, the neuron informing on a correct characteristic pulse period detection, these additional spikes originating from the rebound provides enough energy for LN4 to spike and identify the second stimulus as being at the correct delay from the first stimulus. Inhibition from LN2 in this model is required to prevent detections at delays shorter than the characteristic delay. In fact, with continuous excitation of LN2 and thus LN3 by close or fused spike trains at AN1, LN4 would produce spikes from accumulated energy. The inhibition from LN2 may then suppress the excitation from LN3's spikes that are not timed with LN5's rebound. Naturally, with longer auditory pulse intervals than the characteristic delay to be recognized, the first spikes of the second stimulus and the leaking membrane potential at LN3 do not suffice to produce spikes at LN4 before its inhibition by LN2.

All described neurons, to the exception of LN2 whose processing is more intermediary to the other functions, have a specific role: AN1 is the auditory pathway, LN5 the delay line, LN3 the coincidence detector, and LN4 the feature detector. Combined, this model explains the pulse period detection mechanism in female crickets. Inter-pulse delays lower or higher than the characteristic delay detected by the rebound do not generate spikes at LN4.

In their intracellular experiments for characterization of the proposed biological model, the authors stimulate AN1 with square shaped excitation of 20 ms duration and varying inter-pulse delays. A standard chirp carrier frequency of 4.8 kHz was evaluated as the specie-specific inter-pulse delay to be recognized by the female cricket, resulting in characteristic delays of 20 ms at which LN4 was then observed be the most sensitive.

5.1.2 Proposed Circuit

These ideal biological observations and descriptions provide a well-defined signal processing replicable artificially through slight adjustments. The spiking characteristics of most the neuron (except LN5) suggest the use of the ML *Base* neuron. The transistors and capacitances dimensioning described in [45] does not quite match the slow spiking frequency of the biological model. Therefore, the electrical components were modified such that an even slower spiking neuron would be obtained, then called the ML *Slow* neuron (see section 2.3.1). The possible signal processing is constrained by the ML formalism of the artificial subthreshold neurons due to the simplifications from the complex HH model to a more compact and efficient design.

The adaptation of [127]'s model implemented in this thesis is proposed in Fig. 5.1a with the corresponding neurons activity depicted in Fig. 5.1b,c. The weights of the synapses are identified using a suffix on the synaptic weight notation Wx_{ij} such that "x" is either "e" for excitatory or "i" for inhibitory, and "$\_ij$" refers to the pre- and post-synaptic neurons' identifier. Expanders weights are denoted $W_E x_{ij}$ to specify to which synapse they are linked. The only leak is briefly noted Wl.

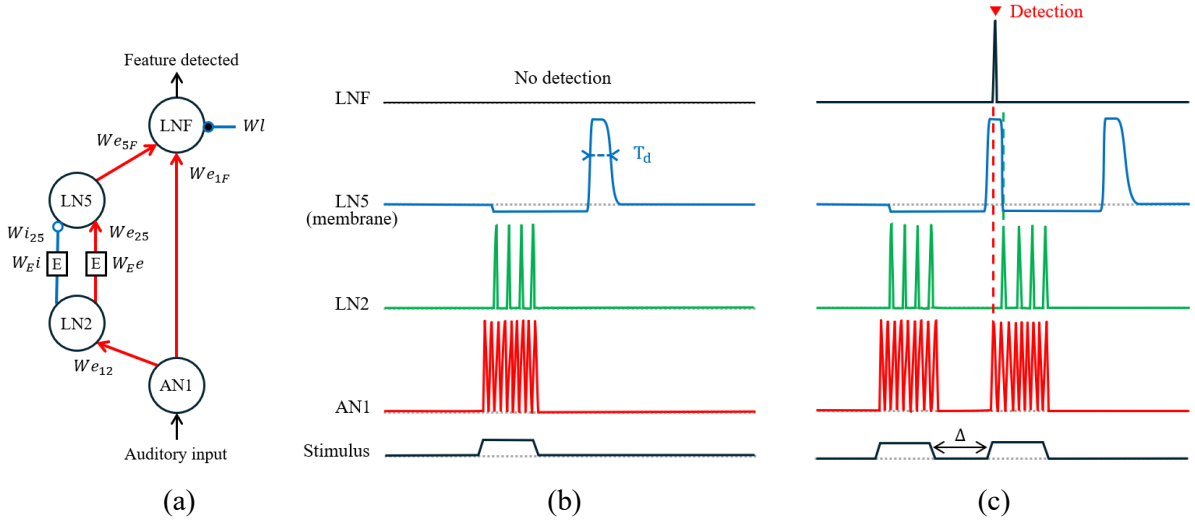


Fig. 5.1 Proposed inter-pulse delay detector. (a) Neuronal circuit adapted from Schöneich and al model of inter-pulse delay detection. Processing performed by the adapted circuit in the case of (b) no delay detected and (c) a single occurrence of the characteristic delay. Except at LN5 that shows the membrane potential (between 0 V and around 115 mV, the potential of the other neurons traced is the generated spike activity (between 0 V and V_{DD}). The delay detection period T_d is highlighted at LN5, and the characteristic delay Δ to be detected is marked on the stimulus waveform. The concurrence of AN1 first spike(s) and LN5 rebound allow LNF to spike.

The processing performed by the proposed circuit is roughly similar to the biological in that it reproduces the delay line, the coincidence and feature detection, but with slight differences attributable to the analog neuromorphic technology used. In order to obtain large time constants and mimic the slow evolution of LN3's membrane potential, larger neuron capacitances would be required, but it would also change the behavior of the analog neuron while increasing the power consumption and die size of the circuit. Consequently, it was decided not to reproduce the slow evolution and, to overcome this situation, to fuse LN3 and LN4. As a result, the coincidence and the feature detection occur at the same time. To avoid any confusion, this neuron is therefore denoted LNF as in “feature”, while the other neurons are named following their biological notation. With this information, the processing is performed as follows.

AN1 receives preprocessed input sounds through a transconductance (conversion of the stimuli from voltage to current driven to neuron), encoding spikes. In the case of an ideal study, the square stimuli are directly fed to AN1. For audio files, the acoustic signals are either kept raw or band-pass filtered with a Butterworth filter type of order 2, central frequency 4.8 kHz with a bandwidth of 300 Hz using the SciPy (v1.11.4) library in Python (v3.9.18). These parameters were chosen according to observations of the chosen audio files of male field crickets calling songs whose central frequency was 4.8 kHz on average. Its order is defined so as to obtain a rather simple filter, and we use a shorter band-pass than the frequency range of cricket calling songs detailed in [130] for better attenuation out of the bandwidth. As further preprocessing, the envelope of the half-wave rectified (non) filtered acoustic signal is then extracted using a R-C filter with time constant of 1 ms for easier processing by the detection circuit. It also allows to remove possible abrupt and short impulses from non-filtered signals.

Lastly, the preprocessed signals are amplified and saturated to V_{DD} (in some cases higher amplification is needed for AN1 to generate spikes).

After spike encoding by AN1, LN2 generates spikes with a frequency halved compared to AN1 by weakening the excitatory synapse We_{12} to create a latency with reference to AN1's bursts. In our hardware circuit, this latency between the onset of AN1 bursts and the onset of LN2's is essential to enable any delay detection, otherwise, the inhibition path (Wi_{25}) would suppress the rebound influence at the time of a successive stimulus. The weight Wi_{25} is set to V_{DD} since no balance is meant to be achieved but a complete inhibition of LNF when both the rebound and successive stimulus occur.

The key element for a delay recognition mechanism, namely the membrane potential dynamics of LN5 inherent to the non-spiking neuronal cell, is not easily reproducible with the current architecture of the ML neurons. In order to keep the same neuron dynamics (same topology) throughout the network, additional electronic elements were used, namely the expanders prior described (see section 2.3.2). The expanders permitted concurrent excitation and inhibition but applied with different durations so as to precisely generate the post-inhibitory rebound in the temporal domain (provided by LN5 membrane potential). Unlike the rebound observed for LN5 in biology, our hardware implementation uses a more clearly defined rebound. It allows us to precisely set the moment from which the short time window T_d occurs (Fig. 5.1b). This window essentially corresponds to the detection selectivity of the circuit and include the characteristic inter-pulse delay Δ to be detected. The lower and upper bounds of T_d are chosen by adjusting the expanders' duration tuning parameters $W_E i_{25}$ and $W_E e_{25}$ respectively. LN5 is kept non-spiking by preventing its membrane potential from reaching the firing threshold, in other words by limiting the excitation path using We_{25}.

At LNF, simultaneous excitation by LN5's post-inhibitory rebound and first spikes of AN1's spike trains (sound pulses) is the only condition in the proposed circuit to generate spikes. In the same manner that LN5 is kept non-spiking so LNF do not generate false detections from the rebound alone, We_{1F} was tuned such that the excitation from AN1 alone would not generate spikes. It is worth noticing that depending on the latency of LN2's first spike, LNF may produce one spike or a burst of two spikes during a correct detection. Compared to the initial architecture, an inhibition from LN2 to LNF was deemed unnecessary because the rebound already receives inhibition from LN2 at LN5. However, in order to experimentally ensure correct detections of the characteristic delay, a leak voltage Wl, set to a NMOS transistor for which the source is connected to ground and the drain to LNF membrane as depicted Fig. 2.10c, is added to LNF to provide more flexibility.

In addition to the above description, one important point is related to the singular excitatory synapse We_{5F}. Indeed, all excitatory synapses in the circuit behave as described in section 2.3.2 but synapse We_{5F}. The reason for which is that LN5 never emits a spike. As a result, its unique output is directly its membrane potential (waveform shown in Fig. 5.1b,c). Hence, in order to provide the extra excitatory current allowing a detection by LNF when a successive stimulus occurs, LN5 membrane voltage potential is simply converted using a transconductance (that is, a single transistor whose gate node is controlled by the membrane voltage potential of LN5) to an excitatory current driven to LNF neuron.

5.1.3 Materials

Design and simulations on LTspice demonstrated the correct operation of the proposed circuit with the analog subthreshold technology. After validation and further chip design on Cadence by the team's research engineer, a demonstrator of this simple circuit was produced for on-chip evaluation.

Hardware and Equipment

The circuit was characterized and tested using a probe station. Probes are placed onto pads (square shape with side width equal to 60 µm), allowing the application of the various circuit voltage inputs. Emulated and real sound signals are configured using a software, and generated by a National Instrument voltage generator to be applied as input for AN1. Other voltages, including excitatory weight and V_{DD} supply voltage, are supplied by Keithley source generators / meters (source measurements units 2600B and 2600A series). Capture and visualization of outputs were performed on a Tektronix oscilloscope. Fig. 5.2 shows the micrograph of the chip and the probe station test bench.

In total, the chip has 22 pads (or pins) as depicted in Fig. 5.2c. It leads to a total die size of 6400 µm² (80 × 80 µm²) and 0.389 mm² (1080 × 360 µm²) approximatively for the functional circuit and including pads, respectively. Among the pins, one is not connected to the circuit, and the rest correspond to the supply current $I_{SS\ ANA}$ and voltage $V_{DD\ ANA}$ of the analog output $LN5_{mem}$, namely the membrane potential V_{mem} of the neuron LN5. The latter three cannot be used because a component was added to prevent the circuit from being damaged by electro-static shocks at the pins, but also resulted in incompatibility with the circuit supply voltages. In fact, visualizing the analog output would render the circuit non-operational. Its use was thus abandoned, and instead the digital buffer output of LNF ($FLAG$) was used to indirectly visualize the rebound generated at LN5.

On the pin layout, supply voltages are highlighted in red, inputs and outputs in blue, and manually tunable weights in green and light grey. A difference is made between the green and light grey parameters to indicate the parameters that can be fixed voltages for a definitive version of the demonstrator. In order to prevent the situation where none of the test chips worked, it was designed with all weights tunable, except the inhibitions that all needed to be maxed. Pads with a notation that includes an *"n"* are described in a second phase, introduced in section 5.3 as a possible enhancement of the circuit's tuning, and will not be discussed until then.

Incidentally, it is to be noted that the neurons and synapses dimensions follows the ones disclosed in Table 2.3, but the capacitance defining the duration range of the expanders C_E is set to 145.6 fF for all expanders in this specific circuit. This leads to a maximum ON duration of the expanders of about 190 ms.

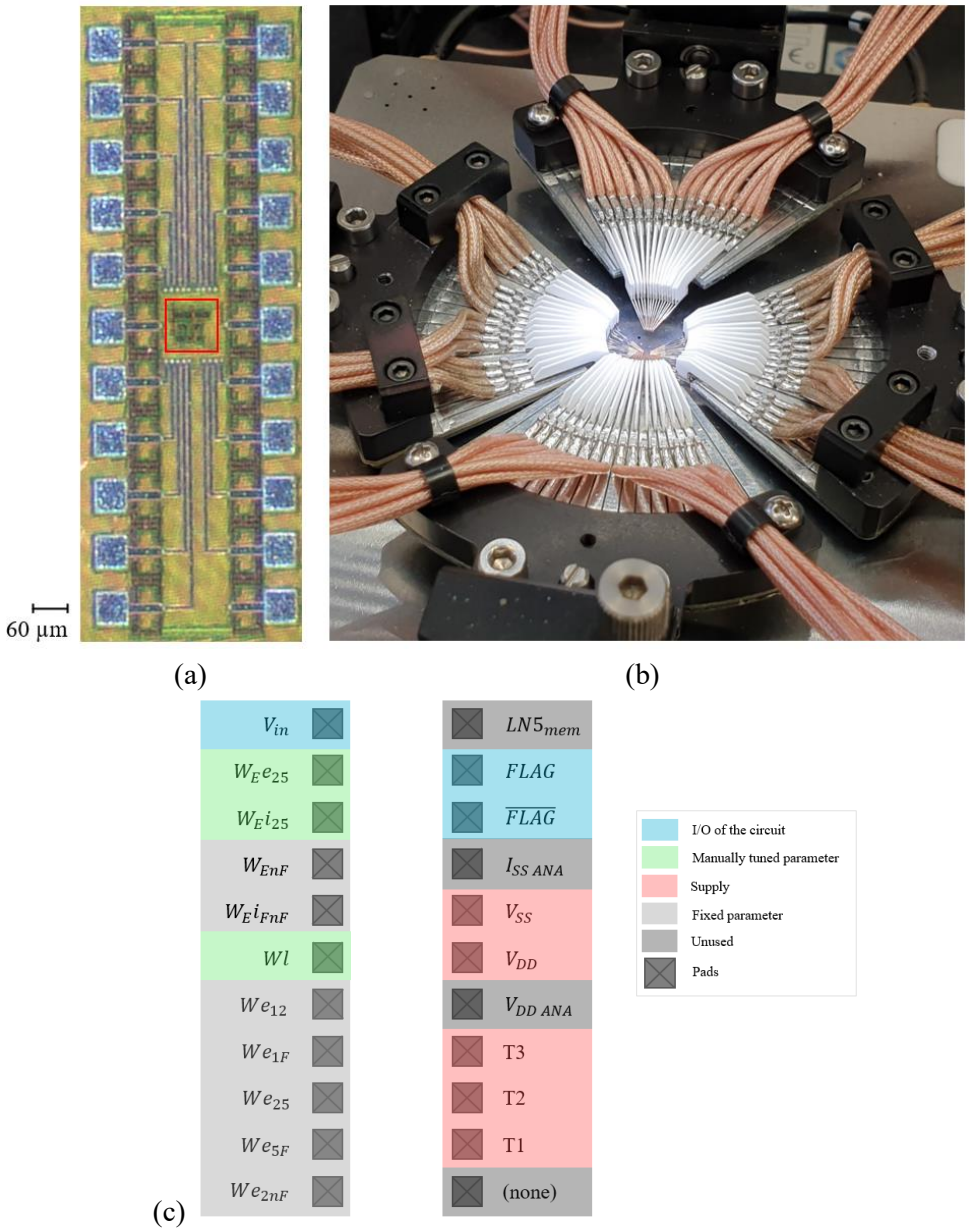


Fig. 5.2 Pictures of (a) the delay detection circuit, where the actual circuit without the pads is highlighted by a red square, and (b) the probe station test bench. A pad has dimensions 60 µm × 60 µm. (c) Pin layout of the inter-pulse delay detection circuit. The supplies Ti are used to record digital outputs on the oscilloscope by amplifying the signal maximum voltage from 300 mV to 800 mV (through a succession of inverters). Except for $I_{SS\ ANA}$, (unused) all pads are voltages input or outputs. V_{SS} is connected to ground.

Power Consumption Measurements

The most crucial contribution of this chapter is the energy efficiency of the demonstrator which is determined from its DC power consumption. In order to resolve the DC power consumption, the circuit depicted in Fig. 5.3 is considered.

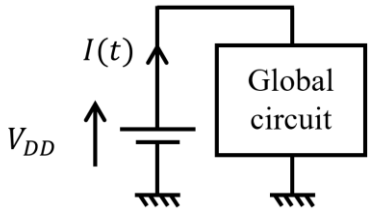


Fig. 5.3 Simplified representation of the circuit on the chip demonstrator. The inter-pulse delay implemented on the chip is represented by a block supplied by one continuous supply voltage V_{DD} and connected to the ground. $I(t)$ is the instantaneous current delivered by V_{DD} and drawn by the global circuit.

The supply voltage V_{DD} is applied to various sub-circuits previously depicted: neurons, excitatory synapses, and expanders. It is to be noted that other voltage sources are used in order to set the synaptic or expanders' duration weights; nevertheless, these are all applied onto the gate of transistors (see Fig. 2.10 and 2.11), therefore across transistors input capacitor. Hence, the DC current flowing out of these voltage sources is zero (null) and do not affect the DC power consumption.

Considering the circuit shown in Fig. 5.3, the overall power consumption is determined through the following equations, with respect to a general case where the current delivered by V_{DD} varies as a function of time:

$$p_i(t) = V_{DD}\, I(t) \tag{5.1}$$

$$\mathcal{P}_{DC} = \frac{1}{T}\int_0^T p_i(t)\, dt = V_{DD}\int_0^T I(t)\, dt = V_{DD}\, \bar{I} = V_{DD}\, I_{DC} \tag{5.2}$$

where $p_i(t)$ stands for the instantaneous power, $\mathcal{P}_{DC}$ the DC power, and $\bar{I}$ the mean (or DC) current observed over a sufficiently long time period T. $\mathcal{P}_{DC}$ is actually the power which is delivered to the circuit (that is, delivered by a supply voltage or a battery). Two scenarios are possible: (i) no signal is applied to the circuit, thus $\bar{I}$ corresponds to the overall DC leakage current (that is, the total amount of transistor leakage currents) which sets the "standby" power; (ii) a signal is applied to the circuit which leads to a time varying current $I(t)$ and an increase of the power consumption. In the latter case, the power consumption corresponds to the aggregation of two contributions: the static (global circuit in standby mode) power to which is added the dynamic power.

The overall power consumption is then experimentally measured from DC current $\bar{I}$ measurements operated through a Keithley 2636A source generator observed on a time period T of at least 4 s. It is to be stressed that this device features a suitable current measurement resolution down to a 100 pA range (basic accuracy of the current measurement of 0.15% + 240 fA on a measuring range of 1 nA).

Test Sound Signals

In order to accurately test the circuit, largely different input sounds were tested: artificial signals that mimicked the tests performed on the initial biological signals in [127], and real-world acoustic emissions of the original cricket specie recorded in quiet, noisy, and multisource scenarios.

Firstly, the delay detection circuit was tested using ideal square signals featuring 20 ms pulse width (Fig. 5.4a), similar to [127]'s experimentations and with varying inter-pulse delays around Δ to which the circuit must be tuned. The signals were either applied continuously or by limiting to 2 or 3 repetitions of the pulses. The feature was also tested with pulse trains of 1 ms pulse width and 20 ms train duration (Fig. 5.4b) with similar inter-pulse delays.

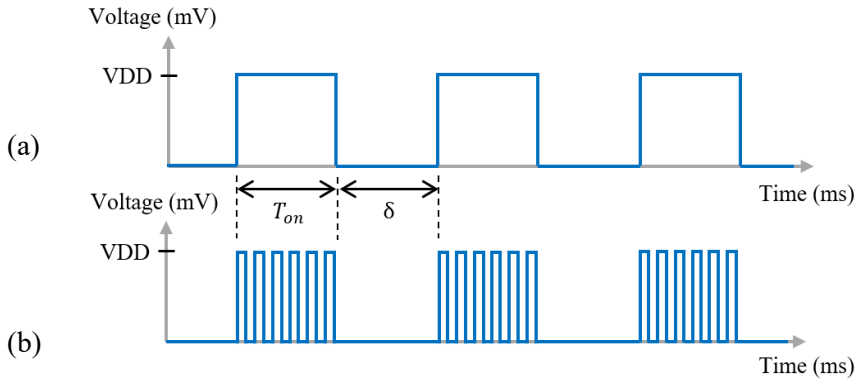


Fig. 5.4 Ideal (a) square pulse and (b) pulse train signals. The duration of the artificial calls emulated by pulses have T_{on} at 20 ms unless specified otherwise. The inter-pulse delay denoted δ on the figure varies around the characteristic delay Δ to be detected.

Secondly, real sound signals raw or preprocessed in Python were loaded as samples files to be reproduced by the voltage generator with sampling rate of 44.1 kHz and 48 kHz. These are audio recordings from the dataset Xeno-Canto available online [111] of calling songs identified as originating from male field crickets (*Gryllus bimaculatus*) in varying conditions. Recordings initially sampled at 96 kHz were resampled to 48 kHz by removing one in two samples. Besides, the recordings were trimmed to the first 150k samples, corresponding to about 16 seconds, since the voltage generator can only handle up to 150k samples.

Table 5.1 resumes the characteristics of the chosen recordings from the Xeno-Canto dataset. The measures in this table were made using Audacity, and in particular the tool *Contrast* was used to compute SNRs. The range or value of the SNR is reported considering the quietest and loudest call in the raw and filtered acoustic signal. Two SNR ranges are reported for multi-sources scenarios, with reference to overlapping sound from non-cricket sources and/or to background noise only, unless the difference is negligible. Fig. 5.5 shows the distribution of the call period on the studied duration of the recordings[5] for all identified cricket calling songs (two singing crickets in XC854026). Depending on the preprocessing and the call period, the characteristic inter-pulse delay Δ may vary, and therefore corresponds to a range of delays unlike with the ideal inputs.

In addition to fully preprocessed wave forms (band-pass filtering and extraction envelope), the raw wave forms were tested to identify the signal integration ability of the circuit. Fig. 5.6 gives the different preprocessed states that were tested, taking as an example the recordings XC867042 and XC922939 respectively. The sound and its processed signals are normalized

[5] XC867042, XC854026, XC853470, XC854155 are under CC BY-SA 4.0; XC922939 is under CC BY-NC-SA 4.0.

according to the maximum absolute value of the initial recording. Besides, some signals were too low for AN1 or LN2 to generate spikes. These signals were thus amplified and saturated to V_{DD}.

Table 5.1 Characteristics of the chosen recordings on the studied duration.

