A Spiking Neural Network Architecture for Localizing Event-Trigger Indoor Moving Sound Sources

Zahra Roozbehi

2024

of Computing and Mathematical Sciences

A thesis submitted to

Auckland University of Technology

in fulfilment of the requirements for the degree of Doctorate of Philosophy

Abstract

Imagine being blindfolded in a room and hearing a voice fading and moving around. How do we track the sound's origin and distance, and how do we distinguish what is being said? What computational methods and techniques exist for addressing this problem?

Sound source localization refers to acoustic methods and technology for determining sound source in a three-dimensional space. However, existing methods struggle in real-world scenarios with background noise and multiple moving sound sources. In the domain of real-time applications, researchers continue to face challenges in sound source tracking and classification

The aim of this thesis is to introduce and evaluate a novel approach based on spiking neural networks to address the challenge of sound localization in dynamic environments. We present the adaptive resonance theory-based reservoir spiking neural network (ART-rSNN) and demonstrate its application in real-time, multi-source sound detection and classification.

Extensive simulations comparing our approach with other conventional machine learning models reveal that these models have some problems in categorizing and detecting sound events with multiple sources in real-time in comparison to our approach. The ART-rSNN can dynamically and autonomously adjust its neuron configuration based on received sound cues. This dynamic characteristic enables it to concentrate computation exclusively in the vicinity of estimated sound sources – a departure from static methods.

Overall, our framework handles the challenges of spatio-temporal data analysis required for this task while demonstrating adaptability in managing changing acoustic environments. What sets our work apart is its reliance on the measured power of sound without necessitating prior spatial sound source data for supervised learning. This distinctive feature improves the performance of our approach, especially in scenarios where other deep-learning approaches struggle to handle multiple sound sources using only time-domain raw signals.

In conclusion, the dynamic adaptability of our ART-rSNN, coupled with its performance in noisy environments and multi-source scenarios, positions it as a promising advancement in the field of AI-based approaches to sound localization and classification.

Contents

List of Figures

List of Tables

List of Appendices

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Zahra Roozbehi	03/02/2024
_____	_____
Signature	Date

Co-Authored Works

Chapter	Publication	Author(%)
Chapter 5	Roozbehi, Z., Narayanan, A., Mohaghegh, M., & Saeedinia, S.-A. (2024 Jan 31). Dynamic-Structured Reservoir Spiking Neural Network in Sound Localization. IEEE Access. doi: 10.1109/ACCESS.2024.3360491.	ZR: 80 AN: 10 MM: 5 SS: 5

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Ajit Narayanan, for his invaluable supervision, support, and tutelage during the course of my PhD degree. His guidance and feedback have been instrumental in shaping my research and academic career. I would also like to thank my second supervisor, Dr.Mahsa Mohaghegh, for her dedicated support and guidance throughout my research project. Her encouragement and enthusiasm have been a constant source of motivation for me.

Finally, I am deeply grateful to both my partner and my broader circle of family and friends, for their combined unwavering support and encouragement throughout my academic journey. Their love and faith have been my anchor during the ups and downs of my PhD.

Ethics Approval

The research conducted as part of this thesis adheres to the ethical standards and guidelines set forth by Auckland University of Technology. While the use of open-source acoustic signal data, which is not related to human subjects, may not fall under the purview of human research ethics, it is essential to ensure that the utilization of such data complies with the terms of use and licensing associated with the specific dataset or source from which the acoustic signal data was obtained.

The ethical considerations encompassed in this research include the respect for rights of privacy and confidentiality, the avoidance of conflict of interest, and the commitment to the principles of the Treaty of Waitangi.

The utilization of open-source acoustic signal data is in line with the university's guidelines and procedures, and it is imperative to demonstrate a responsible approach to the use of such data, ensuring that it aligns with the overarching ethical principles of the academic institution.

1. Introduction

1.1 Background and Scope

Sound source detection, localization, and classification (SSDLC) are essential tasks in signal processing that involve identifying, locating, and distinguishing sound signals originating from various sources. Commonly, microphones or acoustic sensors are employed for sound detection, while algorithms such as spectrogram analysis, time-frequency analysis, and neural network classifiers are utilized for sound classification.

The process of sound source detection begins with devices like microphones or acoustic sensors capturing the acoustic energy and converting it into electrical signals. These electrical signals are then processed using various algorithms to extract essential features such as frequency content, duration, and intensity, which are crucial for sound classification (Nunes et al., 2021).

Sound source localization, the next step in SSDLC, relies on the use of multiple sensors placed strategically in the environment. Triangulation, a technique that measures the time delays between the arrival of the sound at different sensors, is employed to calculate the direction and distance of the sound source (Wang & Soh, 2017). However, factors such as reverberation and noise can make it difficult to accurately determine the location of sound sources.

Various techniques can be employed for sound source classification. Spectrogram analysis, a popular method, offers a visual representation of a sound signal's frequency content over time, enabling the differentiation between distinct types of sounds. Time-frequency analysis, another method, presents a more detailed depiction of a sound signal's time-varying frequency content. Additionally, neural network classifiers leverage machine learning to identify patterns and differentiate between various types of sounds (Karanasiou & Brown, 2021).

Despite considerable advancements, SSDLC still faces numerous challenges, including reverberation, noise, variability of sound sources, complexity of the acoustic

environment, and the requirement for real-time processing in many applications (Zhang et al., 2021).

Emerging SSDLC techniques, such as spiking neural networks (SNNs), show promise in addressing these challenges. SNNs emulate the behaviour of biological neurons and have demonstrated effectiveness in SSDLC tasks. With their ability to handle noisy and variable signals, as well as being implementable in real time on low-power hardware, SNNs are ideally suited for applications in environmental monitoring and robotics (Liu et al., 2009).

The objective of this PhD thesis is to develop a novel approach for SSDLC using a SNN. The proposed method aims to harness the power of SNNs to accurately detect, locate, and classify sound sources in noisy and complex environments. Furthermore, this approach addresses the need for real-time processing, making it suitable for applications such as robotics and environmental monitoring.

In summary, SSDLC is a vital task with numerous applications across various fields, including security systems (Venkatraman et al., 2021), teleconferencing (Zhuo & Cao, 2021), robotics (Rascon & Meza, 2017), and intelligent home services (De Silva et al., 2017). Despite its significance, SSDLC faces several challenges, such as noise, reverberation, and variability of sound sources. The approach proposed in this study, leveraging the capabilities of SNNs, holds the potential to overcome these challenges and deliver accurate and reliable sound source localization and classification in real time.

1.2 Challenges

1.2.1 Theoretical Challenges

In the realm of audio analysis, sound event classification and localization present challenges that have captured the interest of researchers in various domains, such as audio recording, transmission, and intelligent audio processing. Sound events, which encompass a diverse range of audible experiences in daily life – from the rustling of leaves in the wind to the hum of a refrigerator or the laughter of a child – play an essential role in understanding and differentiating sound sources. Localization, which involves pinpointing the exact location of a sound source, is a critical component of audio analysis, fostering innovation and progress in the field.

Sound source localization (SSL) algorithms play a crucial role in analysing the rich tapestry of sounds that fill our environments, from home automation to surveillance systems. Picture a bustling household, with a multitude of sounds – appliances humming, air conditioners whirring, and lively conversations echoing. It is in such complex scenarios that microphones or acoustic sensors strive to accurately capture and analyse these sounds for vital applications, such as monitoring and security. However, the challenge of background noise, combined with other factors like reverberation, broadband speech signals, and the presence of multiple or intermittent and moving sources, hinders SSL algorithms from precisely locating and classifying sound sources (Chen et al., 2017).

Reverberation arises from sound wave reflections off surfaces, corrupting location predictions (Peng et al., 2014). Broadband speech signals present another challenge, as they restrict the application of narrowband algorithms and add complexity (Alam et al., 2017). Intermittency and movement stem from the fact that sound sources are not inherently stationary, posing issues for filters designed for stationary signals within SSL algorithms (Yan et al., 2020). Lastly, the presence of multiple sound sources can cause conventional SSL algorithms to fail (Hammer, 2021).

SNNs use interaural time differences (ITDs) and interaural level differences (ILDs) as binaural cues for sound localization (Francl et al., 2022). Enhancing the efficiency of sound localization can be achieved by fusing other cues with ITDs and ILDs (Francart et al., 2011). For example, Simon et al. (2007) present an algorithm that enhances interaural level differences to improve sound localization in bimodal hearing. Binaural auditory processing is the ability of the binaural system to make use of the interaural difference cues (IID and ITD) in the received sounds (Francart et al., 2011). Therefore, while SNNs can use ITDs and ILDs as binaural cues for sound localization, fusing other cues with them can enhance efficiency.