Sound recording id	SNR (dB RMS) Raw signal	SNR (dB RMS) Filtered signal	Calling song main frequency (kHz)	Number of calls per song	Multi-sources
XC867042	33	45	5.05	3	No
XC922939	7.5 ~ 9	8.5 ~ 14	4.83	3	No
XC854026	0.5 ~ 11 * -0.2 ~ 3 **	10 ~ 27 * 3 ~ 21 **	4.78 (cricket 1) 4.6 (cricket 2)	3	Birds, cricket, insects
XC853470	3.5 * -0.2 ~ 3 **	17 ~ 32	4.7	4.6 (mean) 0.57 (stddev)	Insects
XC854155	0.5 ~ 13 * 0 ~ 7 **	10 ~ 35	4.8	3	Insects

RMS–Root Mean Square; stddev–Standard deviation.

* With reference to the noise only.

** With consideration of other overlapping sound sources.

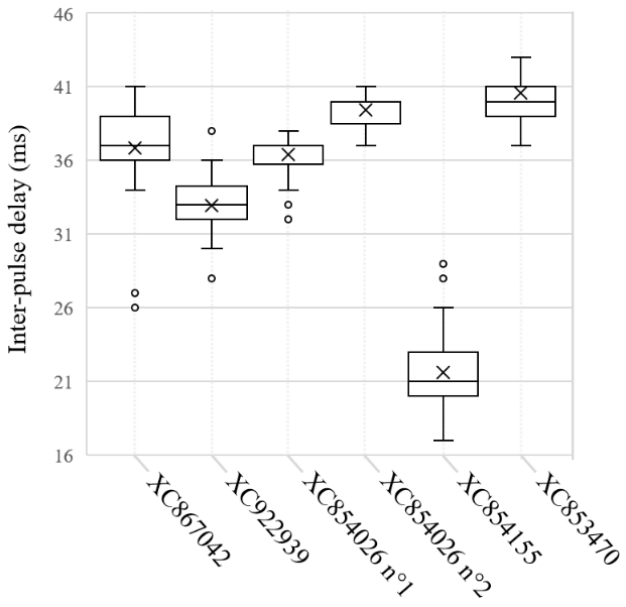


Fig. 5.5 Distribution boxplot of call period per test recording (up to the 150k-th sample) in a song. Recording XC854026 is segmented into the two identified singing field crickets. The boxes show the quartiles, medians marked by lines within the boxes, averages marked by crosses, and extremes marked by circles.

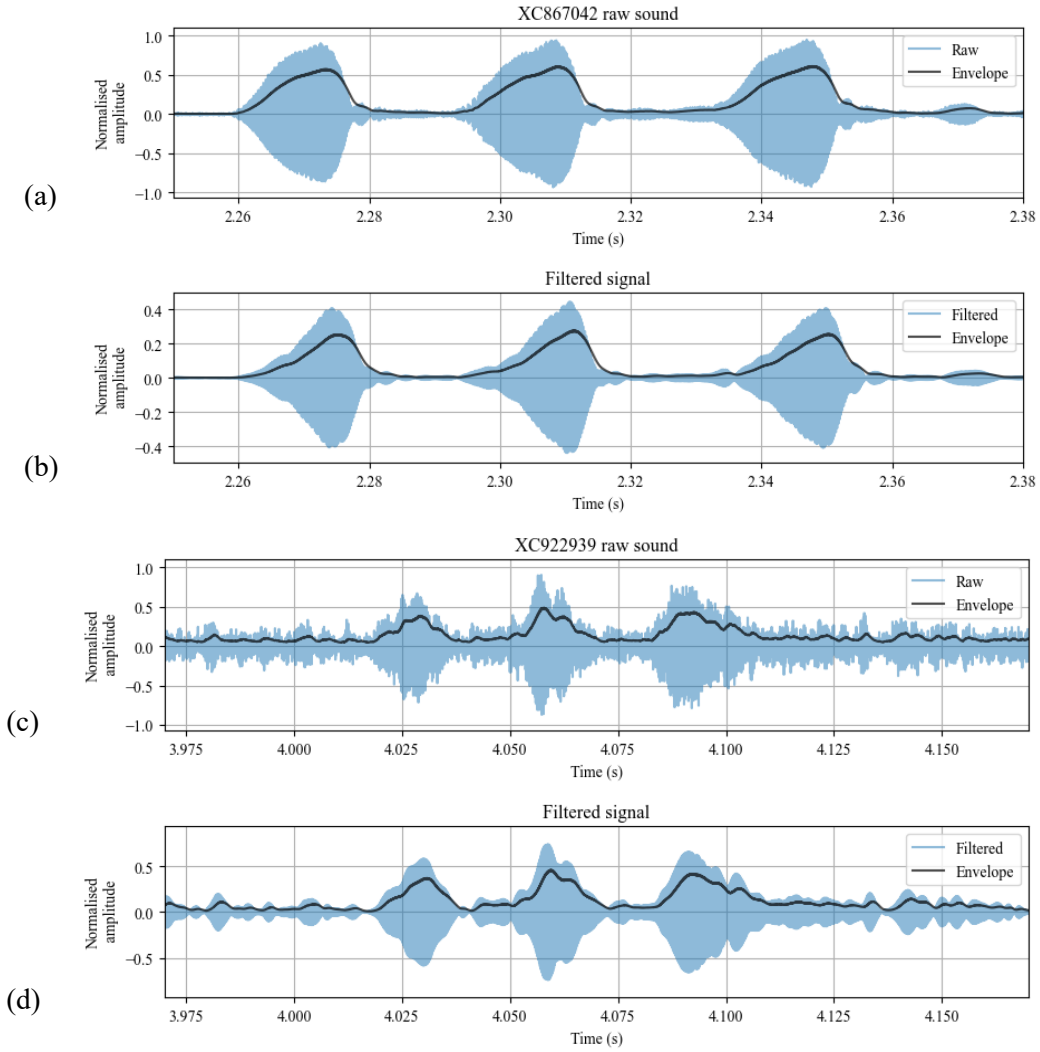


Fig. 5.6 Recordings (a) XC867042 and (c) XC922939 and their preprocessed signals in (b) and (d) respectively. A zoom is made on one calling song. Reproduced from XC867042 CC BY-SA 4.0, and XC922939 CC BY-NC-SA 4.0.

5.1.4 Calibration of the Hardware Circuit under Probe Station

Because neurons membrane potential availability under the probe station was reduced to the generated sound signal input and the feature detection of LNF's output (through a digital buffer, or $FLAG$ of the pin layout), a calibration procedure was carried out for the manual tuning of the synaptic weights. Depending on how the synaptic weights are set, the spikes generated by one neuron can be visible at the output neuron LNF. All neurons are directly or indirectly connected to LNF, so by completely opening (weight at V_{DD}) or shutting (weight connected to ground) excitatory synapses, we can propagate a neuron's spiking activity to LNF and correctly tune the weights one at a time. From AN1 to LNF, the direct path and the one passing through LN2 and LN5 can be tuned separately. The leak Wl of LNF is first set at a non-zero value to be later adjusted.

By opening the synapses of subsequent neurons of a same path and shutting the synapses of the other path, We_{12} is tuned to halve the spiking frequency of LN2 with reference to AN1's

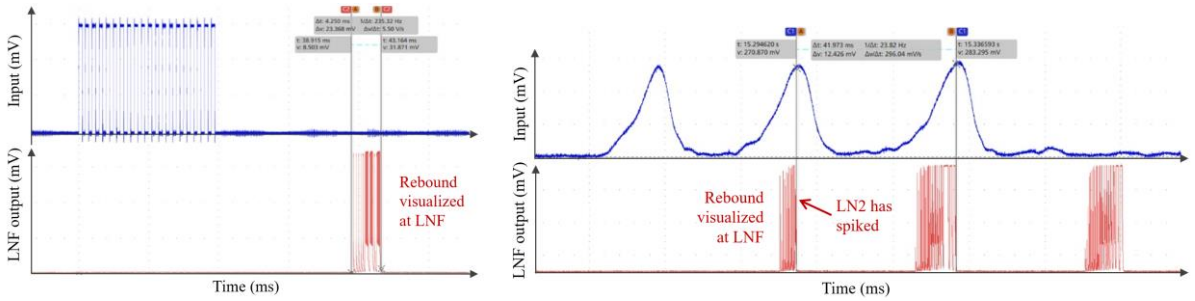
activity and We_{1F} to bring the pre-synaptic spike amplitude close to the spiking threshold of LNF.

Once We_{12} tuned, the spikes generated by LN5 with We_{25} at V_{DD} is exactly the rebound and its detection period T_d. Knowing Δ, the weights $W_E e_{25}$ and $W_E i_{25}$ of the expanders can be tuned to adjust T_d and its latency to the input. To illustrate the calibration process, Fig. 5.7 shows a view of the rebound observed at LNF on the oscilloscope for an ideal and a real sound input during this calibration step. LNF spikes rapidly on the duration of the rebound, revealing the rebound's position in time with reference to the input for correct tuning of T_d.

Then, We_{25} is adjusted to make LN5 non-spiking by decreasing the received excitation below the spiking threshold in the same manner as We_{1F} tuning.

Finally, the tuning is tested by applying successive square pulses with inter-pulse delays around Δ. The number of spikes generated by LNF when the inter-pulse delay matches with T_d is minimised by decreasing the leak. Modifying the leak has for objective to optimise the detection to various noise conditions by moderately facilitating or inhibiting LNF.

The calibration can also be performed directly on real audio recordings by observing the last repetition and all sequence of a calling song for the calibration and testing stage respectively. The parameters offer a great flexibility and a wide range of weight combination. In the end, the most decisive steps of the calibration are the tuning of the expanders and the leak.



(a) (b)

Fig. 5.7 Visualization at LNF of LN5's rebound on oscilloscope for calibration of T_d in the case of (a) an ideal square pulse, and (b) the envelope of the non-filtered audio recording XC867042 (adaptation from XC867042 CC BY-SA 4.0). The rebound is fully observed when only one pulse is given in input, or at the last call of a song such that no following call may inhibit the rebound like in (b). The first rebound in (b) is shorter than the following two rebounds because LN5 receives inhibition from LN2's spikes (reflecting AN1), allowing the observation of the temporal coincidence between LN5's rebound and LN2's (or AN1) spikes generated by the calls. The input voltage provided to AN1 as excitation is in blue, and the output of the digital buffer of LNF in red. Cursors in (a) indicate the width of the rebound corresponding to $T_d = 4.25$ ms, while in (b) they indicate the delay of 41 ms between two calls of the cricket calling song in input.

5.2 Feature Detection Results

In output of the circuit, the detection should be effective only during T_d in coincidence with a following spike train. Detections outside these conditions are considered false positives (FP)

in the same manner that correct detections and incorrect absences of detection are considered true positives (TP) and false negatives (FN). The detection performance of the circuit is evaluated using the precision and the recall calculated from $\frac{TP}{TP+FP}$ and $\frac{TP}{TP+FN}$ respectively.

5.2.1 Ideal Sound Stimuli

Reproducing the experiment input in [127], AN1 is stimulated with a square wave signal to generate successive spike trains. The parameters are tuned to attain 100% recall and precision with the most selective T_d for all ideal stimuli. These signals were studied to characterize the circuit since their duration and period could be easily modified. Overall, detection was successfully made using ideal square pulses and burst-like pulse trains.

The width of T_d remained invariant with the duration of the pulses or length of spike trains, leading to the correct detection of a same inter-pulse delay instead of a pulse period. Observed under probes, the shortest excitation duration of AN1 possible was about 4 ms for LN2 to generate spikes and LN5 to produce a rebound. Naturally AN1 could receive uninterrupted excitation for continuous detection. The minimum value of Δ that can be detected by the chip is 3.4 ms with $T_d \cong 2.5$ ms, and the maximum is 187.4 ms with $T_d \cong 7$ ms. For Δ = 20 ms, a detection selectivity $T_d \cong 4$ ms is achieved. Depending on the variability of the expander duration, T_d varies more or less in time with accentuated instability of T_d at larger Δ, but its value for Δ = 20 ms and a target recall of 100%, as an example, does not exceed 7 ms.

The detection of Δ is immediate in that the circuit detects Δ directly from the second call. In the case of field crickets, only 3 to 5 successive calls are produced, as highlighted in Fig. 5.5, but the circuit can perform continuous detection with the same precision and recall as long as the calls and soundscape conditions remain similar in time.

In order to evaluate the precision and recall of the system, the weights were calibrated according to the procedure aforementioned, and set to an integer value between 0 V and V_{DD}. Several sets of weights were tested, but the combination detailed in Table 5.2 was prioritized. The weights We_{12}, We_{1F}, and We_{25} could be pooled together to 200 mV, with the remaining fixed weight We_{5F} (and Wi_{25}) set to V_{DD} and the leak Wl set to 60 mV, while retaining the circuit's functionality. It is judicious to have fixed weights set to the same voltage since it reduces the number of transformations required, and thus the number of electrical components. In hardware, fixed voltages are obtained from the supply voltage using a voltage divider. The values detailed in Table 5.2 were also used with the real-world sound recordings.

Table 5.2 Pooling of the fixed weights.

Weight	Voltage (mV)
We_{12}	200
We_{1F}	200
We_{25}	200
We_{5F}	V_{DD}*
Wl	60

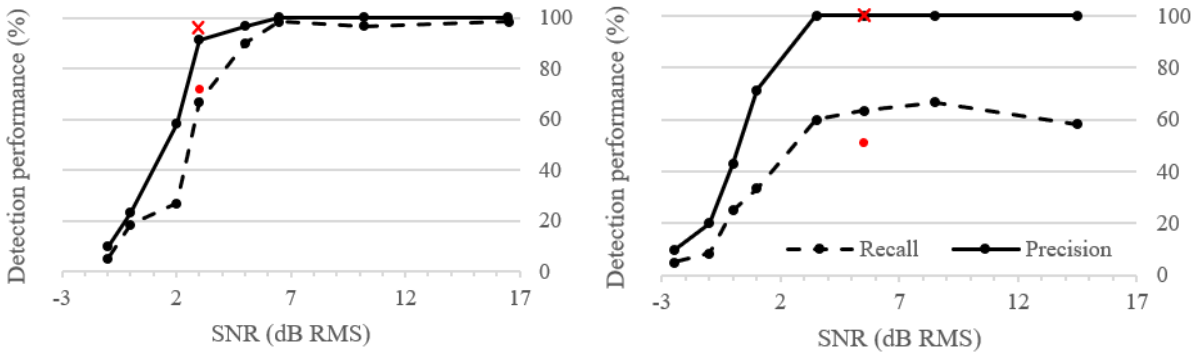
* V_{DD} = 300 mV

5.2.2 Real-World Sound Recordings

The recording XC867042 was chosen for its quiet background and mono-source scenario to first evaluate the circuit under probes with real-world calling songs. A precision of 100% is obtained w/o band-pass filtering on the envelope of the signals. Amplifying and saturating the signals to create square-like pulses allowed to increase the recall that would drop because of the lack of energy to make LN2 spike. A recall of 100% was obtained when amplifying the raw or filtered signal after normalization with $T_d \cong 10$ ms. Moreover, the circuit correctly detected Δ with same precision and recall when AN1 received the half-wave rectified of the raw recording instead of the envelope.

Artificial white noise is then added to the normalized recording XC867042 w/o reduced amplitude by a coefficient α. Here, no amplification of the input signal is performed, and the window T_d is tuned so the selectivity is maximized at SNR = 0 dB root mean square (RMS). Under probes, the envelope of the filtered signal alone is not sufficient to make AN1 and/or LN2 to generate bursting spikes for $\alpha \leq 0.2$. In simulation, the neurons spike more easily, leading to a higher recall than what is observed in practice.

On Fig. 5.8, the evolution of the precision and the recall is plotted according to the SNR for $\alpha = 0.2$ and without attenuation. In the case without attenuation (Fig. 5.8a), the precision and recall remain close to 100% then drop as the SNR deteriorates and false positives are detected. In the case of $\alpha = 0.2$ (Fig. 5.8b), a similar evolution of the precision is observed, whereas the recall increases up to a certain point for SNRs above 5 dB RMS then decreases because of the variation of Δ induced by the noise. In fact, the additional energy provided by the added noise allows the generation of more spikes when the target sound is too low, but also disrupts the detection of the target sounds.



(a) (b)

Fig. 5.8 Evolution of the precision (solid line) and the recall (dashed line) with added white noise in the recording XC867042 for (a) no attenuation and (b) $\alpha = 0.2$. The signal is band-pass filtered, normalized, and no amplification of the extracted envelope is performed. Results from observations under probes are marked with red crosses for the precision and red dots for the recall. RMS–Root Mean Square.

Under probes, a single result is reported per case: it corresponds to the lowest SNR for which the precision remains higher than 95%, with 96.9% precision and 73.3% recall at SNR = 3 dB RMS in Fig. 5.8a, and 100% precision and 55.6% recall at SNR = 5 dB RMS in Fig. 5.8b. Other

values of SNRs were tested and were not recorded, but similar observations are made between simulation and hardware. Especially at $\alpha = 0.2$ in hardware, no spike was generated without any additional noise, thus leading to an increase in recall of 55.6% with 100% precision between SNR = 42 dB RMS and SNR = 5 dB RMS.

Similar to the quiet recording XC867042 with artificially added noise, the recording XC922939 has a stationary noise for which 100% precision and recall could be obtained by amplifying the envelope and increasing T_d to 13.3 ms. Without amplification of the signal after normalization, 100% precision is still obtained but with 51.5% recall.

The circuit was further tested with real-world recordings of field crickets calling songs in the presence of ambient noise. Although detection was still possible using the envelope of the non-filtered recordings, the precision greatly increased with the use of a band-pass filter. It removes most interferences from noisy sound sources out of the bandwidth. Sounds in the bandwidth create a complex scenario where sources cannot be differentiated. In the recording XC854026, birds and insects interfere with the songs of two crickets whose intensities vary differently in time. Fig. 5.9 shows the envelope of the filtered recording XC854026 and the spectrogram of the raw signal; the calls are not overlapped and can be both identified. In Fig. 5.10 on the other hand, the waveform of overlapping calling songs leads to halved inter-pulse delay, hence, detection is not achieved.

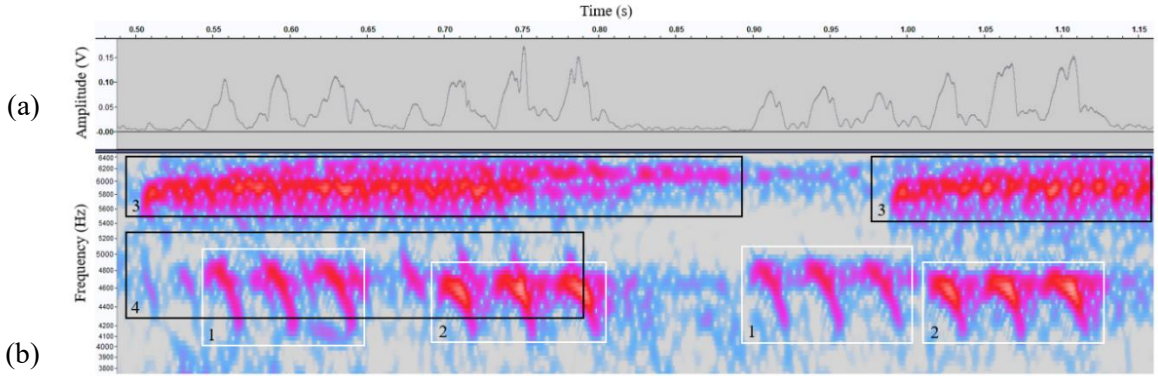
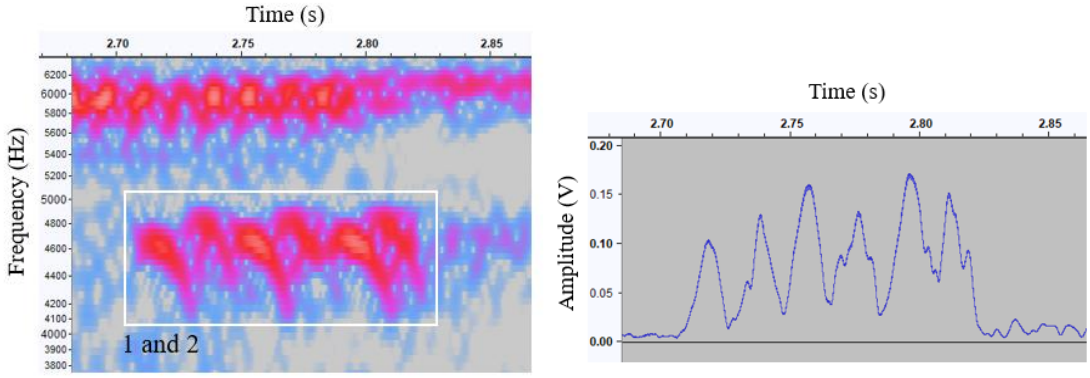


Fig. 5.9 Multi-source scenario. (a) Envelope of the band-pass filtered recording XC854026 and (b) spectrogram of the raw recording with a logarithmic frequency scaling. The white rectangles (1 and 2) are two non-overlapping crickets' song distinguishable according to their dominant call frequency and average call period. The black rectangles (3 and 4) are unidentified bird and/or insect calls. Adaptation from XC854026 CC BY-SA 4.0.

The filter allows the crickets' calling song to be isolated from the other noises. Nevertheless, the integrated signal is a combination of the two perceived as one by the circuit. It results in false negatives as a consequence of overlapping calls loud enough to make LN2 spike, or of destructive interferences in the temporal domain. When considering all characteristic inter-pulse delays, from very quiet calls as well, the best performances on XC854026 were obtained on the amplified envelope of the filtered recording with 100% precision and 36.2% recall. Without filtering and amplifying, the precision dropped to 59.3% and the recall to 13.8%. The recall in the best conditions rose to 39.6% when removing hardly perceptible calls.



(a) (b)

Fig. 5.10 Overlapping calls from two distinct crickets in the recording XC854026 observed on the (a) spectrogram of the raw sound signal and (b) the envelope of the filtered signal. Adaptations from XC854026 CC BY-SA 4.0.

Similar observations were made on the recording XC853470 with 100% precision and 88.1% recall on the amplified envelope of the filtered signal, which dropped to 58.1% precision and 25.8% recall on the amplified envelope of the raw signal. Then, the recording XC854155 was only tested using the amplified envelope of the filtered signal where we reported 100% precision and 41.2% or 80% recall when considering or not very quiet calls respectively.

5.2.3 Power Consumption Performances

During the evaluation of the detector's precision and recall with artificial and real-world sound signals, the current drawn by the circuit under probes from the supply power is measured to compute the DC power consumption according to (5.2).

For all signals tested and with V_{DD} set to 300 mV, the average current delivered during detection by V_{DD} is of 2.5 nA, corresponding to an average power consumption of 750 pW. At most, the current drawn reaches 4 nA, so a maximum of 1.2 nW. Looking at real-world sound signals only, the average current drawn amounts to 2.71 nA, or 813 pW, which is similar to the overall mean. Furthermore, the static (standby) consumption is measured by providing a constant input at 0 V. Then, an average current of 1.9 nA is delivered by the chip, so a static power consumption of 570 pW is reported. These performances are further discussed in section 5.4.2 in comparison to the literature.

5.3 Automatized Tuning

The potential of the core circuit adapted from the biological neuronal circuit for inter-pulse delay detection was validated with an ULP consumption suggesting its use in sensors arrays or as a set of multiple inter-pulse delays detection. As such, it raises the question of user interaction and how weight can be set.

In the demonstrator previously introduced, most of the weights are manually tunable, but the objective for a possible final prototype is to reduce the number of pads and therefore the number of input voltages required. By adding an automatic tuning of the key weights, described as

"manually tuned" in the pin layout (Fig. 5.2c), the "fixed parameters" may be set with voltagedividers from the supply voltage (hence the mutualization of weights) while the remaining tunable weights are set according to the circuit's output in response to the input signal.

In this section, two additional circuits are described for detection optimization with regard to the technology variability and tuning of T_d for adaptation of the detection selectivity to the presented inputs.

5.3.1 Leak Optimization Circuit

Having fixed weight in hardware increases the risk of creating non-functional chips because of the variability of and in electrical components. In order to correctly perform detections, various combinations of the synaptic weights can be used and identified in simulation. However, slight differences with the simulations are observed once integrated on chip due to the variability induced by the subthreshold mode of operation. The transistors are not intended to behave in this operation regime, creating these differences in the whole circuit.

The most likely and problematic issue that was identified concerned the possible lack of excitatory current from AN1 and/or LN5 to LNF, making the feature neuron unable to spike even with T_d correctly configured. Once the weights are fixed in hardware, the response of excitatory synapses cannot be altered. Yet, LNF requires a fine balance between its two excitations to generate spikes only when AN1 and LN5 are spiking simultaneously. There could also be an excess of excitatory current making LNF spike unrelated to a detection from either AN1 or LN5 activity. Since this balance is relatively easy to break, at least one parameter must remain tunable to allow a correction of the balance.

Over multiple tests to determine the possible weight combinations, it appeared that the leak Wl had sufficient leverage on LNF spiking activity to adjust a chosen combination to the circuit's variability. While adding a leak at LNF might not have seemed essential, its usefulness becomes more apparent in the circuit proposed thereafter for handling hardware-related variability.

Proposed Circuit

An adjacent circuit, depicted in Fig. 5.11, is added to the inter-pulse detector for automatic adjustment of Wl through a low-power consuming counter and digital-to-analog converter (DAC). It has the function of progressively decreasing Wl in discrete voltage steps until LNF successfully generates spikes during detections. We will refer to this scheme as an optimization process.

In order to decrement Wl, a flag must be raised stating that a detection was missed. Using an additional neuron denoted LNnF (as in *not-feature*) and inhibition from LNF, LN2's spike bursts are reproduced by LNnF. An expander with weight W_{EnF} extends in the temporal domain the train of spikes to form a single square pulse (Fig. 5.12). Its minimal and maximal durations depend on LN2's spiking frequency and the delay between two input calls, respectively; for inter-pulse delays equal to the inter-spike interval of LN2 for example, a single square signal is produced in output of the expander, hence the dependency.

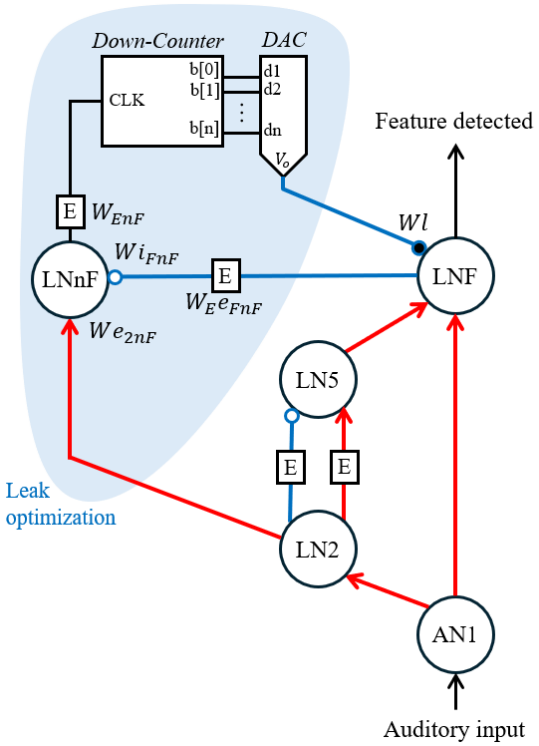


Fig. 5.11 Circuit of Wl tuning by automatic decrement. A decreasing counter receives the output spikes of LNnF (informing on missed detections) which provokes discrete decrements of the counter's value (as a binary number b[0..n]) and output voltage $V_o = Wl$ delivered by a DAC. Only the weights concerned by the leak optimization circuit are shown.

The expander provides an asynchronous clock source so its ascending fronts decreases the value of the counter (and thus of the DAC) until Wl is low enough for LNF to generate spikes. Indeed, only the first rising edge is needed to trigger the counter. Finally, for LNnF to inform on missed detections only, an inhibition from LNF to LNnF prevents the latter from spiking in the event of a detection. Although LNnF reproduces LN2's activity, this neuron is essential. LN2 alone cannot inform on missed detections since receiving inhibition from LNF would also prevent the generation of the rebound and further detection. The missing detection event (rising edge) will be shortly referred to as $\overline{FLAG}$.

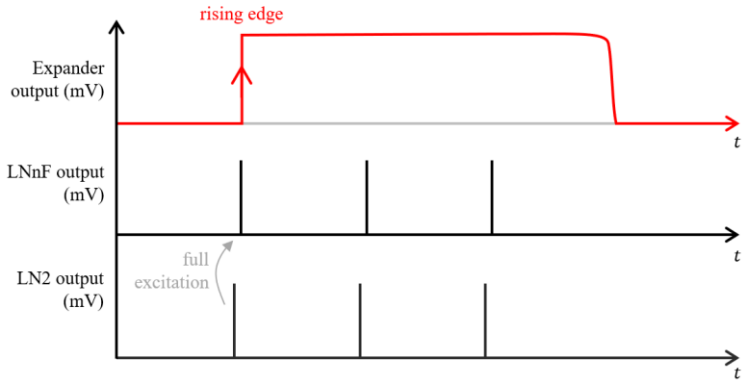


Fig. 5.12 Illustration of LNnF and its expander processing. The duration of the expander is set just long enough to have a single rising edge.

The optimization process is carried out by initializing Wl at a high voltage which then decreases by a voltage step determined by the DAC's resistor ladder at each missed detection. Wl then converges to a lower voltage such that excitations to LNF are slowly facilitated and detections are optimized. To be more precise, the digital counter outputs in a parallel bus a binary number then converted to a voltage by the DAC according to a reference voltage V_{ref} ($\leq V_{DD}$) and the ground (0 V), such that

$$Wl = (N \bmod N_{max}) \frac{V_{ref}}{N_{max}} , \tag{5.3}$$

where N is the decimal value of the decreasing counter, and N_{max} the maximal decimal value of the counter. Because of digital counters' architecture, N is a cyclic variable, initialized at N_0 (could be N_{max}) and resetting to N_{max} (wrapping around). In practice, the counter stops decreasing after LNF starts spiking, so N never wraps around and would require a manual intervention to be reset to N_0. Naturally, this process requires that LNF is able to spike so Wl is not incorrectly set at its minimum, or in other words, the inter-pulse delay presented in input must fall within T_d. The final value of Wl can then be verified by testing inter-pulse delays out of T_d. For a potential user, a switch could be used to launch and stop the optimization process, for example by enabling and disabling the excitation from LN2 to LNnF respectively.

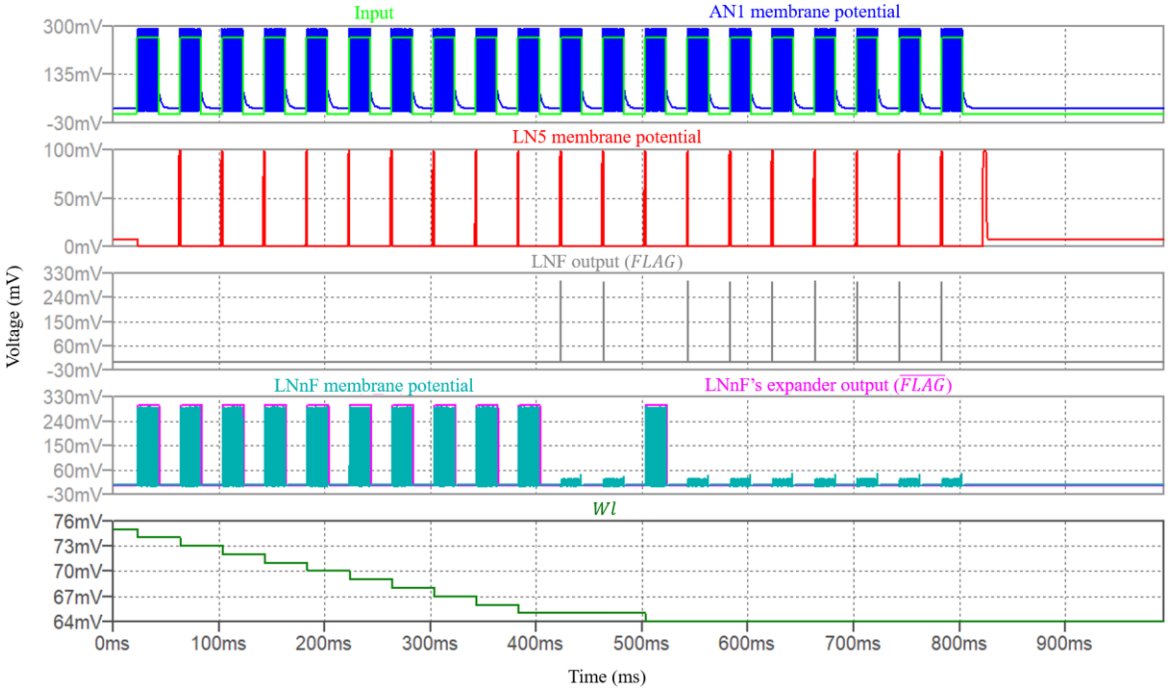
During this process, T_d does not have to be correctly set for detection of a target characteristic delay, but inputs with inter-pulse delay within T_d should be fed to AN1. The goal of the leak optimization circuit is to adjust the effect of the two excitatory synapses connected to LNF with fixed weights, and not to adjust T_d to inputs. It can however be considered as a supervised scheme since the input is controlled by a user. The window T_d can be set to its largest width by setting $W_E i_{25}$ to V_{DD} and $W_E e_{25}$ to its minimal voltage. The minimal voltage for the expanders weight is not 0 V because, at some point, low-pass filtered input spikes have too low amplitudes for the digital buffer to output the correct voltage. In simulation, expanders always output V_{DD} when their weight is set to 0 V even if no input is provided, making them nonoperational.

Simulation

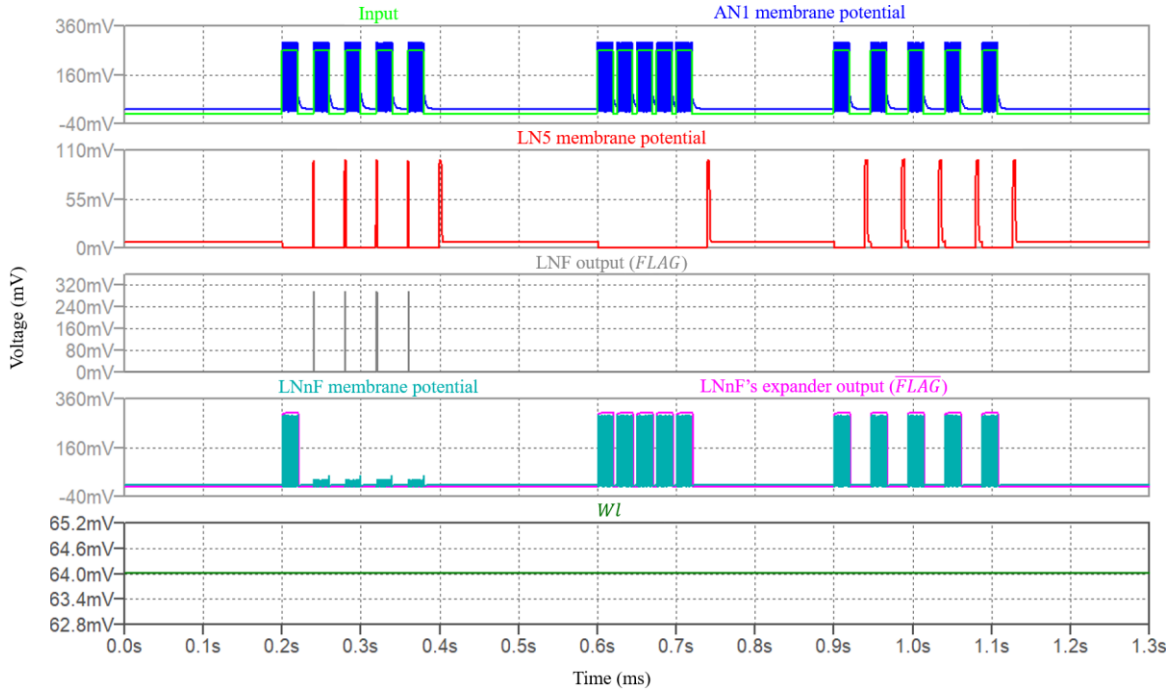
A demonstration of the optimization process is run in simulation. Fig. 5.13 shows $\overline{FLAG}$ and $FLAG$ outputs and control signals in input of the counter for modification of Wl. The leak was lowered just enough to make LNF spike when a correct delay was presented, but not so low that any delay would generate a detection, thus retaining its functionality.

This circuit is partially implemented in the chip shown in Fig. 5.2c. All additional neuromorphic elements were integrated on chip, with the exception of the counter and the DAC. Appendix B.1. describes this partial implementation and observations made under probes.

Currently, this design implies a supervision by the user that would provide a succession of artificial calls with a delay within T_d. In fact, the first call of a calling song currently causes the flag $\overline{FLAG}$ to be raised and Wl to be decremented. Right after the initialization this does not pose a problem, but in an unsupervised context with several short calling songs in succession, unwanted decrements of Wl may occur. We will see in the following circuit for T_d automatic tuning that implementing a more complex design for more detailed flags solves this issue; it was developed after implementation on chip of the demonstrator.



(a)



(b)

Fig. 5.13 Simulation of the optimization process with ideal square inputs and $\Delta = 40$ ms. (a) The leak Wl is automatically tuned by slowly decreasing Wl at each missing detection (1 mV decremental steps). (b) The final value 64 mV of Wl is validated by stopping the optimization and by presenting the correct delay Δ, then delays lower and higher than Δ. LNF successfully generates spike only when the correct delay is presented. The rebound at LN5 is finely tuned for high selectivity.

5.3.2 Training flags for Automatic Tuning of the Detection Window

Being able to automatically handle the chip's variability is useful but does not affect the range of detection. A common feature of detection or recognition systems are their adaptability to the environment. Especially with a recognition mechanism that depends on a straightforward detection window, it is possible to devise a simple processing useful to perform a learning scheme for tuning of the two weights responsible for the detection temporal selectivity.

Motivation

Let us consider a distribution F_Δ of characteristic delays that we would like to detect in the context of biodiversity monitoring. Like it was described in Fig. 5.5 on the few real-world recordings of field crickets calling songs selected as test inputs, depending on the individual studied, the distribution F_Δ varies more or less around a mean.

In the event that F_Δ is known, the detection window T_d can be adjusted through $W_E e_{25}$ and $W_E e_{25}$ using correspondence tables. Once a chip is characterized, the possible bounds of T_d are known according to the expanders' duration weight. A potential user may then supply these weights with the appropriate voltage for a target detection window without the need to visualize the rebound like in the calibration procedure. For example, in simulation, the correspondence between the ON duration of an expander's output according to its duration weight W_E is plotted in Fig. 5.14.

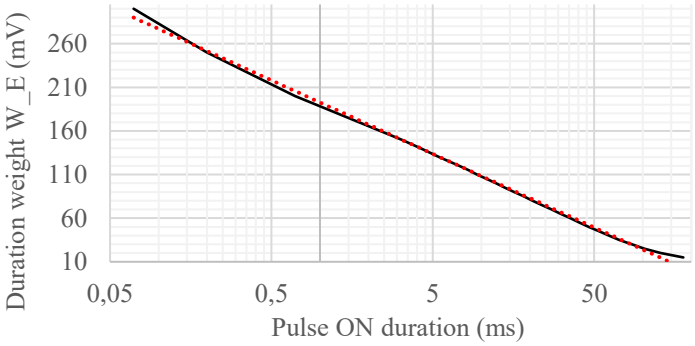


Fig. 5.14 Evolution of the ON duration (measured at 200 mV in simulation) of a spike temporally extended by an expander with tunable duration weight W_E and $V_{DD} = 300$ mV. The abscissa follows a logarithmic scale, and the ordinate a linear scale. The trend curve is shown by a dashed red line whose equation is $W_E = -36.6\ln(x) + 192.6$. Below $W_E = 15$ mV, the output of the expander is always V_{DD}.

For unknown F_Δ, however, a correspondence table may not be sufficient. Typically in an unsupervised learning, where the detector has to adapt to the undetermined inputs, an on-chip automatic adjustment of the expanders' duration weights is necessary to adapt T_d. It also includes the scenario where F_Δ evolves in time; for example if the population of singing insects changes and T_d must be modified in consequence. In the end, it mostly addresses the question of continuous adaptation to the environment and user involvement. To illustrate the latter, more

autonomy and simpler user interactions could be requested, like having few buttons (ON/OFF commands) instead of potentiometers (continuous voltages). Then, an automatization to skip the manual tuning would be relevant.

Regardless of the motivation, an online (on-chip) training is possible but requires several flags to indicate how the weights must be modified for appropriate learning. Hence, a focus is made on the extraction of keys control signals in the form of flags which will constitute a base for a future (un)supervised tuning of T_d.

Training Flags Extraction

For any learning scheme, simple or complex, providing the correct information on the state of the system is essential. Otherwise, the system cannot carry out the appropriate update of the parameters. In conventional and neuromorphic computing, deep learning-based methods rely on loss metrics, or on positive/negative feedback signals in the case of reinforcement learning, to determine the changes to apply on the weights [131], [132].

In this analog circuit, a continuous loss metric cannot be easily computed, unlike in digital systems. Instead, simple flags are extracted that indicate where the presented inter-pulse delays Δ are temporally situated from the window T_d. The three possible cases are $\Delta \in T_d$, $\Delta < T_d$, and $\Delta > T_d$, with Δ following F_Δ. The former, namely $\Delta \in T_d$, corresponds to correct detections and is indicated by the output $FLAG$. Concerning the other two scenarios, additional processing is necessary.

Already in the context of the leak optimization, the flag $\overline{FLAG}$ for missed detection was proposed and tested in hardware. In this section, we take the granularity further by splitting this information into the two flags $\Delta < T_d$, and $\Delta > T_d$ for a more detailed status of the tuning. It can be noted that the following additional circuit was designed after the demonstrator implementation and can be factorized with the leak optimization circuit where $\overline{FLAG}$ would become the output of a neuron taking in input the flags $\Delta < T_d$, and $\Delta > T_d$ for full propagation in an OR logic (or can simply be an OR gate). This would thus allow the leak optimization circuit to bypass the first call of calling songs thanks to the additional processing provided by this design.

The adjacent neuromorphic processing for training flags extraction is proposed in Fig. 5.15, tested in simulation. Efforts were made to limit the number of neuromorphic and electrical components by using signals already available within the circuit. In Fig. 5.15, the leak optimization circuit is also depicted for a full view of the proposed circuits. The aggregated automatization design is described in Appendix B.2.

For $\Delta < T_d$, a neuron duplicate of AN1, denoted AN1b (as in *bis*), receives excitation from AN1 and is self-inhibited in order to reproduce only the first spike of AN1 at the onset of the input call. An integrator with a small temporal constant is used to slightly delay the inhibition enabling the first spike to be generated. Then, using a two-input AND gate between AN1b's single spike and the output of the inhibitory expander (controlled by $W_E i_{25}$), AN1b's spike is propagated if it is generated before LN5's rebound. In other words, it corresponds to the case where AN1b's spike does not coincide with LN5's rebound but with the output of the inhibitory expander that contributes in shaping the rebound. This is especially suitable as the rebound is

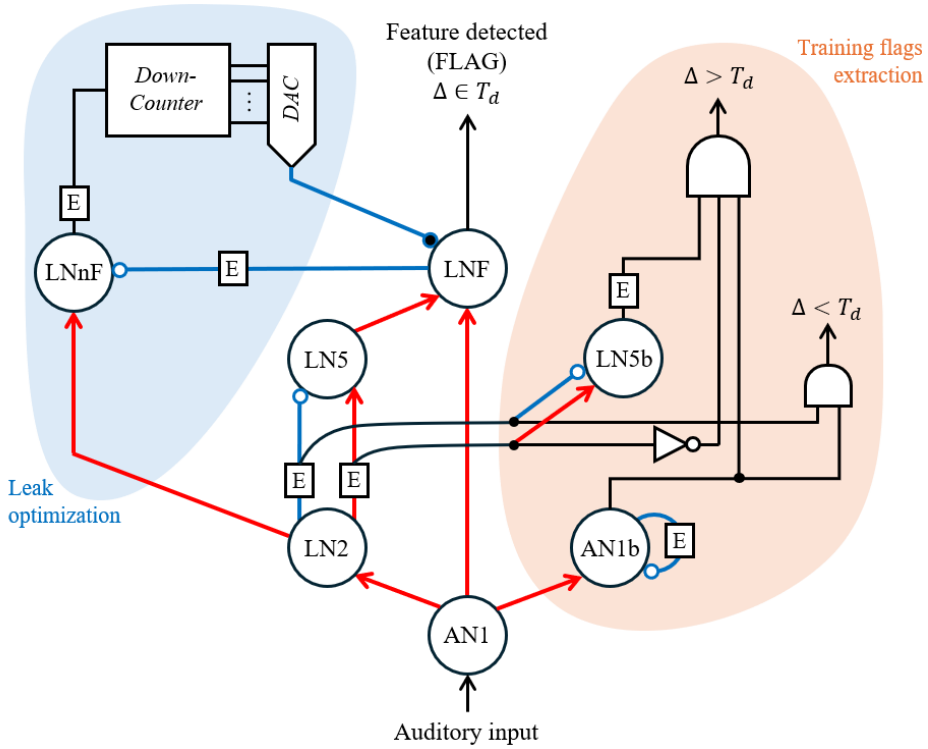


Fig. 5.15 Training flags extraction circuit, highlighted in orange. The leak optimization circuit is shown in blue. Black dots are electric lines connections.

inhibited and not generated when a successive call with a Δ lower than T_d follows. Consequently, this flag indicates that the lower bound of T_d should be lowered in order to expand towards the Δ it just missed, so $W_E i_{25}$ should be increased.