In addition, SNNs can be used for adaptive sound localization and classification because they are biologically inspired and can simulate the sound localization ability of mammalian auditory pathways. SNNs can process temporal information and perform computations in real time, making them suitable for sound event analysis. SNNs can also adapt their topology to the locations of sound events, which helps to reduce the

computational cost of sound event analysis and improve the efficiency of the process (Chan et al., 2010; He et al., 2019).

Despite advances in time domain sound localization, SNNs continue to encounter challenges in accurately localizing sound, particularly with respect to the azimuth sign issue. This issue arises when a network is trained on a specific set of sounds, and subsequently tested on an entirely different set. Consequently, the network may come to rely on the sign of the azimuth angle to determine the location of a sound source, leading to potentially erroneous localization if the sign is flipped. Although some studies have proposed solutions, including the use of azimuth-frequency representation and CNNs (Chun et al., 2020), SNNs still face significant obstacles in sound localization. Recent studies have shown that deep neural network models can be trained to map binaural audio to the location of the sound source, specified by its azimuth and elevation relative to the model's location bins (Francl et al., 2022). Another study used machine learning with SNNs for binaural sound source localization (Al-Abboodi, 2019).

1.2.2 Practical Challenges

Practical challenges in SSL commence with hardware limitations. Capturing and processing audio are demanding tasks that necessitate a balance between the complexity of audio processing and available hardware resources (Belloch, 2015). Audio monitoring also imposes specific hardware platform requirements, such as large memory capacities and high computational capabilities (Denk, 2019; Fabregat, 2020).

Measurement errors, the second practical challenge, result from environmental uncertainties, noise, and reverberations. Sound waves are susceptible to signal diffraction, echoes, reflection, deflection, and diffractions, leading to measurement errors (Shaukat, 2021). Node synchronization issues can also cause errors in methods like time of arrival (TOA) and time difference of arrival (TDOA), yielding inaccurate location predictions. It is crucial to select a method that minimizes measurement errors and demonstrates resistance to noise (Cobos, 2017).

Power dissipation, a third challenge, can lead to power loss in battery-powered microphone hardware, necessitating power consumption optimization (Liaquat et al., 2021).

Deployment issues, a fourth challenge, involve specific hardware requirements for each SSL method. For instance, TOA necessitates node synchronization, while energy-based methods require calibrated gains. Some methods mandate physical administration, and their calibration can be time consuming (Shaukat, 2021).

The fifth challenge, scalability, entails adjusting both hardware and software based on the size of the SSL implementation environment. SSL algorithms should be scalable, and their parameters must be tuned according to the application (Liaquat et al., 2021).

The sixth challenge is implementing adaptive SNN for sound event detection and localization, which involves several issues.

One of the main challenges is the lack of labelled data for training the SNNs. The process of labelling the data is time consuming and requires domain expertise. Another challenge is the computational cost of training the SNNs. The adaptive topology network approach helps to reduce the computational cost, but it requires careful tuning of the parameters to achieve optimal performance. Additionally, the performance of the SNNs is affected by the quality of the input data, such as the signal-to-noise ratio and the reverberation of the sound sources. The SNNs' performance is also affected by the complexity of the sound events, such as the number of sound sources and their locations. Finally, the SNNs' performance is affected by the choice of the optimization algorithm and the learning rate. Addressing these challenges requires careful consideration of the SNNs' architecture, the training process, and the evaluation metrics (Pan et al., 2021).

1.3 Motivation

The motivation for this PhD thesis stems from the potential of spiking neural networks (SNNs) to overcome challenges in sound event classification and localization (SECL), a crucial field with applications in areas such as surveillance systems, smart homes, and human-machine interaction. In this study, a "sound event" is defined as the moment a sound signal is detected by the system’s sensors, triggering real-time localization and classification processes. This sensor-based approach allows the system to dynamically adapt as signals are received, facilitating rapid response and accurate detection.

A key distinction within the study is between Sound Source Localization (SSL) and Sound Event Classification and Localization (SECL). SSL is focused solely on identifying the spatial origin of a sound source, whereas SECL extends this capability by also categorizing the type of sound event, thus addressing both "what" and "where" a sound is occurring.