For $\Delta > T_d$, a neuron duplicate of LN5, denoted LN5b (as in *bis*), reproduces the rebound with full excitation from LN2, so as to generate a spike burst during T_d. An expander in output of LN5b with maximum duration creates a time window of the inter-pulse delay that will be considered by the flag $\Delta > T_d$. It allows the differentiation between large values of Δ and silences between two calling songs presented to the circuit. A three-input AND gate between AN1b, the expanded output of LN5b, and the inverted output of the excitatory expander (controlled by $W_E e_{25}$), propagates AN1b's spike only if it is generated after the rebound. In other words, the AND gate informs on the coincidence between the elongated rebound minus the actual rebound corresponding to T_d with the onset of a successive call. Consequently, this flag indicates that the upper bound of T_d should be raised, so $W_E e_{25}$ should be decreased.

In the end, both flags are generated at the onset of input successive calls of a song when a detection should have occurred. Naturally, no flag is generated at a song's first call as it should not. A simulation of the training flags extraction is visualized in Fig. 5.16. Zooms on AND gates processing are provided for better understanding of the extraction process of each flag.

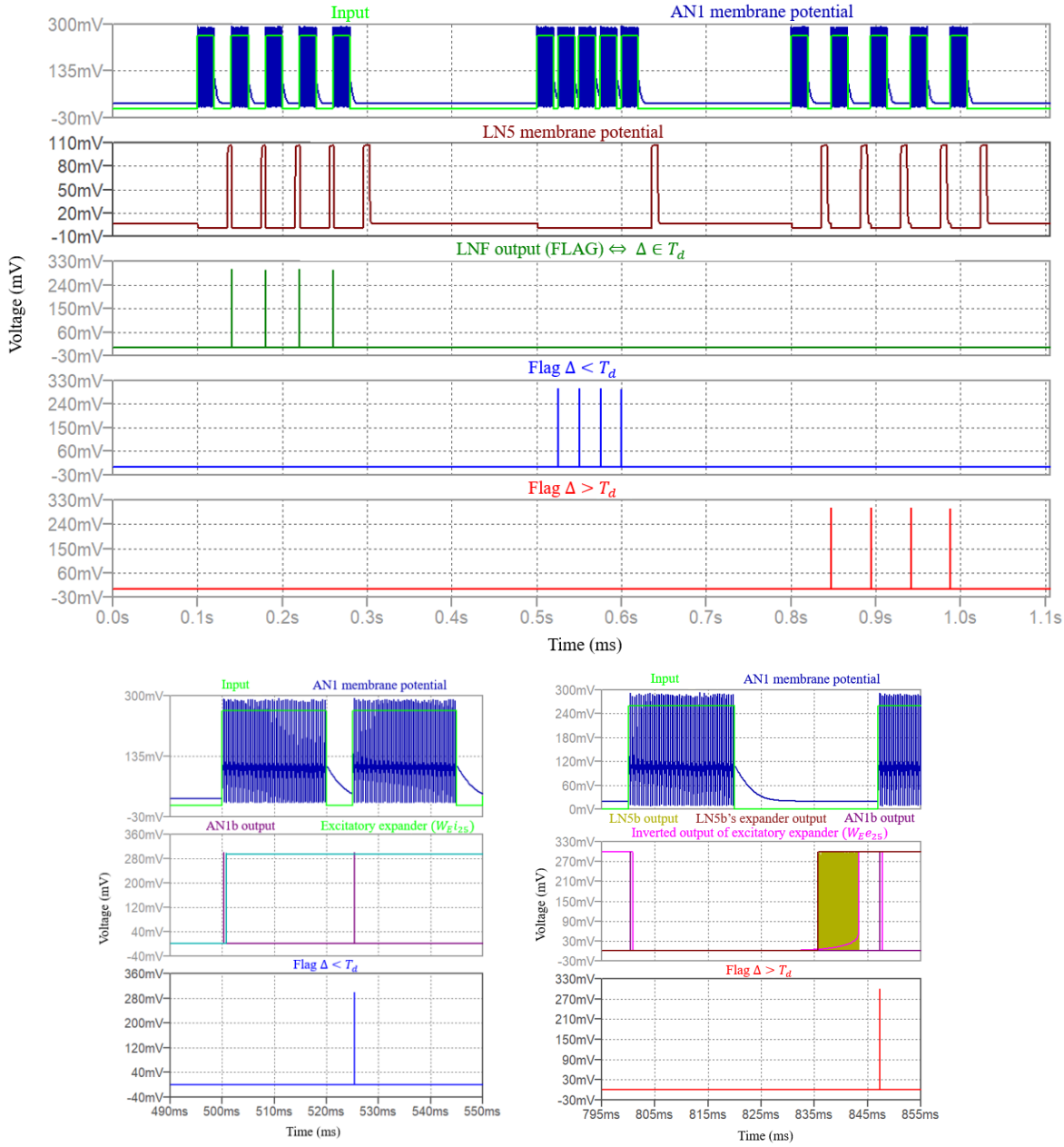


Fig. 5.16 Extraction of the training flags in simulation. (a) Generation of flags to indicate the scenarios $\Delta \in T_d$, $\Delta < T_d$, and $\Delta > T_d$ performed in simulation. The rebound is visualized at LN5's membrane potential. (b)–(c) Zoom on the inputs (2nd plot panel) and outputs (3rd plot panel) of the AND gates that contribute to the creation of the flags $\Delta < T_d$ and $\Delta > T_d$ respectively. LN5b generates spike at a high spiking rate, the width of plot lines do not allow visualizing the individual spikes. The circuit is tuned for $\Delta = 40$ ms.

The rules stated for these flags may only widen T_d at this stage, and they suppose that T_d is already in the vicinity of the delay Δ to detect. Indeed, a satisfactory selectivity cannot be obtained for T_d initialized far from F_Δ mean since T_d would only expand. In this manner and in addition to the above, a control signal is needed for narrowing T_d. Since the system does not have access to the minimum and maximum delays of the distribution it is slowly adapting to,

only the feature flag ($FLAG$) indicating $\Delta \in T_d$ is suitable for this task. In the scenario where T_d encompasses all Δ, there is no way to know by how much the selectivity can be narrowed down because the output $FLAG$ is an event, a binary information (detection or no detection). Therefore, the output $FLAG$ can be used as a leaky parameter on the weights involved in shaping T_d, such that T_d would slowly narrow towards its central value at each detection. The duration weight would then be modified using up-down counters updated by the training flags.

Having now the necessary flags to indicate which adjustments to make, it is possible to automatically tune $W_E e_{25}$ and $W_E i_{25}$. The flags $\Delta < T_d$ and $\Delta > T_d$ describe missed detections while $\Delta \in T_d$ describes a detection from which T_d is adjusted through narrowing and widening. Fig. 5.17 illustrates the training flags' effect on T_d, previously mentioned as possible basic rules for the update of the detection window according to shown inputs.

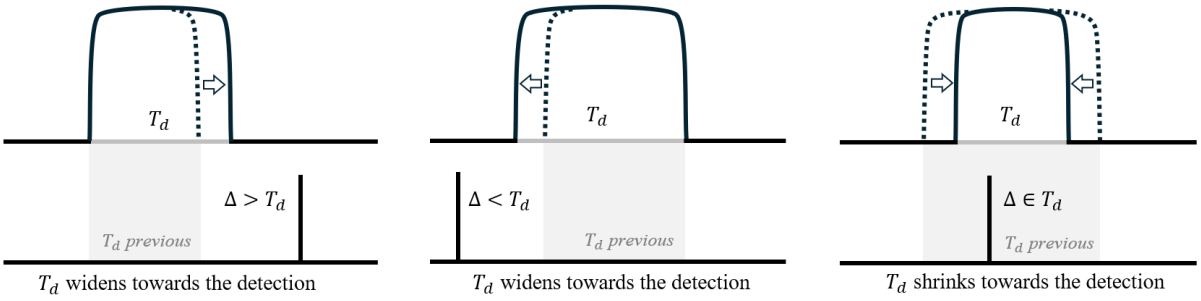


Fig. 5.17 Tuning of the detection window from the extracted training flags illustrated. Vertical black line are the spike generated in output of LNF ($FLAG$). (a) For $\Delta > T_d$, the upper bound of T_d is extended by decreasing $W_E e_{25}$. (b) For $\Delta < T_d$, the lower bound of T_d is further lowered by increasing $W_E i_{25}$. (c) For $\Delta \in T_d$, namely when a detection is observed at $FLAG$, both bounds of T_d are narrowed around $t$he central value of T_d. The temporal window T_d is exactly the rebound generated at LN5.

These cues can then be passed to another circuit, analog or digital, that will make the necessary adjustments with suitable processing and learning schemes. Implementing these simple rules alone would not result in a satisfying training speed and final selectivity, but they provide straightforward guidelines for the potential use of the training flags.

5.4 Discussion

The several tests performed on the demonstrator using software generated (ideal) and real-world sound signals provide great insights on the full potential of this circuit and the subthreshold CMOS neuromorphic technology, as well as their operation capabilities once integrated on chip.

5.4.1 Hardware Variability from Sub-Threshold Mode of Operation

In hardware, most of the components' variability is related to the low supply voltage, lower than the source-to-gate threshold voltage of transistors (it is to be stressed that the electrical models of the 65nm CMOS technology are guaranteed above the threshold voltage). Yet, we

demonstrated that even in these challenging conditions, and also with added complexity from noisy real-world inputs, the circuit was able to perform with 100% precision and T_d below 13.4 ms. The variability affects the expanders, and so the rebound period T_d and the delays detected, resulting in lower detection selectivity. The tuning of T_d uses $W_E i_{25}$ and $W_E e_{25}$ to adjust the lower and upper bound respectively of the rebound. However, under probes, the modification of $W_E i_{25}$ for the lower bound would typically affect the upper bound without modification of $W_E e_{25}$. In practice, this phenomenon is explained by small interferences between lines carrying the spiking activity, creating higher dependency between weights, and is deterministic once characterised. For set values of the expanders' duration weight, the variability of T_d timing and width, resulting from the low supply voltage, is not impacted.

A more biological looking rebound could be obtained by using only integrators (by removing double inverters of expander), but it would also create more uncertainty in determining the effective detection period and decrease the detection rate because potentials output by the integrators would not be pulled to V_{DD}. Excitation We_{5F} cannot go higher than V_{DD} to potentially compensate for the lack of energy, and additional energy would only be able to be supplied by AN1 spikes at the expense of systematic false detections. In the end, our choices align with our motivation to take inspiration from biology, as opposed to mimicking biology, while taking advantage of the flexibility and strengths provided by an electronic design.

Incidentally, the use of logic gates in the proposed circuit of training flags extraction follows the same approach. While using neurons to perform AND logic is possible, it involves capacitances and more transistors than traditional logic gates designs. Additionally, the weights would need to be set so only the coincidence of all inputs would propagate spikes, which is likely to be disrupted by the hardware variability. The leak optimization circuit has this exact purpose of balancing the excitations received at LNF to perform AND logic.

The use of the leak at LNF is particularly interesting as it allows weights to be pooled. Weights We_{12}, We_{2F}, and We_{25} could be set to the same voltage 200 mV, which means that the number of electronic components (resistive and pads) and the total die size of the circuit may be reduced in the definitive design. Unlike the expanders' weight, these weights must in fact be tuned with consideration of the thermal noise and variability of the electronic components. The leak thus takes over as a more global parameter for optimization of the circuit.

5.4.2 Detection Performances with Real-World Recordings

In the present circuit, no adjacent neural processing is proposed to identify a rhythmic call drowned in non-stationary noise or mixed with other rhythmic sounds present in the same frequency band, as it was demonstrated with the two singing crickets in the recording XC854026. However, it does not prevent the circuit from reporting correct detections with 100% precision, even with a recall at 32.2%, when the signal is sufficiently amplified. Very quiet calls are recorded in XC854026 audio file that the circuit could not recognize without impacting the precision. If very quiet calls hardly noticeable in the time domain are removed from the formula, the recall reaches 39.6% and 80% in the recordings XC854026 and XC854155 respectively. Like most auditory systems, bandpass filtering is essential to obtain

reliable results in noisy and multi-source conditions. Besides, such a passive circuit may be added at a negligeable cost in practice.

By fine tuning T_d, high detection precision is reported for all recordings. The detection period T_d is decisive for this recognition task since the inter-pulse delay is the only discriminative characteristic identified by the circuit. Without setting boundaries on the delay selectivity T_d, 100% precision and with high recall can easily be obtained using appropriate amplification of the filtered signal envelope on all recordings listed in Table 5.1. When the noise is correctly attenuated within the filter's bandwidth, the major contributor to a great detection performance is the calls' amplitude. The delay selectivity becomes more decisive and relevant when non-target sounds from the surroundings create inter-pulse delays close to Δ.

The inter-pulse delays could also be detected in the presence of noise and even facilitated some detections where the input signal was not strong enough. By using the envelope of the acoustic signal, the noise provided a bias current to the neuron. Depolarization by noise is in fact a phenomenon observed in neurology [133]; a constant and small excitation increases the resting potential of the neuron's membrane such that it requires less energy from other excitatory synapses to spike. This trick can be useful in quiet soundscapes where target sounds have low intensity. It would increase the power consumption, but it would also enable detections. Unlike a thresholding of the voltage through amplifiers (as suggested for the HRD-based coincidence detector), the input in the inter-pulse delay detector is directly given to a neuron as a current. Therefore, the current can be controlled by a weight (voltage set at the gate of a transistor) for implementation of an adaptative thresholding.

5.4.3 Ultra-Low Power Consumption

The number of transistors and power consumed are exceptionally low in this fully analog design compared to the literature. Although mixed-signal implementations allow to reprogram the processor chip using a computer, smaller circuits are not at an advantage within such versatile and bulky hardware. Their integration with digital systems is facilitated at the price of an increased processing load and necessary digital interface for setting weights.

The proposed circuit is similar to [128]'s in that it relies on synapses to emulate a post-inhibitory rebound, but our choice of implementing ULP biomimetic Morris-Lecar neurons and compact synapses in place of more complex formalism allowed us to reduce the number of transistor/capacitance to 12/2 per neuron including the digital buffer and 8/1 or 2/0 per synapse, including the expander or not respectively, compared to 29/3 per AdEx neuron [134] and 13/2 per differential-pair-integrator synapse [44], [135] used in the processor [34]. The number of transistors and value of the capacitance impact the total die size of the circuit and therefore the cost of the chip, so smaller chips are preferred.

The same authors report in [129] an energy efficiency of 883 pJ for the generation of a spike that corresponds to a supply voltage of 1.8 V according to the processor's specifications [34]. In contrast, our implemented neurons feature an energy consumption lower than 100 fJ per spike, leading to a design more energy efficient by at least 3 orders of magnitude. Overall for the demonstrator, we report a total power consumption of about 1.2 nW at most or 750 pW on average during detection, and about 570 pW on average when AN1 receives no stimulation.

Actually, these power consumptions are in line with those reported in [45] for the on-chip implementation of single ML neurons, bearing in mind the four neurons used in our basic circuit.

The partial implementation of the leak optimization circuit (that is, one neuron and two expanders added to the demonstrator besides the core circuit) is also supplied by V_{DD}, but the neuron LNnF remained inactive during the DC current measurements by shunting the excitation from LN2 to LNnF, and expanders' duration was set to be minimal for minimizing the dynamic power consumption.

Little difference is observed between low and high spiking activity from quiet or noisy sound inputs since the static consumption is dominant in the total consumption. In simulation, V_{DD} could be lowered to 200 mV for a more energy efficient system and better similarity to the biological observations, but resulted in unstable neurons under probes. The variability of the components is accentuated with lower supply voltage, so bringing V_{DD} to 300 mV was favored to propose first and foremost a stable demonstrator.

It goes without saying that when including the whole digital hardware required for handling spikes propagation and weight memory in [128], the number of components quickly grows, whereas our work is limited to the essential neuromorphic elements and unpowered electronic components such band-pass filter used for pre-processing. This statement is also valid for the power consumption assessment not reported in detail in [129] aside from the main operations quantified in [34].

Table 5.3 Comparison of the main characteristics of the delay detection circuits.

	[128]	**This work**
Use a processor	Yes	No
Technology CMOS	0.18 µm	65 nm
System implementation	Mixed-signal	Analog
Number of transistors / capacitances per neuron	29 / 3	12 / 2
Number of transistors / capacitances per synapse	13 / 2	8 / 1 (with expander) 2 / 0 (without expander)
Die size [*1]	43.79 mm²	0.389 mm² (1080 × 360 µm²)
Maximal Δ detected	100 ms	187 ms
Detection selectivity [*2]	–	4 ms – 7 ms
Supply voltage	1.8 V	300 mV
Energy consumed per spike generated	883 pJ 8.6 nJ [*3]	< 100 fJ
Total power consumption during detection	–	1.2 nW (at most) 750 pW (on average)
Standby power consumption	–	570 pW

[*1] Includes pads and whole processor if any.

[*2] For systematic detection, measured in hardware with ideal artificial inputs.

[*3] Consider the following hardware operations in [34]: *Generate one spike*, *Encode one spike and append destinations*; *Broadcast event to same core*.

Table 5.3 resumes the main characteristics of this work's circuit under probes with comparison to [128]'s circuit.

Moreover, compared to the sound recognition systems [120] and [124] where the lowest power consumptions reported are 80 mW and 71.2 µW with consideration of the static and dynamic consumption, the circuit presented in our work largely surpasses these energy performances (to be noted that because their dynamic power consumptions stand above the µW, this also holds for [121], [122], [123]). Even if the circuit is not comparable to a full SNN like [120], [121], [122], [123], [124], it is still a recognition mechanism although limited to very simple acoustic signals. Its low energy cost is enabled by its simple architecture and processing. In fact, to be truly comparable to these systems, the inter-pulse delay detector would need to be enhanced and integrated to a larger SNN. Multiple inter-pulse delay detectors, tuned according to target sounds, could be combined in a feature layer for the output spikes to be interpreted at a higher level of abstraction. Then, a more meaningful comparison to the literature's sound recognition systems would be possible.

5.4.4 User Interaction

The proposed base circuit requires manual interactions with a user to make the initial calibration and potential renewal of the calibration to adapt the detector to other characteristic delays. Several solutions are available with more or less intervention by the user, electronic feedback, and electronic components. As far as a user is concerned, the targeted simplicity and user involvement in the system configuration varies depending on the user's needs.

Instead of manually tuning the weights, both leak optimization and detection window automatic tuning allow the use of simple switch and buttons besides providing an adaptation to the environment. However, reducing the configuration procedure complexity means a higher number of electronic components and a more complex circuit architecture, bringing into light the compromise between power consumption and user-friendliness, driven by the target application.

Allowing a (un)supervised online learning of the detection window is useful in a situation where the distribution of the inter-pulse delays we want the circuit to learn is unknown. Otherwise, a correspondence table between the weights and the temporal bounds of T_d may be sufficient, provided that the table takes into account the component variability. In the hardware implementation, the values of capacitances or resistors may slightly vary from the initial design. A set of potentiometers would then give to the user the access to the tuning of these weights. Also valid for Wl tuning, a potentiometer could replace the leak optimization circuit so the user may take full control of the circuit's tuning.

5.4.5 Sparse Architecture for Automatized Weight Tuning

Nevertheless, an interesting use-case of the inter-pulse delay detector investigated in this chapter is the complete autonomy of the circuit's tuning for a supervised (dataset shown by the user) or unsupervised learning (constant adaptation to the environment and convergence to the most shown inter-pulse delays without user intervention). A rather straightforward and simple

rule set was described for a potential training with the motivation to emphasize the relevance of the training flags extracted. A more in-depth investigation of learning schemes using these flags could be conducted to identify more complex update rules compatible with the subthreshold CMOS neuromorphic technology.

The circuit does not have to be limited to mature technology. Emerging components could be envisioned for weight tuning, that would use the training flags provided, for example using memristive elements in place of the suggested up-down counters. Each neuron of the core circuit has a defined role and connections, creating a network more akin to a signal processing system than a neuron pool organized in layers or reservoirs, but SNNs' neuromorphic computing can still provide leads on how to perform an adequate learning [132]. The extraction of training flags is the first step to performing a successful training which was validated in simulation.

In the end, the architecture of the additional circuits could be further improved and tested. For example, the leak optimization and flag extraction circuits can be combined in order to reduce the number of components and the redundancy of missed detections flags, as well as fix the leak optimization circuit's inability to perform an unsupervised training (Appendix B.2). Besides, considering the variability in subthreshold CMOS hardware implementations, it is difficult to determine if the proposed autonomous tuning of all parameters would be fully operational. A prototype is required to evaluate the robustness and flexibility/adaptability of the additional features under probes. Like it was observed in the demonstrator with non-anticipated interferences, behaviors that can only be observed in hardware may bring further modification of the design.

5.5 Conclusion

In this chapter, outstanding inter-pulse delay detection performances are reported in ideal and challenging conditions, both in simulation and in hardware, by taking inspiration from a biological neuronal model of the temporal pattern recognition in female field crickets. Using the ULP neuromorphic technology, our demonstrator reached an ULP power consumption of 1.2 nW at most with a supply voltage of 300 mV for a sound recognition task in real-word conditions by relying on rhythmic characteristics of the sound source. Furthermore, its standby power consumption of 570 pW makes it an ideal candidate for embedded sensing.

At the moment, our hardware implementation is a prototype of the core delay recognition neuronal circuit with any excitatory weight tunable. Nevertheless, investigation on possible automatization of key weights tuning was conducted. Extraction of training flags reveals the possibility of an online (un)supervised learning of inter-pulse delays and opens on autonomous applications.

The circuit was tested in multi-source scenarios where encouraging results were reported. With several of these delay detection circuits, more than one rhythmic pattern can be identified such that a resulting temporal feature vector of characteristic delays could be processed for complex calls, songs, or non-animal signals recognition with ULP consumption. Putting aside the exploration of automatic tuning, the base inter-pulse delay detection circuit provides a

recognition mechanism whose compatibility with the subthreshold neuromorphic technology has been validated. Besides, the base circuit and evaluation its detection performances have been published in [Au3]. Furthermore, it can also be combined with a DOA estimator for a smarter SSL and tracking in a multisource context.

6
Perspectives and Conclusion

In the previous chapters 4 and 5, the core circuit of these precomputing tools was evaluated and several enhancements were investigated to solve the systems' shortcomings and extend their capabilities, involving the enhancement of localization performances and facilitation of the user experience. Nonetheless, the perspective tasks and applications in which these tools may be used had yet to be fully identified and explored.

While these localization and recognition tools can be used separately, the major interest and motivation of the design of these tools was their combination for performing a more complex study that is monitoring in a multisource scenario. Without recognition mechanisms, localization of sound source is but a succession of detections and estimated positions whose evolution is not always a fine indicator of the sources' count and movements. Instead of using complex mathematics to extract information on the sources from a series of DOAs estimations only, which incorporates correct estimations and errors, adding a joint recognition contributes in labelling the estimated locations.

This chapter provides a non-exhaustive overview of the application field available to these tools, centered around biodiversity monitoring and exploring both individual and combined perspective uses, before concluding on the overall work that was performed in this thesis.

6.1 Perspective Applications

The core of the two precomputing neuromorphic circuits described and studied in chapters 4 and 5 can be enhanced to improve the performance of the circuits' present features or propose the extraction and spike encoding of different information. These tools were introduced in the context of acoustic monitoring, yet any signal can be passed in input as long as it is compatible with the voltage dynamics of the subthreshold technology. They can be integrated to systems working with other modalities, whether it is another biological sense such as vision and touch, or with artificial signals like radio waves. Broadening this work's scope, transmodal

applications can take advantage of the constraints and potential already identified to benefit from the insights gained in this work and confirmed ULP consumption of the subthreshold neuromorphic and analog technology.

Focusing back on acoustic monitoring, already plenty of perspective applications are addressable with further research. An exhaustive and complete list of all possibilities can hardly be provided, but several identified variants of the present circuits open on a larger panel of use cases. Of the two neuromorphic tools, the inter-pulse delay detector allows the extraction of more different features relevant to source monitoring, contrary to the HRD-based coincidence detector that has most of its potential uses unraveled. While the delay detector provides information on a single input changing over time, the coincidence detector makes a comparison between two inputs, limited moreover to signals with sharp amplitude changes.

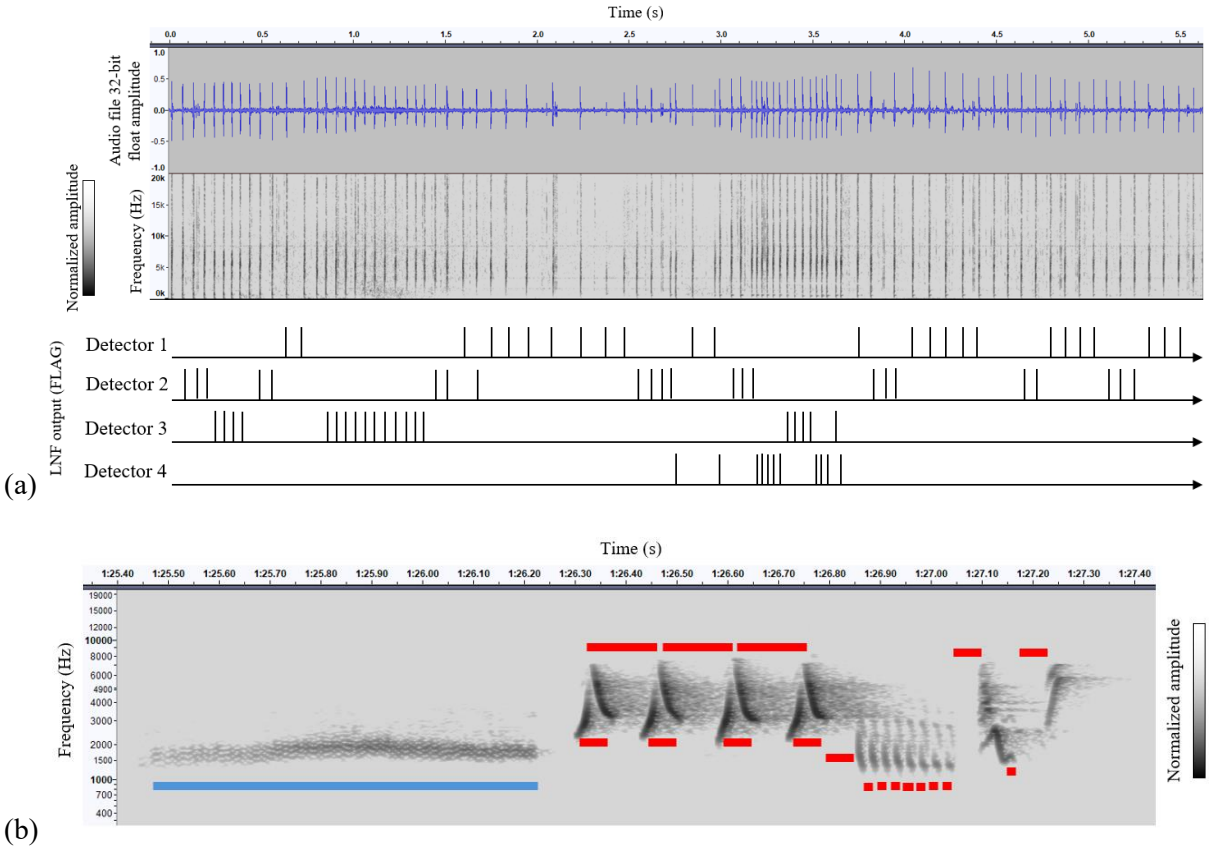


Fig. 6.1 Extraction of multiple temporal characteristics for acoustic pattern representation. (a) Waveform and spectrogram of a sperm whale click train. The inter-click delays are detected by four inter-pulse delay detectors tuned to different delay ranges. (b) Song of a common nightingale. Its signature is identifiable from its tonal changes, rhythm, as well as short and sustained tones. Determining the silence (red lines) and sound (blue lines) duration in specific or across frequency channels allows the generation of a fingerprint. Reproduced from [136], [111], CC BY-NC-SA 4.0.

Starting with use cases unrelated to SSL, the inter-pulse delay detector can be modified to perform a detection not only of silence durations (inter-pulse delays) but also of sound durations in a same circuit, and in another variant, of rhythms based on onsets. By incorporating few additional neuromorphic components, detection of other temporal features characteristic of the

sound sources are envisioned. Furthermore, using a multitude of those variants in parallel and configured to detect different durations enables the generation of a feature vector that will provide a discriminative spiking signature of relevant rhythmic sounds. For example, bird songs are recognizable thanks to their characteristic melody, rhythms, and short or sustained tones. Incidentally, it is applicable to music in general in addition to singing animals. Fig. 6.1a gives a schematization of the vector-like feature output of multiple inter-pulse delay detectors obtainable from a sperm whale's click train (first 5 s of the recording 8301900C of the Watkins Marine Mammal Sound Database [136]). Also, Fig. 6.1b highlights distinct temporal durations in a song of the common nightingale (recording XC999121[6] of the Xeno-Canto online dataset [111]).

A rhythm or sound/silence duration detector is not constrained to a single channel, but can combine several narrow frequency band for temporal analysis of frequency changes (Fig. 6.1b). Overall, the possible variants of the inter-pulse delay detector are especially promising in view of the core circuit's ULP consumption rigorously evaluated. Further study on the core circuit is required to quantify the compromise between the range of detectable delays and the energy cost (determined by capacitances and transistors sizing), but delays in the order of few seconds are expected to be detectable with sub-microwatt power consumption.

Besides the inter-pulse delay detector use cases in recognition, it is applicable to SSL and navigation. Firstly, the rebound mechanism is not only present in female field crickets neuronal processing, but also in echolocating bats for positioning and hunting [137]. In bats echolocating with frequency modulated biosonar signals for example, specific detection neurons tuned for particular delays fire spikes when inhibitory rebounds from the call and its echo coincide. It closely resembles the processing performed by the inter-pulse delay detector, therefore proving a perspective application in navigating systems. Among the animals that use echolocation, bats and cetaceans are the most commonly studied echolocating mammals, but humans with complete or partial blindness have also developed the ability to differentiate objects' location, material, or shape by using acoustic echolocation [138]. Hence, integration of the inter-pulse delay detector in an echolocation system based on temporal delays would benefit hearing aid devices since small chips and energy efficient processing are required.

Secondly, the detection of temporal delay is strongly similar to a Jeffress model. With multiple inter-pulse delay detectors, the delay between the onset spikes of two channels from a binaural microphone pair can be determined in a classification manner where one output neuron corresponds to a narrow range of ITDs. Naturally, the drawback of the Jeffress model over an HRD-based model is still present. The number of neurons greatly increases, especially for a coarse mapping of the ITD range and upstream frequency segmentation with an artificial cochlea. Also, the generation of the onset spikes would still depend on the sound detection method, a threshold in this work's case.

In the end, the HRD-based coincidence detector is still the best candidate for the extraction and spike encoding of ITDs. Through numerous experiments, the circuit was deemed incompatible with long onsets because of the threshold detection method employed. No matter

[6] CC BY-NC-SA 4.0

the isolation of a frequency, the location and frequency-dependent acoustic reception of the microphones rendered real-world tonal signals impossible to work with. Possibly without completely changing the ITD extraction model, the only solution identified as of now is to limit the frequency range studied in input and employ a phase coding scheme in a more biomimetic approach. Instead of detecting onsets with a non-zero threshold, a spike can be generated with zero-cross detection at each positive half-wave of the input signal. Considering the maximal ITD allowed by a binaural pair's baseline, the maximal frequency is determined according to the sound wave's period at which ITDs can be non-ambiguously distinguished. In that manner, even sounds with very slowly increasing amplitudes would be localized as long as they contain low frequency components. In fact, phase coding for ITD extraction is the method used in biology which limit the ITD cue usage to frequencies below 1.6 kHz (see section 3.1.1). Therefore, this variant does not necessarily extend the range of sound type with which the coincidence detector would be compatible with since many sounds are above the 1.6 kHz, but is still suitable for specific applications, especially where power consumption is a great concern.

Both neuromorphic tools have advantages and drawbacks that reflect a compromise between precision, power consumption, and application scope. However, their intended use is not to be standalone precision devices. Without the multilateration technique, the DOA estimator studied in chapter 4 does not directly estimate a location. Also, the inter-pulse delay detector raises a flag when a characteristic delay is presented but does not directly output the detection of a sound source from its temporal pattern. At the very least, a counter is needed to indicate that, for example, a male cricket is detected from two successive inter-pulse delays around 20 ms at 4.8 kHz. These precomputing tools are expected to be integrated possibly with additional preprocessing and mostly upstream of an SNN (or conventional processor) for a smarter processing of their outputs, in particular when using multiple instances of these tools.

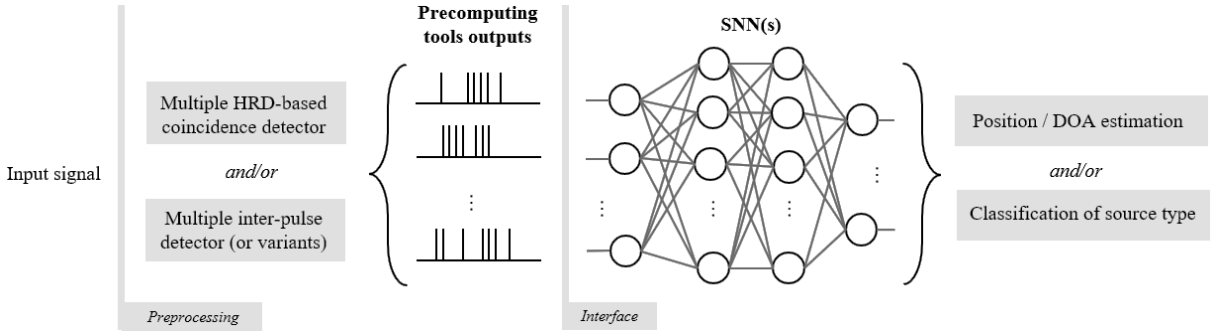


Fig. 6.2 The neuromorphic precomputing tools are interfaced with more complex systems for further and smarter processing. SNNs are more relevant downstream but conventional computing could also be used. An interface is necessary to translate spike into numeric values, for example using counters, or to introduce another spike encoding.

Moreover, their low power consumption and imperfect estimations make them suitable for providing an incertitude metric such that a more power consuming but precise device in sleep is awoken when necessary. The main objective of the proposed neuromorphic tools was to (or be able to) reach ULP consumption for which the precision performances were then evaluated.

Finally, the most anticipated (and under consideration) perspective applications of the localization and recognition precomputing tools are their combination in a single SSL system for tracking and source counting in a multisource context. In accordance with the main scope of the thesis on SSL, the inter-pulse delay detector was chosen as a simple and straightforward joint recognition mechanism for smarter localization. Concurrently, location estimations allow a better distinction between sound sources, and even more so when their signature are similar or identical. Localization and recognition tasks are complementary and carried out by the proposed circuits at a precomputing level. Fig. 6.3 illustrates a tracking and source counting task in a multisource context. In this situation, a recognition mechanism of temporal inter-click delays facilitates the extraction of the targe source's position where statistics and/or artificial intelligence would have been used to track the position.

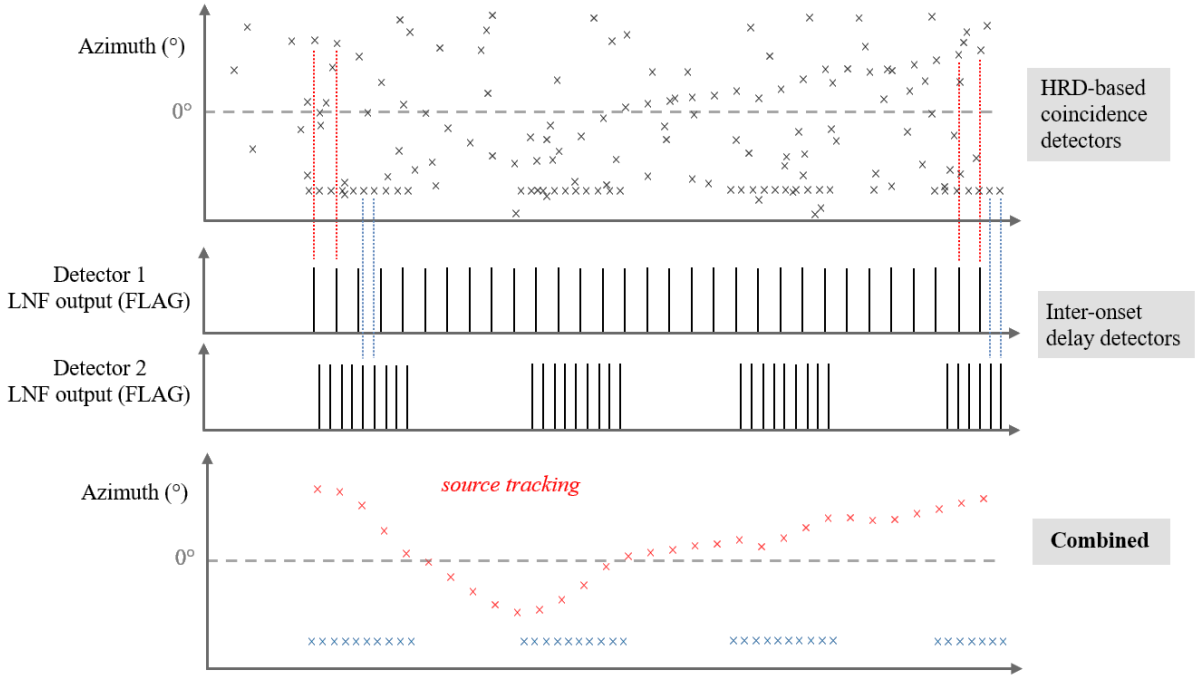


Fig. 6.3 Combined DOA estimation and rhythmic detection for sound source tracking and counting in a multisource context. Here, two sound sources (red and blue) are identified and distinguished from the noise thanks to their distinct location and temporal pattern. Resulting tracking and recognition are ideal.

6.2 Summary and Conclusion

This thesis investigates in the context of acoustic monitoring the potential of the mature subthreshold neuromorphic technology which encompasses CMOS ML neurons, synapses, and components operating in the subthreshold regime for high energy efficiency of its analog processing. More specifically, this work deals with the design, simulation, and on-chip implementation of simple SSL and/or recognition precomputing tools with this neuromorphic technology, then tested with artificial and real-world recordings.

From an analysis of the neuromorphic SSL literature and state-of-the-art systems, extracting ITD localization cues for estimation of a position produces the best accuracies. Following an HRD model of coincidence detection, a sparse architecture is obtained which extracts ITDs as a spike count and requires few neurons as opposed to the widely used Jeffress model. Using simplified multilateration technique based on hyperbolic intersection, localization performances of the proposed model are evaluated with diverse sounds played at distances between 24 cm and 10 m in 2-D and 3-D space. While experiments with tonal sounds do not lead to any coherent results, the HRD-based coincidence detector shows great DOA precision on click-like sounds in light of its simple architecture and limitations, with mean accuracies above 73.9% (±2.5°) and 77% (±5°) in the 2-D space between 1 m and 3 m with a collectively and individually suitable detection threshold V_{sat}.

Concurrently and with the motivation to address multisource scenarios, a joint recognition mechanism is adapted from a biological sparse neuronal processing observed in female field crickets. Unlike the common approach in neuromorphic literature of employing densely connected SNNs, an inter-pulse delay detection allows the recognition of simple characteristic temporal features in songs.

Made compatible with the subthreshold neuromorphic technology and using the ML *Slow* neurons, this bioinspired neuronal circuit is successfully integrated on chip, and also, high detection precision is reported on real-world cricket stridulation recordings in quiet and noisy soundscapes, with a power consumption fulfilling an ultra-low characterization of at most 1.2 nW and 570 pW in standby. Moreover, automatization of the weight tuning is investigated for an online adaptation of the circuit parameters to its environment with (un)supervised training schemes. Since this thesis objectives are focused on the application, user interaction must be included in this tool's discussion. The selectivity of the delays is tunable and depends on the sounds of interest, thus automatized tuning of the key parameters and extraction of training flags are proposed for further integration with an SNN performing online learning. Enhanced variants of the circuit suggest a wide range of use cases for multiple temporal feature extraction as an input vector to a subsequent neural network.

The studies initiated in this thesis open on numerous perspectives including the aforementioned potential enhancements and applications. While state-of-the-art DOA estimation performances are not achieved, the proposed ITD extraction circuit is assured to reach even lower power consumptions than the inter-pulse delay detector, that is below the nanowatt, thanks to an even sparser architecture and use of the more energy efficient ML *Fast* neurons. Among reported power consumptions in neuromorphic recognition, the implementation of the inter-pulse delay detector surpasses the best performance by a 10^3 factor, implying the possible implementation of a temporal feature extraction layer w/o a complementary pattern detector for a cost below the microwatt.

Validating models of localization and recognition in the subthreshold neuromorphic technology opens the path to ULP applications for the acoustic monitoring of biodiversity activities, but not only. Acoustic input signal may originate from animal or anthropogenic activities. In fact, any signal can be passed as input to both circuits to extract the corresponding temporal features. Combined together, the two neuromorphic precomputing tools unlock

tracking and source counting applications, essential in monitoring. Additional work is required to fully assess their real potential with multiple sound sources, but first experimentations with artificial soundscapes reveal possibilities and challenges of the monitoring tasks considered.

Appendix A

This appendix provides supplementary information on chapter 4 regarding 3-D localization method and additional results.

A.1 Multilateration Technique for a 4-Microphone Rectangular Array

With 4 microphones, the spatial location is computed from three ITD pairs following the same calculation steps. A rectangular configuration of the microphones shown in Fig. A.1 is considered, where the source coordinates are (x, y, z). The reference microphone is marked M, and the three other microphones E_i with $i \in \{1, 2, 3\}$. The binaural pairs (M, E_i) have perpendicular baselines D_i. We note δ_i the ITDs divided by the speed of sound of the binaural pairs with the microphone M and E_i microphones. Besides, we note $\delta_A = \delta_1 - \delta_2$, $\delta_B = \delta_1 - \delta_3$, and $\delta_C = \delta_2 - \delta_3$.

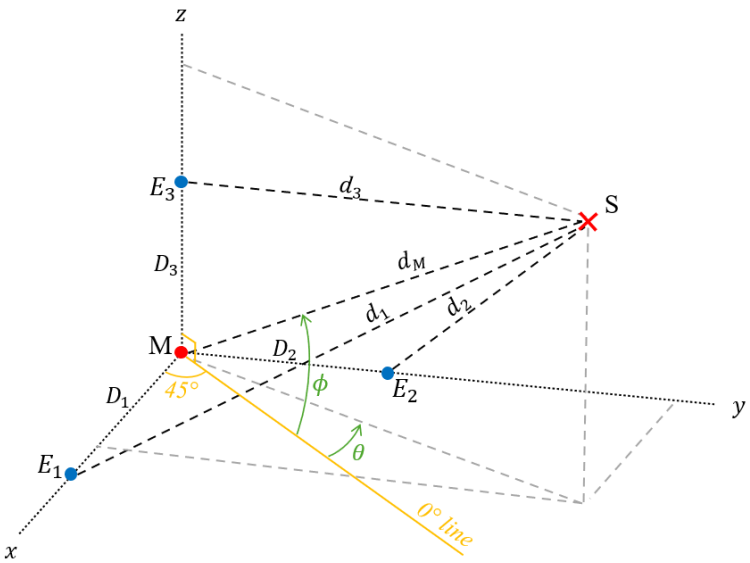


Fig. A.1 A 4-microphone rectangular configuration in 3-D space. Microphone M is the reference for estimation of the source's azimuth θ, elevation ϕ, and distance d_M, (M, E_i) are binaural pairs with baselines D_i. Projection lines in grey show the actual 3-D location of the source S.

With the microphones in a normalized configuration, namely M at coordinates $(0,0)$, E_1 at $(D_1,0,0)$, E_2 at $(0,D_2,0)$ and E_3 at $(0,0,D_3)$, δ_i^2 are computed to obtain two relations linking the three squared ITDs, the baselines, and the source coordinates, reorganized to express x and y as functions of z and the remainder,

$$x = \frac{\delta_1 D_3}{\delta_3 D_1} z - \frac{\delta_1 D_3^2}{2\delta_3 D_1} - \frac{\delta_1 \delta_B}{2D_1} + \frac{D_1}{2}\,, \tag{4.15}$$

$$y = \frac{\delta_2 D_3}{\delta_3 D_2} z - \frac{\delta_2 D_3^2}{2\delta_3 D_2} - \frac{\delta_2 \delta_C}{2D_2} + \frac{D_2}{2}. \tag{4.16}$$

The source coordinate z is found by using the hyperbolas of the three binaural pairs (M, E_i), leading to the characteristic expression of a 2nd order polynomial equation $Az^2 + Bz + C = 0$ with

$$\begin{aligned} A &= \frac{1}{a_3^2 c_1^2} - \frac{1}{a_1^2 c_3^2} + \frac{\delta_2^2 D_3^2}{\delta_3^2 D_2^2}\left(\frac{1}{a_3^2 b_1^2} + \frac{1}{a_1^2 b_3^2}\right) \\ B &= \left(-\frac{\delta_2^2 D_3 \delta_C}{\delta_3 D_2^2} - \frac{\delta_2^2 D_3^3}{\delta_3^2 D_2^2} + \frac{\delta_2 D_3}{\delta_3}\right)\left(\frac{1}{a_3^2 b_1^2} + \frac{1}{a_1^2 b_3^2}\right) + \frac{\delta_1 D_3}{a_1^2 a_{3_3^{2\delta}}} + \frac{D_3}{a_1^2 c_3^2} \\ C &= \left(\frac{1}{a_3^2 b_1^2} + \frac{1}{a_1^2 b_3^2}\right)\left(\frac{\delta_2^2 \delta_C^2}{4D_2^2} + \frac{\delta_2^2 D_3^4}{4\delta_3^2 D_2^2} + \frac{D_2^2}{4} + \frac{\delta_2^2 \delta_C D_3^2}{2\delta_3 D_2^2} - \frac{\delta_2 D_3^2}{2\delta_3} - \frac{\delta_2 \delta_C}{2}\right) \\ &\quad + \frac{D_1}{a_3^2 a_1^2}\left(\frac{D_1}{2} - \frac{\delta_1 D_3^2}{2\delta_3 D_1} - \frac{\delta_1 \delta_B}{2D_1}\right) - \frac{D_1^2}{4a_3^2 a_1^2} - \frac{D_3^2}{4a_1^2 c_3^2} + \frac{1}{a_1^2} + \frac{1}{a_3^2}\,, \end{aligned} \tag{4.17}$$

where $a_i = \frac{\tau_i v_{son}}{2}$ and $b_i^2 = a_i^2 - \left(\frac{D_i}{2}\right)^2$ the semi-axes of the hyperboloids, with τ_i the time delay of the binaural pair with baseline D_i. Expressions A, B, and C do not have a more compact factorized form. Again, the correct z coordinate is selected knowing the sign of the time delays, then x and y coordinates of the source are deduced using (4.15) and (4.16).

Conversion to distance, azimuthal, and elevation angles is performed considering the 0° line like in 2-D. such that,

$$d_M = \sqrt{x^2 + y^2 + z^2} \tag{4.18}$$

$$\phi = \arcsin\left(\frac{z}{d_M}\right) \tag{4.19}$$

$$\theta = \theta_{max} - \operatorname{atan2}\left(\frac{y}{x}\right) \tag{4.20}$$

with d_M the distance from the reference microphone to the source, θ the source azimuth from the 0° azimuth, ϕ the source elevation from the plan (x, y) of the configuration, and θ_{max} the absolute azimuth value at maximum ITD equal to 45° for the rectangular configuration. An example of the resulting position estimation in 3-D space is given in Fig. A.2 for baselines of 17 cm and a given ITD set.

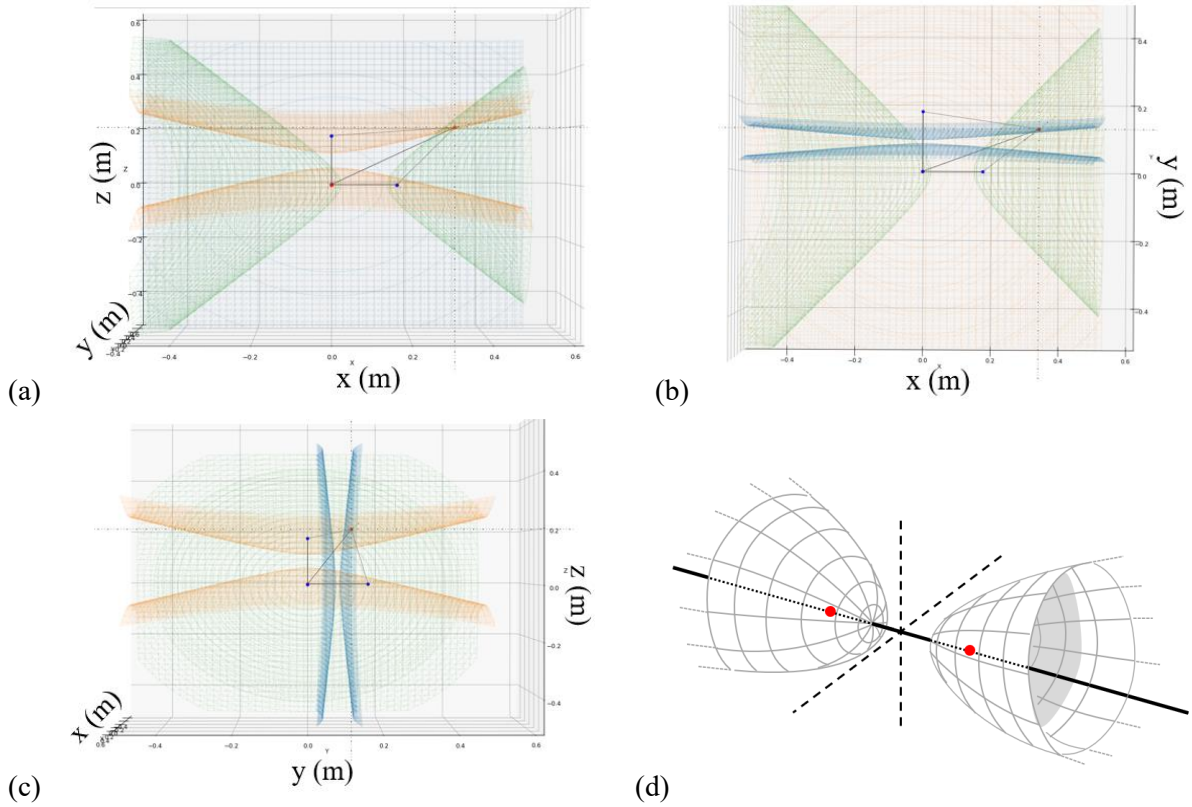


Fig. A.2 Position estimation in 3-D from hyperboloid intersection. Views of the (a) xz-plane, (b) xy-plane, and (c) yz-plane. Hyperboloids are plotted as plane meshes in blue, red, and green, corresponding to ITDs -354 µs, -43 µs, and -158 µs respectively, such that the intersection is at 40 cm from microphone M, azimuth 25° and elevation 30°. $D_i = 17$ cm. (d) Illustration of a hyperboloid. The binaural pair's collectors are red dots. Construction lines indicate the perspective.

A.2 Encoding at a Lower V_{DD}

The model is tested with ideal parametrization and especially with a high encoding supply voltage V_{DD} at encoding. For hardware considerations, 3-D localization performances are also evaluated at $V_{DD} = 300$ mV such that a unique power supply remains.

It is less power consuming to supply the ML *Fast* with $V_{DD} = 300$ mV, since fewer spikes mean a lower dynamic power consumption. Although lower than 1 MHz, the spiking frequency already reach around 270 kHz, leading to a time resolution of 3.7 µs. At V_{DD} = 300 mV, an excitation of 35 µs and 500 µs long (at amplitude 300 mV) of the ML *Fast* results in 9 and 135 spikes. Because f_s is still lower than f_{spike}, then the baseline at 17 cm for binaural angular resolution of 1° remains valid.

For evaluation of the performances with these parameters, differences are reported for the 3-D recording set G described in section 4.2.3 with saturation threshold V_{sat3}. Results with both V_{DD} are resumed in Table A.1 for comparison. Higher precision is observed with a higher temporal resolution, but results for V_{DD} at 300 mV do not fall far behind. It confirms the possibility to switch without much performance drops to a lower encoding V_{DD}, that may be more suitable for a hardware implementation.

Table A.1 Comparison of DOA performances with lower V_{DD} in set D.

Encoding V_{DD} (mV)	**400**	**300**
Temporal resolution (µs)	1	3.7
Azimuth accuracy ±2.5° (%)	31.24	29.67
Azimuth accuracy ±5° (%)	68.44	66.62
Azimuth accuracy ±10° (%)	73.65	73.21
Elevation accuracy 2.5° (%)	67.82	67.75
Elevation accuracy ±5° (%)	73.09	72.71
Elevation accuracy ±10° (%)	74.72	74.28

It is not so surprising to observe similar DOA estimation precision with encoding V_{DD} at 300 mV. The temporal resolution deteriorates but only by 3.7 µs. In the end, the resulting drop in angular resolution is small in front of the HRD-based coincidence detection's error, and is smaller than the recording sampling rate (5.2 µs steps for $f_s = 192$ kHz). The difference between the actual ITD detected and the discretization induced by the spike encoding may increase the ITD error or may slightly correct it towards the ground truth, thus improving the DOA estimations.

A.3 Adaptable Saturation Threshold

The strong dependence of the DOA estimation performances is problematic. For V_{sat} set too low, noises or small sound waves preceding the onset of the target sounds are detected, creating incorrect front matching or sound detections. For V_{sat} set too high, the influence of ILDs are stronger which may prevent solutions to the multilateration method or create missing detections of sounds with low intensities (quiet or far).

Chosen manually according to the noise level, a valuable enhancement would be the addition of an adaptable V_{sat} that slowly evolves according to the mean acoustic signal intensity collected by the microphones over a representative time frame. In the recordings created for this chapter, clicks are separated by long periods of silence (ambient noise) compared to the clicks' duration. Integration of the rectified input signal by a R-C filter with a large time constant extracts its envelope with low-pass filtering of the fluctuations, or in other words impulsive noises and sounds are smoothed. For a quick evaluation of the proposed enhancement, we define the adaptative saturation threshold as

$$V_{sat} = a\hat{e}(t'), \tag{4.23}$$

where $\hat{e}(t')$ is the mean value of the signal's envelope extracted using an R-C filter on a sliding time window defined by $t' \in [t-\tau, t]$. a is a bias parameter related to the envelope extraction parameters. t is the time variable, and τ the length of the time window.

Fig. A.3 gives an example of the adaptative V_{sat} evolution in time computed in Python on one recording of set G. Only the channel from microphone M is used for the computations. For a 10 s sliding time window to compute the mean and a set adequately, the adaptative V_{sat} allows to follow the noise floor unlike fixed V_{sat}. With a fixed threshold, the evolution of the noise cannot be taken into account. On Fig. A.3a, V_{sat1} detects many noises, and V_{sat2} appears to be

the most suited, but do not take into account the increase in noise intensity at the beginning of the recording. Visible on Fig. A.3b, higher values of fixed V_{sat} are not subject to noise but to ILDs; also, quiet clicks are not detected. At $a = 4$, the adaptative V_{sat} also has difficulty in handling the increase in noise at the beginning, however we can observe V_{sat} rising to handle the noise increase. With $a = 5$, all clicks are still detected, and the initial noise increase is bypassed.

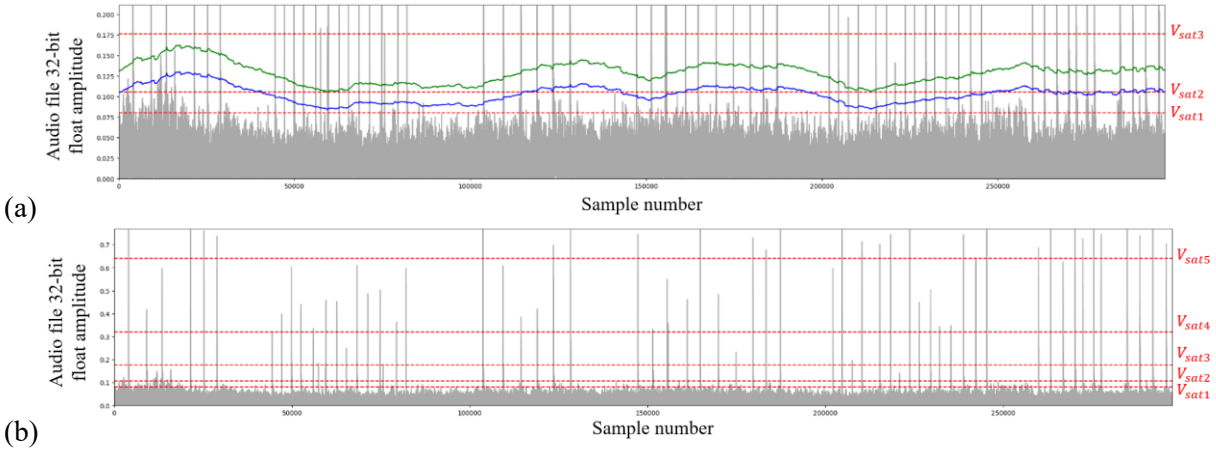


Fig. A.3 Adaptative V_{sat} according to (4.23) for $\tau = 10$ s, with $a = 4$ in blue, and $a = 5$ in green. (a) Zoom on the adaptative V_{sat} curve. (b) Fixed V_{sat} only. The solid blue line is the adaptative value, whereas the dashed lines are the fixed threshold of Table 4.2. The grey waveform is a superposition of all channels' envelope of a recording in set D. Values are normalized according to the envelope max amplitude.

Table A.2 Detection and DOA average performances with adaptative and fixed V_{sat}.

V_{sat}	Adaptative $a = 4$	Adaptative $a = 5$	Fixed V_{sat2}	Fixed V_{sat3}
Click detections (%) *	92.1	**98.4**	84.1	**98.4**
Number of correct detections	49	**54**	43	52
Number of incorrect detections	9	**8**	10	10
Number of noise detections	19	**3**	24	**3**
Azimuth accuracy ±5° (%)	63.64	**84.62**	63.53	78.79
Azimuth accuracy ±10° (%)	63.64	**86.15**	65.88	81.82
Elevation accuracy ±2.5° (%)	63.64	**83.08**	60.0	80.3
Elevation accuracy ±5° (%)	63.64	**83.08**	62.35	81.82
Elevation accuracy ±10° (%)	67.53	**86.15**	67.06	83.33

* On 63 clicks to detect in the chosen recording of set D.

For analysis, we take as example a single recording from the set G, namely at -10° azimuth and 40° elevation. Table A.2 resumes the keys results obtained with adaptative and fixed V_{sat}. Correct or incorrect detections refer to wavefront matching. Accuracies are specified in percentages, and best results are highlighted in bold font. Azimuth accuracy at tolerances ≤2.5°

is lower than 5% for all thresholds (supposed placement error previously mentioned in section 4.2.3) and are thus not added to the table. DOA performances show better accuracies when using a suited adaptative threshold. There is a less noticeable difference between V_{sat2} and the mean value of adaptative V_{sat} with $a = 4$, but the adaptative value have better click detection by 8%. For lower V_{sat} values, detection performances are lower, but DOA estimations are better, which can be observed between adaptative values only or fixed values only. Consequently, it is safe to say that implementing an adaptative saturation threshold would improve the overall performances. Although a better formula could be proposed, a compromise between target sound detection and DOA estimation performances would certainly still need to be made.

Appendix B

This appendix provides supplementary information on chapter 5 regarding the inter-pulse delay detection hardware implementation and additional circuit design of the proposed automatization.

B.1 Partial Hardware Implementation of Leak Optimization

The leak optimization circuit was partially implemented in hardware to test its correct operation. In the demonstrator, no counter or DAC is implemented and Wl is kept manually tunable. Instead, the control signal $\overline{FLAG}$ is routed to a pad (see Fig. 5.2c) and is visualized on the oscilloscope. Unexpectedly, a supposed crosstalk between electrical lines on the chip led to a dependence between excitatory currents and membrane potential of the neurons LN2, LN5, and LNnF. LNnF and its expander's activity reduces the effect of the excitation from AN1 to LN2, as well as the maximal duration of the expanders from LN2 to LN5.

With these interferences, the rebound is greatly impacted as shown in Fig. B.1. For lower values of W_{EnF} (other weights are unchanged), LNnF's spikes are more extended temporally, but this results in fewer spikes at LN2, lower inter-pulse delays detected, and narrower T_d. In Fig. B.1, the control signal $\overline{FLAG}$ is not tuned to generate a single square pulse otherwise T_d falls to 0 ms width, supposedly from a complete inactivity of LN2 and disappearance of the rebound (LN2 and LN5 activity is not directly observable).

Nevertheless, the operation of the leak optimization circuit could still be partially validated. On Fig. B.2, weights are tuned to successfully perform a detection of 20 ms inter-pulse delays. Activity at $\overline{FLAG}$ reproduces the activity of LN2. The neuron LNF correctly detects the 20 ms inter-pulse delay of the ideal square pulse input and inhibits LNnF such that $\overline{FLAG}$ is at 0 mV. However, LNF does not detect the following inter-pulse delay as a result of inhibiting LNnF.

It cannot be verified on the oscilloscope with the few outputs accessible through the pads, but it is likely that LN2 becomes dependent on LNnF's activity when the excitatory synapse from LN2 to LNnF is enabled. Therefore, when LNnF is inhibited during a correct detection, LN2 does not spike, and no rebound is generated.

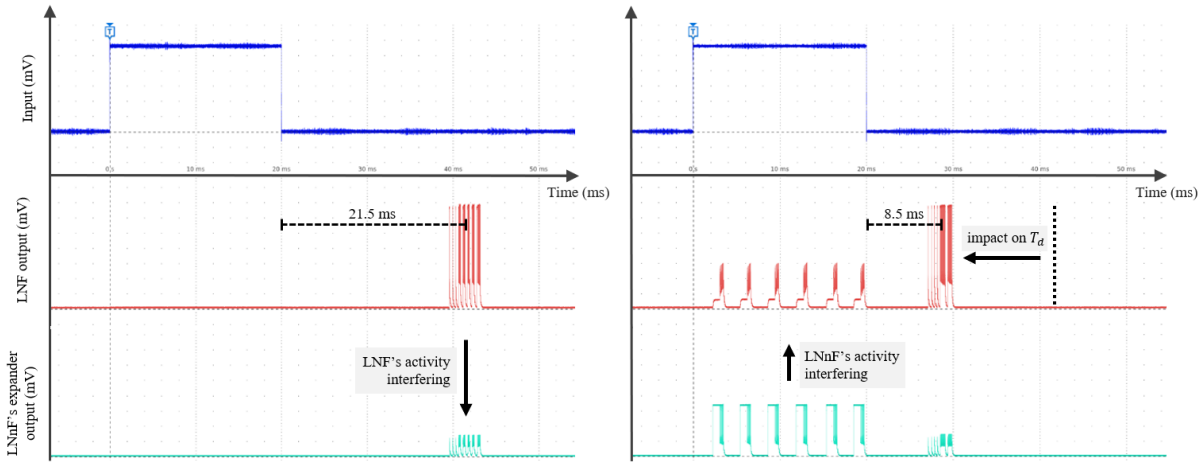


Fig. B.1 Crosstalk between the leak optimization circuit and LN2 and LN5's activity observed on oscilloscope for an ideal square pulse input. The rebound is visualized at LNF according to the calibration procedure explained in section 5.1.4. (a) No excitation is provided to LNnF, the rebound is tuned to detect inter-pulse delay between 19 ms and 23 ms approximately. The arrow shows activity from LN5 is visible at $\overline{FLAG}$. (b) Activity of LNnF (visualized at LNnF's expander digital output) reduces the rebound's width and delay from the input. We_{2nF} and W_{EnF} are set to V_{DD} such that the excitation from LN2 to LNnF is enabled and the expander in output of LNnF simply reproduces LN2's spikes.

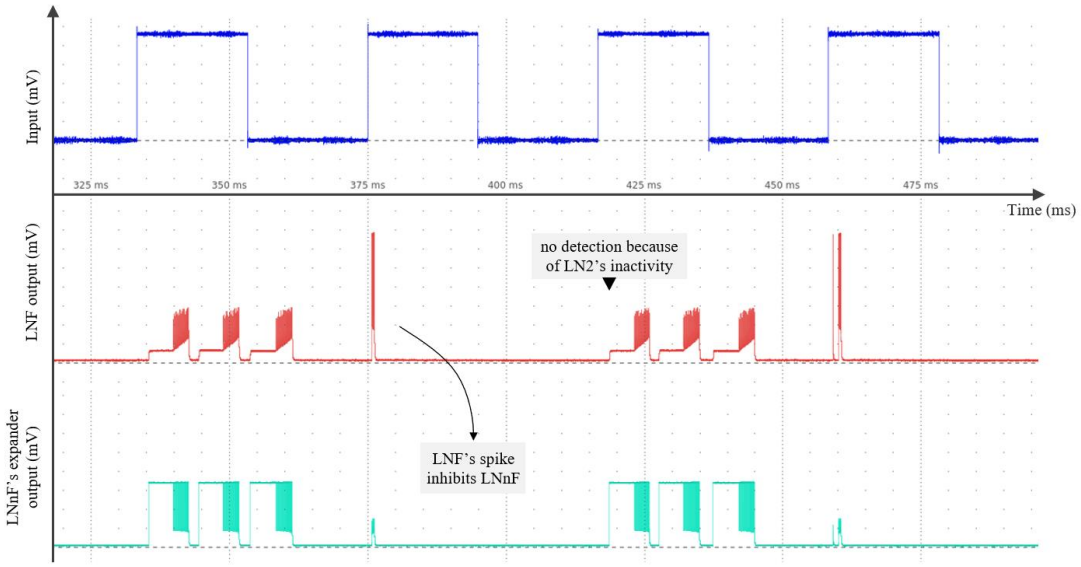


Fig. B.2 Partial validation of the control signal's operation in leak optimization circuit integrated in the demonstrator. An ideal square pulse is shown to the circuit. The circuit alternates between detections and missed detections as a consequence of crosstalk in the chip. The expander in output of LNnF is not set to create a single square signal per call as a control signal but several, otherwise the overall circuit stops operating correctly.

In the end, the proposed optimization process cannot be used as is. Yet, its validation in simulation, partial validation in hardware, and the observations made on a possible crosstalk suggest that a successful implementation of this optimization circuit is possible with some alterations.

B.2 Combining Automatization Circuits

Two circuits have been proposed to respond at automatized tuning problematics. However, a redundancy is present between the output of LNnF's expander informing on missed detections and the flags $\Delta < T_d$ and $\Delta > T_d$. In fact, both automatization circuits provide a signal corresponding to $\Delta \notin T_d$. The two proposed circuits for automatization of the inter-pulse delay detector's tuning can be aggregated into one as depicted in Fig. B.3.

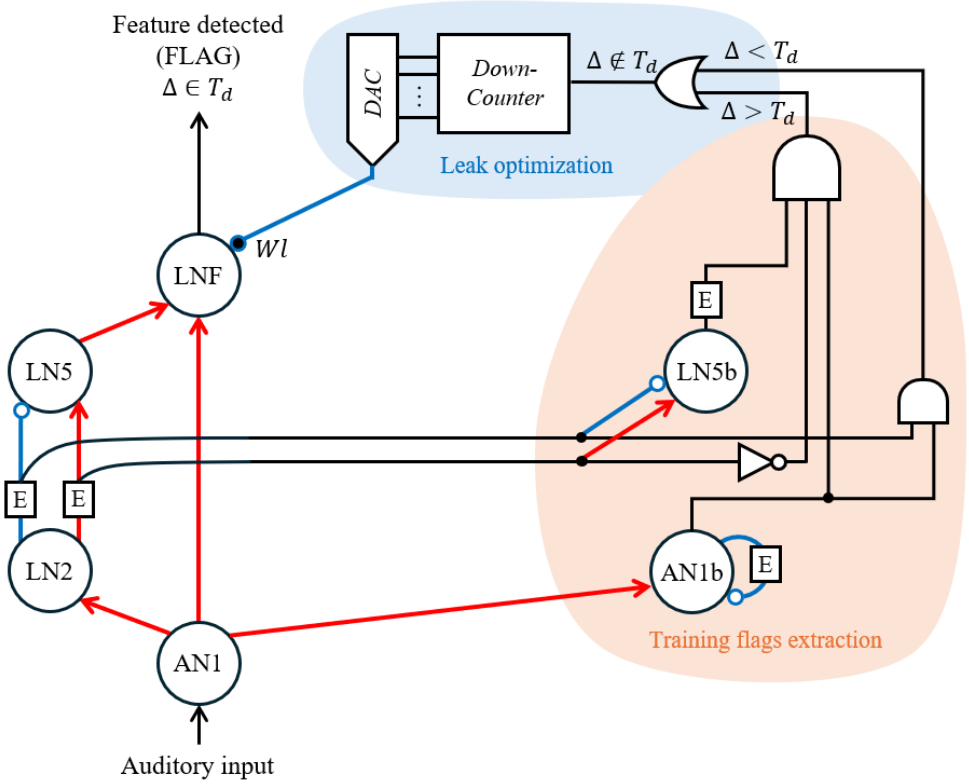


Fig. B.3 Combination of the leak optimization circuit with the training flag extraction circuit for a more compact design. The flags $\Delta < T_d$ and $\Delta > T_d$ are in fact a decomposed "missed detection" flags, that way they can both be used to decrease the value of Wl. The OR-gate can also be a neuron receiving full excitation from the two AND-gates output.

Moreover, the process for the extraction of the flags $\Delta > T_d$ and $\Delta < T_d$ allows to optimize Wl in an unsupervised manner. Initially, the first call also made the counter decrease as if the maximal Δ was infinite. By using the processing of the training flags extraction circuit, Wl now only decreases for $\Delta \leq \Delta_{max}$, with Δ_{max} the maximal delay considered as an *inter-pulse* delay determined by the expander in output of LN5b.

References

[1] "Executive summary – Energy and AI – Analysis," IEA. [Online]. Available: https://www.iea.org/reports/energy-and-ai/executive-summary

[2] S. Datta, W. Chakraborty, and M. Radosavljevic, "Toward attojoule switching energy in logic transistors," Science, vol. 378, no. 6621, pp. 733–740, Nov. 2022, doi: 10.1126/science.ade7656.

[3] W. Cao et al., "The future transistors," Nature, vol. 620, no. 7974, pp. 501–515, Aug. 2023, doi: 10.1038/s41586-023-06145-x.

[4] Y. Hu, Y. Liu, and Z. Liu, "A Survey on Convolutional Neural Network Accelerators: GPU, FPGA and ASIC," in 2022 14th International Conference on Computer Research and Development (ICCRD), Jan. 2022, pp. 100–107. doi: 10.1109/ICCRD54409.2022.9730377.

[5] R. Muralidhar, R. Borovica-Gajic, and R. Buyya, "Energy Efficient Computing Systems: Architectures, Abstractions and Modeling to Techniques and Standards," ACM Comput Surv, vol. 54, no. 11s, p. 236:1-236:37, Sept. 2022, doi: 10.1145/3511094.

[6] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition. Cambridge University Press, 2014.

[7] A. Cappy, Neuro-inspired Information Processing. John Wiley & Sons, 2020.

[8] S. E. Hyman, "Neurotransmitters," Curr. Biol., vol. 15, no. 5, pp. R154–R158, Mar. 2005, doi: 10.1016/j.cub.2005.02.037.

[9] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," Bull. Math. Biol., vol. 52, no. 1, pp. 25–71, Jan. 1990, doi: 10.1016/S0092-8240(05)80004-7.

[10] Y. Wei, G. Ullah, and S. J. Schiff, "Unification of Neuronal Spikes, Seizures, and Spreading Depression | Journal of Neuroscience," J. Neurosci., vol. 34, no. 35, pp. 11733–11743, Aug. 2014, doi: https://doi.org/10.1523/JNEUROSCI.0516-14.2014.

[11] R. Fitzhugh, "Thresholds and Plateaus in the Hodgkin-Huxley Nerve Equations," J. Gen. Physiol., vol. 43, no. 5, pp. 867–896, May 1960, doi: 10.1085/jgp.43.5.867.

[12] R. FitzHugh, "Impulses and Physiological States in Theoretical Models of Nerve Membrane," Biophys. J., vol. 1, no. 6, pp. 445–466, July 1961, doi: 10.1016/S0006-3495(61)86902-6.

[13] J. Nagumo, S. Arimoto, and S. Yoshizawa, "An Active Pulse Transmission Line Simulating Nerve Axon," Proc. IRE, vol. 50, no. 10, pp. 2061–2070, Oct. 1962, doi: 10.1109/JRPROC.1962.288235.

[14] C. Morris and H. Lecar, "Voltage oscillations in the barnacle giant muscle fiber," Biophys. J., vol. 35, no. 1, pp. 193–213, July 1981, doi: 10.1016/S0006-3495(81)84782-0.

[15] L. F. Abbott, “Lapicque’s introduction of the integrate-and-fire model neuron (1907),” Brain Res. Bull., vol. 50, no. 5–6, pp. 303–304, Nov. 1999, doi: 10.1016/S0361-9230(99)00161-6.

[16] A. N. Burkitt, “A Review of the Integrate-and-fire Neuron Model: I. Homogeneous Synaptic Input,” Biol. Cybern., vol. 95, no. 1, pp. 1–19, July 2006, doi: 10.1007/s00422-006-0068-6.

[17] R. Jolivet, T. J. Lewis, and W. Gerstner, “Generalized Integrate-and-Fire Models of Neuronal Activity Approximate Spike Trains of a Detailed Model to a High Degree of Accuracy,” J. Neurophysiol., vol. 92, no. 2, pp. 959–976, Aug. 2004, doi: 10.1152/jn.00190.2004.

[18] L. Badel, S. Lefort, R. Brette, C. C. H. Petersen, W. Gerstner, and M. J. E. Richardson, “Dynamic I-V Curves Are Reliable Predictors of Naturalistic Pyramidal-Neuron Voltage Traces,” J. Neurophysiol., vol. 99, no. 2, pp. 656–666, Feb. 2008, doi: 10.1152/jn.01107.2007.

[19] P. E. Latham, B. J. Richmond, P. G. Nelson, and S. Nirenberg, “Intrinsic Dynamics in Neuronal Networks. I. Theory,” J. Neurophysiol., vol. 83, no. 2, pp. 808–827, Feb. 2000, doi: 10.1152/jn.2000.83.2.808.

[20] E. M. Izhikevich, “Simple model of spiking neurons,” IEEE Trans. Neural Netw., vol. 14, no. 6, pp. 1569–1572, Nov. 2003, doi: 10.1109/TNN.2003.820440.

[21] J. Zhu, T. Zhang, Y. Yang, and R. Huang, “A comprehensive review on emerging artificial neuromorphic devices,” Appl. Phys. Rev., vol. 7, no. 1, p. 011312, Feb. 2020, doi: 10.1063/1.5118217.

[22] E. M. Izhikevich, “Which model to use for cortical spiking neurons?,” IEEE Trans. Neural Netw., vol. 15, no. 5, pp. 1063–1070, Sept. 2004, doi: 10.1109/TNN.2004.832719.

[23] C. Schuman, S. Kulkarni, M. Parsa, J. Mitchell, P. Date, and B. Kay, “Opportunities for neuromorphic computing algorithms and applications,” Nat. Comput. Sci., vol. 2, pp. 10–19, Jan. 2022, doi: 10.1038/s43588-021-00184-y.

[24] A. Javanshir, T. T. Nguyen, M. A. P. Mahmud, and A. Z. Kouzani, “Advancements in Algorithms and Neuromorphic Hardware for Spiking Neural Networks,” Neural Comput., vol. 34, no. 6, pp. 1289–1328, May 2022, doi: 10.1162/neco_a_01499.

[25] M. Stimberg, R. Brette, and D. F. Goodman, “Brian 2, an intuitive and efficient neural simulator,” eLife, vol. 8, p. e47314, Aug. 2019, doi: 10.7554/eLife.47314.

[26] J. Schemmel, D. Brüderle, A. Grübl, M. Hock, K. Meier, and S. Millner, “A wafer-scale neuromorphic hardware system for large-scale neural modeling,” in 2010 IEEE International Symposium on Circuits and Systems (ISCAS), May 2010, pp. 1947–1950. doi: 10.1109/ISCAS.2010.5536970.

[27] S. B. Furber et al., “Overview of the SpiNNaker System Architecture,” IEEE Trans. Comput., vol. 62, no. 12, pp. 2454–2467, Dec. 2013, doi: 10.1109/TC.2012.142.

[28] F. Akopyan et al., “TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip,” IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 34, no. 10, pp. 1537–1557, Oct. 2015, doi: 10.1109/TCAD.2015.2474396.

[29] M. Davies et al., “Loihi: A Neuromorphic Manycore Processor with On-Chip Learning,” IEEE Micro, vol. 38, no. 1, pp. 82–99, Jan. 2018, doi: 10.1109/MM.2018.112130359.

[30] M. Davies et al., “Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook,” Proc. IEEE, vol. PP, pp. 1–24, Apr. 2021, doi: 10.1109/JPROC.2021.3067593.

[31] H. Bos and D. Muir, “Sub-mW Neuromorphic SNN Audio Processing Applications with Rockpool and Xylo,” in Embedded Artificial Intelligence, River Publishers, 2023.

[32] P. Blouw, X. Choo, E. Hunsberger, and C. Eliasmith, “Benchmarking Keyword Spotting Efficiency on Neuromorphic Hardware,” in Proceedings of the 7th Annual Neuro-inspired Computational Elements Workshop, in NICE ’19. New York, NY, USA: Association for Computing Machinery, Mar. 2019, pp. 1–8. doi: 10.1145/3320288.3320304.

[33] C. Ostrau, C. Klarhorst, M. Thies, and U. Rückert, “Benchmarking Neuromorphic Hardware and Its Energy Expenditure,” Front. Neurosci., vol. 16, June 2022, doi: 10.3389/fnins.2022.873935.

[34] S. Moradi, N. Qiao, F. Stefanini, and G. Indiveri, “A Scalable Multicore Architecture With Heterogeneous Memory Structures for Dynamic Neuromorphic Asynchronous Processors (DYNAPs),” IEEE Trans. Biomed. Circuits Syst., vol. 12, no. 1, pp. 106–122, Feb. 2018, doi: 10.1109/TBCAS.2017.2759700.

[35] O. Richter et al., “DYNAP-SE2: a scalable multi-core dynamic neuromorphic asynchronous spiking neural network processor,” Neuromorphic Comput. Eng., vol. 4, no. 1, p. 014003, Jan. 2024, doi: 10.1088/2634-4386/ad1cd7.

[36] F. Jebali et al., “Powering AI at the edge: A robust, memristor-based binarized neural network with near-memory computing and miniaturized solar cell,” Nat. Commun., vol. 15, no. 1, p. 741, Jan. 2024, doi: 10.1038/s41467-024-44766-6.

[37] B. Gökgöz, F. Gül, and T. Aydın, “An overview memristor based hardware accelerators for deep neural network,” Concurr. Comput. Pract. Exp., vol. 36, no. 9, p. e7997, Apr. 2024, doi: 10.1002/cpe.7997.

[38] X. Duan et al., “Memristor-Based Neuromorphic Chips,” Adv. Mater., vol. 36, no. 14, p. 2310704, Apr. 2024, doi: 10.1002/adma.202310704.

[39] A. Mehonic et al., “Roadmap to Neuromorphic Computing with Emerging Technologies,” July 05, 2024, arXiv: arXiv:2407.02353. doi: 10.48550/arXiv.2407.02353.

[40] R. Li et al., “Photonics for Neuromorphic Computing: Fundamentals, Devices, and Opportunities,” Adv. Mater., vol. 37, no. 2, p. 2312825, 2025, doi: 10.1002/adma.202312825.

[41] X. Guo, J. Xiang, Y. Zhang, and Y. Su, “Integrated Neuromorphic Photonics: Synapses, Neurons, and Neural Networks,” Adv. Photonics Res., vol. 2, no. 6, p. 2000212, 2021, doi: 10.1002/adpr.202000212.

[42] A. S. Sedra and K. C. Smith, Microelectronic Circuits 7th Edition, International Edition. Oxford University Press, 2015.

[43] G. Indiveri et al., “Neuromorphic Silicon Neuron Circuits,” Front. Neurosci., vol. 5, May 2011, doi: 10.3389/fnins.2011.00073.

[44] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri, “Neuromorphic Electronic Circuits for Building Autonomous Cognitive Systems,” Proc. IEEE, vol. 102, no. 9, pp. 1367–1388, Sept. 2014, doi: 10.1109/JPROC.2014.2313954.

[45] I. Sourikopoulos et al., “A 4-fJ/Spike Artificial Neuron in 65 nm CMOS Technology,” Front. Neurosci., vol. 11, Mar. 2017, doi: 10.3389/fnins.2017.00123.

[46] D. O. Hebb, The Organization of Behavior: A Neuropsychological Theory. New York: Psychology Press, 2005. doi: 10.4324/9781410612403.

[47] G. Bi and M. Poo, “Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type,” J. Neurosci., vol. 18, no. 24, pp. 10464–10472, Dec. 1998, doi: 10.1523/JNEUROSCI.18-24-10464.1998.

[48] C. Loyez, K. Carpentier, I. Sourikopoulos, and F. Danneville, “Subthreshold neuromorphic devices for Spiking Neural Networks applied to embedded A.I,” in 2021 19th IEEE International New Circuits and Systems Conference (NEWCAS), June 2021, pp. 1–4. doi: 10.1109/NEWCAS50681.2021.9462779.

[49] F. Hassan et al., “State-of-the-Art Review on the Acoustic Emission Source Localization Techniques,” IEEE Access, vol. 9, pp. 101246–101266, 2021, doi: 10.1109/ACCESS.2021.3096930.

[50] D. Desai and N. Mehendale, “A Review on Sound Source Localization Systems,” Arch. Comput. Methods Eng., vol. 29, no. 7, pp. 4631–4642, Nov. 2022, doi: 10.1007/s11831-022-09747-2.

[51] G. Jekateryńczuk and Z. Piotrowski, “A Survey of Sound Source Localization and Detection Methods and Their Applications,” Sensors, vol. 24, no. 1, p. 68, Dec. 2023, doi: 10.3390/s24010068.

[52] A. R. Palmer and A. Rees, Eds., The Oxford Handbook of Auditory Science: The Auditory Brain, 1st ed. Oxford University Press, 2010. doi: 10.1093/oxfordhb/9780199233281.001.0001.

[53] M. N. Kunchur, “The human auditory system and audio,” Appl. Acoust., vol. 211, p. 109507, Aug. 2023, doi: 10.1016/j.apacoust.2023.109507.

[54] A. Carlini, C. Bordeau, and M. Ambard, “Auditory localization: a comprehensive practical review,” Front. Psychol., vol. 15, p. 1408073, July 2024, doi: 10.3389/fpsyg.2024.1408073.

[55] J. Schnupp, I. Nelken, and A. King, “Nerual Basis of Sound Localization,” in Auditory Neuroscience: Making Sense of Sound, MIT Press: Cambridge, MA, USA, 2011, pp. 177–221.

[56] B. Zonooz, E. Arani, K. P. Körding, P. A. T. R. Aalbers, T. Celikel, and A. J. Van Opstal, “Spectral Weighting Underlies Perceived Sound Elevation,” Sci. Rep., vol. 9, no. 1, p. 1642, Feb. 2019, doi: 10.1038/s41598-018-37537-z.

[57] R. Lyon, “A computational model of filtering, detection, and compression in the cochlea,” in ICASSP ’82. IEEE International Conference on Acoustics, Speech, and Signal Processing, Paris, France: Institute of Electrical and Electronics Engineers, 1982, pp. 1282–1285. doi: 10.1109/ICASSP.1982.1171644.

[58] M. S. A. Zilany and I. C. Bruce, “Representation of the vowel /ε/ in normal and impaired auditory nerve fibers: Model predictions of responses in cats,” J. Acoust. Soc. Am., vol. 122, no. 1, pp. 402–417, July 2007, doi: 10.1121/1.2735117.

[59] V. Chan, S.-C. Liu, and A. Van Schaik, “AER EAR: A Matched Silicon Cochlea Pair With Address Event Representation Interface,” IEEE Trans. Circuits Syst. Regul. Pap., vol. 54, no. 1, pp. 48–59, Jan. 2007, doi: 10.1109/TCSI.2006.887979.

[60] J. Liu, D. Perez-Gonzalez, A. Rees, H. Erwin, and S. Wermter, “A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation,” Neurocomputing, vol. 74, no. 1–3, pp. 129–139, Dec. 2010, doi: 10.1016/j.neucom.2009.10.030.

[61] A. Jiménez-Fernández et al., “A Binaural Neuromorphic Auditory Sensor for FPGA: A Spike Signal Processing Approach,” IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 4, pp. 804–818, Apr. 2017, doi: 10.1109/TNNLS.2016.2583223.

[62] T. Schoepe et al., “Closed-loop sound source localization in neuromorphic systems,” Neuromorphic Comput. Eng., vol. 3, no. 2, p. 024009, June 2023, doi: 10.1088/2634-4386/acdaba.

[63] C. Schauer, T. Zahn, P. Paschke, and H.-M. Gross, “Binaural sound localization in an artificial neural network,” in 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100), Istanbul, Turkey: IEEE, 2000, pp. II865–II868. doi: 10.1109/ICASSP.2000.859097.

[64] G. Ashida and C. E. Carr, “Sound localization: Jeffress and beyond,” Curr. Opin. Neurobiol., vol. 21, no. 5, pp. 745–751, Oct. 2011, doi: 10.1016/j.conb.2011.05.008.

[65] J. Lazzaro and C. A. Mead, “A Silicon Model Of Auditory Localization,” Neural Comput., vol. 1, no. 1, pp. 47–57, Mar. 1989, doi: 10.1162/neco.1989.1.1.47.

[66] E. I. Knudsen and M. Konishi, “A Neural Map of Auditory Space in the Owl,” Science, vol. 200, no. 4343, pp. 795–797, May 1978, doi: 10.1126/science.644324.

[67] R. F. Lyon and C. Mead, “An analog electronic cochlea,” IEEE Trans. Acoust. Speech Signal Process., vol. 36, no. 7, pp. 1119–1134, July 1988, doi: 10.1109/29.1639.

[68] J. A. Feldman and D. H. Ballard, “Connectionist Models and Their Properties,” Cogn. Sci., vol. 6, no. 3, pp. 205–254, July 1982, doi: 10.1207/s15516709cog0603_1.

[69] T. Horiuchi, “An Auditory Localization and Coordinate Transform Chip,” in Advances in Neural Information Processing Systems, G. Tesauro, D. Touretzky, and T. Leen, Eds., MIT Press, 1994.

[70] T. C. T. Yin, “Neural Mechanisms of Encoding Binaural Localization Cues in the Auditory Brainstem,” in Integrative Functions in the Mammalian Auditory Pathway, vol. 15, D. Oertel, R. R. Fay, and A. N. Popper, Eds., in Springer Handbook of Auditory Research, vol. 15. , New York, NY: Springer New York, 2002, pp. 99–159. doi: 10.1007/978-1-4757-3654-0_4.

[71] K. Voutsas and J. Adamy, “A Biologically Inspired Spiking Neural Network for Sound Source Lateralization,” IEEE Trans. Neural Netw., vol. 18, no. 6, pp. 1785–1799, Nov. 2007, doi: 10.1109/TNN.2007.899623.

[72] M. Kugler, K. Iwasa, V. A. P. Benso, S. Kuroyanagi, and A. Iwata, “A Complete Hardware Implementation of an Integrated Sound Localization and Classification System Based on Spiking Neural Networks,” in Neural Information Processing, vol. 4985, M. Ishikawa, K. Doya, H. Miyamoto, and T. Yamakawa, Eds., in Lecture Notes in Computer Science, vol. 4985. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 577–587. doi: 10.1007/978-3-540-69162-4_60.

[73] K. Iwasa, M. Kugler, S. Kuroyanagi, and A. Iwata, “A Sound Localization and Recognition System using Pulsed Neural Networks on FPGA,” in 2007 International Joint Conference on

Neural Networks, Orlando, FL, USA: IEEE, Aug. 2007, pp. 902–907. doi: 10.1109/IJCNN.2007.4371078.

[74] B. Glackin, J. Wall, T. M. McGinnity, L. P. Maguire, and L. J. McDaid, “A spiking neural network model of the medial superior olive using spike timing dependent plasticity for sound localization,” Front. Comput. Neurosci., 2010, doi: 10.3389/fncom.2010.00018.

[75] D. J. Tollin and K. Koka, “Postnatal development of sound pressure transformations by the head and pinnae of the cat: Monaural characteristics,” J. Acoust. Soc. Am., vol. 125, no. 2, pp. 980–994, Feb. 2009, doi: 10.1121/1.3058630.

[76] V. Y.-S. Chan, C. T. Jin, and A. V. Schaik, “Adaptive Sound Localization with a Silicon Cochlea Pair,” Front. Neurosci., vol. 4, 2010, doi: 10.3389/fnins.2010.00196.

[77] H. Finger and S.-C. Liu, “Estimating the location of a sound source with a spike-timing localization algorithm,” in 2011 IEEE International Symposium of Circuits and Systems (ISCAS), Rio de Janeiro, Brazil: IEEE, May 2011, pp. 2461–2464. doi: 10.1109/ISCAS.2011.5938102.

[78] J. A. Wall, T. M. McGinnity, and L. P. Maguire, “A comparison of sound localisation techniques using cross-correlation and spiking neural networks for mobile robotics,” in The 2011 International Joint Conference on Neural Networks, San Jose, CA, USA: IEEE, July 2011, pp. 1981–1987. doi: 10.1109/IJCNN.2011.6033468.

[79] B. Schrauwen and I. Van Campenhout, “BSA, a fast and accurate spike train encoding scheme,” in Proceedings of the International Joint Conference on Neural Networks, 2003., Portland, Oregon USA: IEEE, 2003, pp. 2825–2830. doi: 10.1109/IJCNN.2003.1224019.

[80] V. Y.-S. Chan, C. T. Jin, and A. van Schaik, “Neuromorphic audio–visual sensor fusion on a sound-localizing robot,” Front. Neurosci., vol. 6, 2012, doi: 10.3389/fnins.2012.00021.

[81] P. K. J. Park et al., “Fast neuromorphic sound localization for binaural hearing aids,” in 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka: IEEE, July 2013, pp. 5275–5278. doi: 10.1109/EMBC.2013.6610739.

[82] M. M. Faraji, S. B. Shouraki, and E. Iranmehr, “Spiking neural network for sound localization using microphone array,” in 2015 23rd Iranian Conference on Electrical Engineering, Tehran, Iran: IEEE, May 2015, pp. 1260–1265. doi: 10.1109/IranianCEE.2015.7146409.

[83] C. Beck, G. Garreau, and J. Georgiou, “Sound Source Localization through 8 MEMS Microphones Array Using a Sand-Scorpion-Inspired Spiking Neural Network,” Front. Neurosci., vol. 10, Oct. 2016, doi: 10.3389/fnins.2016.00479.

[84] J. Encke and W. Hemmert, “Extraction of Inter-Aural Time Differences Using a Spiking Neuron Network Model of the Medial Superior Olive,” Front. Neurosci., vol. 12, p. 140, Mar. 2018, doi: 10.3389/fnins.2018.00140.

[85] R. Luke and D. McAlpine, “A Spiking Neural Network Approach to Auditory Source Lateralisation,” in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK: IEEE, May 2019, pp. 1488–1492. doi: 10.1109/ICASSP.2019.8683767.

[86] T. Schoepe, D. Gutierrez-Galan, J. P. Dominguez-Morales, A. Jimenez-Fernandez, A. Linares-Barranco, and E. Chicca, “Neuromorphic Sensory Integration for Combining Sound Source

Localization and Collision Avoidance," in 2019 IEEE Biomedical Circuits and Systems Conference (BioCAS), Nara, Japan: IEEE, Oct. 2019, pp. 1–4. doi: 10.1109/BIOCAS.2019.8919202.

[87] Z. W. Song, S. Y. Xiang, Z. X. Ren, S. H. Wang, A. J. Wen, and Y. Hao, "Photonic spiking neural network based on excitable VCSELs-SA for sound azimuth detection," Opt. Express, vol. 28, no. 2, p. 1561, Jan. 2020, doi: 10.1364/OE.381229.

[88] Z. Pan, M. Zhang, J. Wu, J. Wang, and H. Li, "Multi-Tone Phase Coding of Interaural Time Difference for Sound Source Localization With Spiking Neural Networks," IEEEACM Trans. Audio Speech Lang. Process., vol. 29, pp. 2656–2670, 2021, doi: 10.1109/TASLP.2021.3100684.

[89] S. Zhong, Y. Zhang, H. Zheng, F. Yu, and R. Zhao, "Spike-Based Spatiotemporal Processing Enabled by Oscillation Neuron for Energy-Efficient Artificial Sensory Systems," Adv. Intell. Syst., vol. 4, no. 9, p. 2200076, Sept. 2022, doi: 10.1002/aisy.202200076.

[90] X. Chen, Q. Yang, J. Wu, H. Li, and K. C. Tan, "A Hybrid Neural Coding Approach for Pattern Recognition With Spiking Neural Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 46, no. 5, pp. 3064–3078, May 2024, doi: 10.1109/TPAMI.2023.3339211.

[91] M. B. Milde, O. J. N. Bertrand, H. Ramachandran, M. Egelhaaf, and E. Chicca, "Spiking Elementary Motion Detector in Neuromorphic Systems," Neural Comput., vol. 30, no. 9, pp. 2384–2417, Sept. 2018, doi: 10.1162/neco_a_01112.

[92] D. Gutierrez-Galan, T. Schoepe, J. P. Dominguez-Morales, A. Jimenez-Fernandez, E. Chicca, and A. Linares-Barranco, "An Event-Based Digital Time Difference Encoder Model Implementation for Neuromorphic Systems," IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 5, pp. 1959–1973, May 2022, doi: 10.1109/TNNLS.2021.3108047.

[93] J. A. Wall, L. J. McDaid, L. P. Maguire, and T. M. McGinnity, "Spiking Neural Network Model of Sound Localization Using the Interaural Intensity Difference," IEEE Trans. Neural Netw. Learn. Syst., vol. 23, no. 4, pp. 574–586, Apr. 2012, doi: 10.1109/TNNLS.2011.2178317.

[94] F. Xiao and D. Weibei, "A biologically plausible spiking model for interaural level difference processing auditory pathway in human brain," in 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada: IEEE, July 2016, pp. 5029–5036. doi: 10.1109/IJCNN.2016.7727862.

[95] X. Feng and W. Dou, "A Biologically Plausible Spiking Model of Human Auditory Pathways for Sound Localization," in Proceedings of the 2016 5th International Conference on Measurement, Instrumentation and Automation (ICMIA 2016), Shenzhen, China: Atlantis Press, 2016. doi: 10.2991/icmia-16.2016.67.

[96] E. C. Escudero, F. P. Peña, R. P. Vicente, A. Jimenez-Fernandez, G. J. Moreno, and A. Morgado-Estevez, "Real-time neuro-inspired sound source localization and tracking architecture applied to a robotic platform," Neurocomputing, vol. 283, pp. 129–139, Mar. 2018, doi: 10.1016/j.neucom.2017.12.041.

[97] T. Oess, M. Lohr, C. Jarvers, D. Schmid, and H. Neumann, "A Bio-Inspired Model of Sound Source Localization on Neuromorphic Hardware," in 2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), Genova, Italy: IEEE, Aug. 2020, pp. 103–107. doi: 10.1109/AICAS48895.2020.9073935.

[98] D. Schmid, T. Oess, and H. Neumann, “Listen to the Brain–Auditory Sound Source Localization in Neuromorphic Computing Architectures,” Sensors, vol. 23, no. 9, p. 4451, May 2023, doi: 10.3390/s23094451.

[99] J. Dávila-Chacón, S. Heinrich, J. Liu, and S. Wermter, “Biomimetic Binaural Sound Source Localisation with Ego-Noise Cancellation,” in Artificial Neural Networks and Machine Learning – ICANN 2012, vol. 7552, A. E. P. Villa, W. Duch, P. Érdi, F. Masulli, and G. Palm, Eds., in Lecture Notes in Computer Science, vol. 7552. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 239–246. doi: 10.1007/978-3-642-33269-2_31.

[100] D. F. M. Goodman and R. Brette, “Spike-Timing-Based Computation in Sound Localization,” PLoS Comput. Biol., vol. 6, no. 11, p. e1000993, Nov. 2010, doi: 10.1371/journal.pcbi.1000993.

[101] B. Gao et al., “Memristor-based analogue computing for brain-inspired sound localization with in situ training,” Nat. Commun., vol. 13, no. 1, p. 2026, Apr. 2022, doi: 10.1038/s41467-022-29712-8.

[102] Y. Xu et al., “An organic electrochemical synaptic transistor array for neuromorphic computation of sound localization,” Appl. Phys. Lett., vol. 123, no. 13, p. 133701, Sept. 2023, doi: 10.1063/5.0167865.

[103] Y. Li, J. Zhao, X. Xiao, R. Chen, and L. Wang, “Brain-Inspired Binaural Sound Source Localization Method Based on Liquid State Machine,” in Neural Information Processing, vol. 14449, B. Luo, L. Cheng, Z.-G. Wu, H. Li, and C. Li, Eds., in Lecture Notes in Computer Science, vol. 14449. , Singapore: Springer Nature Singapore, 2024, pp. 198–213. doi: 10.1007/978-981-99-8067-3_15.

[104] W. Maass, T. Natschläger, and H. Markram, “Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations,” Neural Comput., vol. 14, no. 11, pp. 2531–2560, Nov. 2002, doi: 10.1162/089976602760407955.

[105] Z. Roozbehi, A. Narayanan, M. Mohaghegh, and S.-A. Saeedinia, “Dynamic-Structured Reservoir Spiking Neural Network in Sound Localization,” IEEE Access, vol. 12, pp. 24596–24608, 2024, doi: 10.1109/ACCESS.2024.3360491.

[106] L. Deng et al., “Rethinking the performance comparison between SNNS and ANNS,” Neural Netw., vol. 121, pp. 294–307, Jan. 2020, doi: 10.1016/j.neunet.2019.09.005.

[107] F. Ottati et al., “To Spike or Not to Spike: A Digital Hardware Perspective on Deep Learning Acceleration,” IEEE J. Emerg. Sel. Top. Circuits Syst., vol. 13, no. 4, pp. 1015–1025, Dec. 2023, doi: 10.1109/JETCAS.2023.3330432.

[108] B. Hassenstein and W. Reichardt, “Der Schluß von Reiz-Reaktions-Funktionen auf System-Strukturen,” Z. Für Naturforschung B, vol. 8, no. 9, pp. 518–524, Sept. 1953, doi: 10.1515/znb-1953-0910.

[109] B. Hassenstein and W. Reichardt, “Systemtheoretische Analyse der Zeit-, Reihenfolgen- und Vorzeichenauswertung bei der Bewegungsperzeption des Rüsselkäfers Chlorophanus,” Z. Für Naturforschung B, vol. 11, no. 9–10, pp. 513–524, Oct. 1956, doi: 10.1515/znb-1956-9-1004.

[110] Y. T. Chan and K. C. Ho, “A simple and efficient estimator for hyperbolic location,” IEEE Trans. Signal Process., vol. 42, no. 8, pp. 1905–1915, Aug. 1994, doi: 10.1109/78.301830.

[111] B. Planqué and W.-P. Vellinga, “Xeno-canto:: Sharing bird sounds from around the world.” July 2017. [Online]. Available: http://www.xeno-canto.org/

[112] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, “The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer,” J. Acoust. Soc. Am., vol. 115, no. 2, pp. 833–843, Feb. 2004, doi: 10.1121/1.1639908.

[113] O. O. Abayomi-Alli, R. Damaševičius, A. Qazi, M. Adedoyin-Olowe, and S. Misra, “Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review,” Electronics, vol. 11, no. 22, p. 3795, Nov. 2022, doi: 10.3390/electronics11223795.

[114] S. Bhattacharya, N. Das, S. Sahu, A. Mondal, and S. Borah, “Deep Classification of Sound: A Concise Review,” in Proceeding of First Doctoral Symposium on Natural Computing Research, vol. 169, V. H. Patil, N. Dey, P. N. Mahalle, M. Shafi Pathan, and Vinod. V. Kimbahune, Eds., in Lecture Notes in Networks and Systems, vol. 169. , Singapore: Springer Singapore, 2021, pp. 33–43. doi: 10.1007/978-981-33-4073-2_4.

[115] X. Wu, B. Dang, T. Zhang, X. Wu, and Y. Yang, “Spatiotemporal audio feature extraction with dynamic memristor-based time-surface neurons,” Sci. Adv., vol. 10, no. 14, p. eadl2767, Apr. 2024, doi: 10.1126/sciadv.adl2767.

[116] X. Li, Y. Liu, L. Zheng, and W. Zhang, “A Lightweight Convolutional Spiking Neural Network for Fires Detection Based on Acoustics,” Electronics, vol. 13, no. 15, p. 2948, July 2024, doi: 10.3390/electronics13152948.

[117] S. Kshirasagar, B. Cramer, A. Guntoro, and C. Mayr, “Auditory Anomaly Detection using Recurrent Spiking Neural Networks,” in 2024 IEEE 6th International Conference on AI Circuits and Systems (AICAS), Abu Dhabi, United Arab Emirates: IEEE, Apr. 2024, pp. 278–281. doi: 10.1109/AICAS59952.2024.10595878.

[118] S. Kshirasagar, A. Guntoro, and C. Mayr, “Impact of Sliding Window Variation and Neuronal Time Constants on Acoustic Anomaly Detection Using Recurrent Spiking Neural Networks in Automotive Environment,” Algorithms, vol. 17, no. 10, p. 440, Oct. 2024, doi: 10.3390/a17100440.

[119] Z. Roozbehi, A. Narayanan, M. Mohaghegh, and S.-A. Saeedinia, “Enhanced Multiple Sound Event Detection and Classification Using Physical Signal Properties in Recurrent Spiking Neural Networks,” IEEE Access, vol. 13, pp. 81312–81325, 2025, doi: 10.1109/ACCESS.2025.3563346.

[120] F. Martinelli, G. Dellaferrera, P. Mainar, and M. Cernak, “Spiking Neural Networks Trained With Backpropagation for Low Power Neuromorphic Implementation of Voice Activity Detection,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain: IEEE, May 2020, pp. 8544–8548. doi: 10.1109/ICASSP40776.2020.9053412.

[121] G. Dellaferrera, F. Martinelli, and M. Cernak, “A Bin Encoding Training of a Spiking Neural Network Based Voice Activity Detection,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain: IEEE, May 2020, pp. 3207–3211. doi: 10.1109/ICASSP40776.2020.9054761.

[122] J. Wu, E. Yılmaz, M. Zhang, H. Li, and K. C. Tan, “Deep Spiking Neural Networks for Large Vocabulary Automatic Speech Recognition,” Front. Neurosci., vol. 14, p. 199, Mar. 2020, doi: 10.3389/fnins.2020.00199.

[123] M. Bensimon, S. Greenberg, and M. Haiut, “Using a Low-Power Spiking Continuous Time Neuron (SCTN) for Sound Signal Processing,” Sensors, vol. 21, no. 4, p. 1065, Feb. 2021, doi: 10.3390/s21041065.

[124] C.-C. Yang and T.-S. Chang, “A 71.2-μW Speech Recognition Accelerator With Recurrent Spiking Neural Network,” IEEE Trans. Circuits Syst. Regul. Pap., vol. 71, no. 7, pp. 3203–3213, July 2024, doi: 10.1109/TCSI.2024.3387993.

[125] P. Blouw, X. Choo, E. Hunsberger, and C. Eliasmith, “Benchmarking Keyword Spotting Efficiency on Neuromorphic Hardware,” in Proceedings of the 7th Annual Neuro-inspired Computational Elements Workshop, Albany NY USA: ACM, Mar. 2019, pp. 1–8. doi: 10.1145/3320288.3320304.

[126] S. A. Kotz, A. Ravignani, and W. T. Fitch, “The Evolution of Rhythm Processing,” Trends Cogn. Sci., vol. 22, no. 10, pp. 896–910, Oct. 2018, doi: 10.1016/j.tics.2018.08.002.

[127] S. Schöneich, K. Kostarakos, and B. Hedwig, “An auditory feature detection circuit for sound pattern recognition,” Sci. Adv., vol. 1, no. 8, p. e1500325, Sept. 2015, doi: 10.1126/sciadv.1500325.

[128] F. Sandin and M. Nilsson, “Synaptic Delays for Insect-Inspired Temporal Feature Detection in Dynamic Neuromorphic Processors,” Front. Neurosci., vol. 14, p. 150, Feb. 2020, doi: 10.3389/fnins.2020.00150.

[129] M. Nilsson, F. Liwicki, and F. Sandin, “Spatiotemporal Pattern Recognition in Single Mixed-Signal VLSI Neurons with Heterogeneous Dynamic Synapses,” in Proceedings of the International Conference on Neuromorphic Systems 2022, Knoxville TN USA: ACM, July 2022, pp. 1–8. doi: 10.1145/3546790.3546794.

[130] B. G. Hedwig, “Sequential Filtering Processes Shape Feature Detection in Crickets: A Framework for Song Pattern Recognition,” Front. Physiol., vol. 7, Feb. 2016, doi: 10.3389/fphys.2016.00046.

[131] J. Terven, D.-M. Cordova-Esparza, J.-A. Romero-González, A. Ramírez-Pedraza, and E. A. Chávez-Urbiola, “A comprehensive survey of loss functions and metrics in deep learning,” Artif. Intell. Rev., vol. 58, no. 7, p. 195, Apr. 2025, doi: 10.1007/s10462-025-11198-7.

[132] Z. Yi, J. Lian, Q. Liu, H. Zhu, D. Liang, and J. Liu, “Learning rules in spiking neural networks: A survey,” Neurocomputing, vol. 531, pp. 163–179, Apr. 2023, doi: 10.1016/j.neucom.2023.02.026.

[133] Y. Yarom and J. Hounsgaard, “Voltage Fluctuations in Neurons: Signal or Noise?,” Physiol. Rev., vol. 91, no. 3, pp. 917–929, July 2011, doi: 10.1152/physrev.00019.2010.

[134] N. Qiao et al., “A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses,” Front. Neurosci., vol. 9, Apr. 2015, doi: 10.3389/fnins.2015.00141.

[135] C. Bartolozzi and G. Indiveri, “Synaptic Dynamics in Analog VLSI,” Neural Comput., vol. 19, no. 10, pp. 2581–2603, Oct. 2007, doi: 10.1162/neco.2007.19.10.2581.

[136] L. Sayigh et al., “The Watkins Marine Mammal Sound Database: An online, freely accessible resource,” presented at the Fourth International Conference on the Effects of Noise on Aquatic Life, Dublin, Ireland, 2016, p. 040013. doi: 10.1121/2.0000358.

[137] M. J. Beetz and J. C. Hechavarría, “Neural Processing of Naturalistic Echolocation Signals in Bats,” Front. Neural Circuits, vol. 16, p. 899370, May 2022, doi: 10.3389/fncir.2022.899370.

[138] L. Thaler and M. A. Goodale, “Echolocation in humans: an overview,” WIREs Cogn. Sci., vol. 7, no. 6, pp. 382–393, Nov. 2016, doi: 10.1002/wcs.1408.

List of Publications

A part of the work described in chapters 3, 4, and 5 resulted in the publication of three papers [Au1], [Au2], and [Au3].

[Au1] provides a review of bioinspired sound source localization systems and echolocation-based navigation systems that use spiking neurons. Then, [Au2] reports the results of sound source localization with click-like sounds using the neuromorphic HRD-like coincidence detector. Finally, the inter-pulse delay detector and its subthreshold CMOS implementation on chip are detailed and evaluated in [Au3].

This thesis was also the subject of poster presentations at BioComp 2023 symposium [Au4] and at the 23rd JNM conference [Au5].

International Peer-Reviewed Journal Articles

[Au1] E. Dalmas, F. Danneville, F. Elbahhar, M. Bocquet, and C. Loyez, "A Review of Neuromorphic Sound Source Localization and Echolocation-Based Navigation Systems," Electronics, vol. 13, no. 24, p. 4858, 2024, doi: 10.3390/electronics13244858.

[Au2] E. Dalmas, F. Danneville, M. Bocquet, and C. Loyez, "Neuromorphic Coincidence Detector for Interaural Time Difference Encoding and Sound DOA Estimation," IEEE Trans. Instrum. Meas., vol. 73, pp. 1–11, 2024, doi: 10.1109/TIM.2024.3460950.

[Au3] E. Dalmas, C. Loyez, K. Carpentier, and F. Danneville, "Bioinspired recognition of cricket calling songs in sub-nanowatt inter-pulse delay detector", Bioinspiration & Biomimetic, vol. 20, no. 6, p. 066009, 2025, doi: 10.1088/1748-3190/ae0aa8.

National Conferences without Proceedings

[Au4] E. Dalmas, C. Loyez, and F. Danneville, "Spiking Neural Networks and Artificial Cochlea for Acoustic Pattern Recognition", [Poster], in Symp. BioComp 2023, France, Banyuls-sur-Mer, Dec. 2023.

[Au5] E. Dalmas, F. Danneville, K. Carpentier, and C. Loyez, "Dispositifs neuromorphiques faible consommation pour l'I.A. embarquée", [Poster], in 23rd Journées Nationales Microondes – JNM 2024, France, Antibes Juan-Les-Pins, June 2024.

Title: Bioinspired Ultra-Low Power Architectures for Sound Source Localization and Recognition

Keywords: spiking neurons, neuromorphic technology, ultra-low power, sound source localization, acoustic recognition

Abstract: Biomimetism is an increasingly widespread approach in various scientific fields. It regularly gives rise to new paradigms and, in recent years, has driven neuromorphic computing and technologies which promise significant advances in the field of information theory and unprecedented energy efficiency. With this approach, artificial spiking systems inspired by the neuronal impulse processing in the brain can process signals from various modalities. In the context of acoustic monitoring of biodiversity, this thesis investigates the potential of an analog neuromorphic technology integrating metal-oxide-semiconductor field-effect transistors operating in the subthreshold regime with ultra-low power (ULP) consumption. Keeping ULP constraints under consideration, bioinspired energy-efficient precomputing tools are designed for sound source localization and recognition and their performances assessed. Firstly, an original interaural time difference (ITD) extractor is modelled after the Hassenstein-Reichardt detector of motion detection, chosen for its low number of neurons, and applied to acoustic signals for estimation of sound sources' direction of arrival. The ITD extractor is evaluated in simulation on the basis of 2-D and 3-D indoor recordings at distances between 24 cm and 10 m of click-like sounds in particular. A simplified hyperbolic multilateration technique enables the analysis of the ITD extractor's localization performances, resulting in encouraging azimuth accuracies, in view of its potential ULP consumption, of 73.9% (±2.5°) and 77% (±5°) between 1 m and 3 m for click-like sounds. Then, with the aim to address multisource scenarios, a detector of temporal characteristics inspired by the calling song recognition mechanism of female field crickets is designed and successfully implemented on chip using the subthreshold neuromorphic technology. Tested under a probe station with artificial and real-world cricket calling songs, the detector reaches a sub-nanowatt total power consumption in quiet or noisy scenarios with high precision and encouraging recall. Finally, combining multiple instances of these two precomputing tools enables one to envision acoustic source tracking and counting applications.

Titre: Architectures Bioinspirées Ultra-Faible Consommation pour la Localisation et la Reconnaissance de Sources Sonores

Mots clés: neurones à spikes, technologie neuromorphique, ultra-faible consommation, localisation de sources sonores, reconnaissance acoustique

Résumé: Le biomimétisme est une approche de plus en plus répandue dans les différents domaines scientifiques. Il est régulièrement la source de nouveaux paradigmes et, depuis quelques années, a impulsé le traitement et les technologies neuromorphiques qui portent la promesse d'avancées significatives dans le domaine de la théorie de l'information et une efficacité énergétique sans précédent. Avec cette approche, les systèmes artificiels à impulsions inspirés du traitement neuronal dans le cerveau peuvent traiter des signaux issus de diverses modalités. Dans le contexte de la surveillance acoustique de la biodiversité, cette thèse explore le potentiel d'une technologie neuromorphique analogique intégrant des transistors à effet de champ métal-oxyde-semi-conducteur fonctionnant en régime sous-le-seuil avec une puissance consommée ultra-faible (ULP). En tenant compte des contraintes ULP, des outils de pré-traitement bioinspirés et économes en énergie sont conçus pour la localisation et la reconnaissance des sources sonores et leur performances sont évaluées. Tout d'abord, un extracteur original de différence interaurale de temps (ITD) est modélisé d'après le détecteur de mouvement Hassenstein-Reichardt, choisi pour son faible nombre de neurones, et appliqué aux signaux acoustiques pour estimer la direction d'arrivée des sources sonores. L'extracteur d'ITD est évalué en simulation sur la base d'enregistrements en intérieur en 2D et 3D à des distances comprises entre 24 cm et 10 m, en particulier pour des sons de type clic. Une technique simplifiée de multilatération hyperbolique permet d'analyser les performances de localisation de l'extracteur d'ITD, et qui résulte en des précisions azimutales encourageantes, compte tenu de sa consommation potentielle d'ULP, de 73,9 % (±2,5°) et 77 % (±5°) entre 1 m et 3 m pour les sons de type clic. Ensuite, dans le but de traiter des scénarios multisources, un détecteur de caractéristiques temporelles inspiré du mécanisme de reconnaissance du chant d'appel des criquets femelles a été conçu et intégré sur puce à partir de la technologie neuromorphique sous-le-seuil. Testé sur un banc sous pointes avec des signaux artificiels et des chants d'appel de criquets réels, le détecteur atteint une consommation totale inférieure au nanowatt dans des scénarios calmes ou bruyants, avec une grande précision et un recall encourageant. Finalement, la combinaison de plusieurs instances de ces deux outils de pré-traitement permet d'envisager des applications de suivi et de comptage des sources acoustiques.

IEMN (UMR 8520), Cité Scientifique, Avenue Henri Poincaré, CS 60069, 59652 Villeneuve d'Ascq, France