Despite the importance of SECL, conventional deep learning models used for this purpose often face limitations in terms of accuracy, computational efficiency, and their ability to operate in real time on edge devices. SNNs have emerged as a promising alternative, offering the following key advantages:

- Superior accuracy: SNNs have demonstrated better performance compared to conventional deep learning models in SEL tasks, enabling more precise sound event detection and localization (Wu et al., 2020).
- Low computational cost: Due to their event-driven nature, SNNs require less computational power (Yamazaki et al., 2022), making them highly suitable for real-time SEL tasks, even on resource-constrained edge devices.
- Temporal context capture: SNNs can model time-based event sequences in acoustic data (Pan et al., 2020). This unique ability enables them to perceive patterns that are typically unattainable through conventional models (Bing et al., 2019). The significance of this capability lies in its contribution towards the precise tracking of mobile sound sources.
- Energy efficiency: The event-triggered SNN approach holds promise for reducing energy consumption in SEL applications (Yu et al., 2023), an important consideration for battery-powered devices.

Reservoir structures have emerged as a specific area of interest in SNNs as they enable efficient processing of spatio-temporal data and exhibit competitive classification performance (Schliebs et al., 2011). This gives them significant potential in the efficient tracking of spatially moving sound sources in dynamic scenes. Reservoir computing in SNNs (rSNNs) offers fast learning capabilities and low training costs (Reynolds, 2019; Cucchi et al., 2022; Ferreira et al., 2013).

Furthermore, reservoir structures without adaptive updating can be more hardware-friendly, allowing for better implementation in real-world applications (Tanaka et al., 2019).

However, concerns remain regarding the efficiency of network architecture and coding schemes used for audio signals in event-triggered SNNs (Pan et al., 2019; Valenti et al., 2017). To maximize efficiency, the network size and structure must be optimized to minimize computation time and increase precision. Our research introduces a novel recurrent spiking neural network (RSNN) with a dynamic framework for both triggering and categorizing sound events. RSNNs have been employed as reservoirs for liquid state machines (LSMs) in spatio-temporal pattern recognition tasks (Schliebs et al., 2011), which motivates us to use an RSNN with a malleable structure. Having a dynamic structure can enhance network architecture and coding schemes, leading to the advancement of energy efficient and precise sound event localization systems while taking advantage of benefits offered by RSNNs.

In summary, the motivation for this research lies in harnessing the potential of SNNs, specifically event-triggered SNNs with reservoir structures, to address challenges in sound event classification and localization. The structure presented in our research includes changes in the topology and size of the network in both temporal and spatial alignments, and these changes are updated according to the amount of energy estimation error. This provides an architecture with adaptive dynamics that can enhance network architecture. By developing our dynamic SNN architecture, this research aims to improve the efficiency, accuracy, and energy consumption of SEL systems, ultimately benefiting applications in surveillance, smart homes, and human-machine interaction.

1.4 Research Questions

The primary goal of this research is to uncover a functional architecture capable of detecting, localizing, and classifying sound sources in noisy environments. This ambitious endeavour seeks to tackle the complex challenges associated with energy estimation error propagation in algorithms and to develop a comprehensive structure that efficiently combines detection, localization, and classification algorithms. Inspired by the intricate workings of the brain, where neurons convey diverse information through the timing of spikes produced by neuronal populations (Averbeck, 2006), this research aims to design an efficient structure that can balance accuracy and computational cost, which are dependent on its size and connection density (Yudanov, 2010).

The overarching research question in this study is "What is a workable architecture for detecting, localizing, and classifying multiple sound sources in noisy environments?". To systematically address this question, we have organized our inquiry into four main themes, each with its own set of sub-questions:

1. Sound Localization
    - How can the accuracy of sound localization be improved by synergizing energy-based cues with traditional interaural time difference (ITD) and interaural intensity difference (IID) methods in spiking neural network (SNN) structures?
    - What factors influence the determination of the azimuth sign when locating a sound?
    - How should neurons be arranged and connected in an rSNN structure to effectively process spatio-temporal data in incoming signals?

2. Sound Event Triggering
    - What methods can be used to encode audio data into electrical impulses suitable for rSNNs in noisy environments?
    - What threshold should be considered to initiate neuronal activity in rSNNs?

3. Sound Event Classification
    - What is the most effective approach for classifying sound event patterns in the developed network structure?

4. System Evaluation
    - How does the dynamic spatio-temporal rSNN architecture perform in sound localization compare to conventional approaches?
    - How do the adaptive size increase of the neural network and dynamically assigned neuron positions impact the learning process and overall performance of the model?

By addressing these research questions, this PhD research aims to develop a new RSNN architecture optimized for energy efficiency, computational speed, and accuracy, and to compare its performance with existing state-of-the-art techniques to determine potential applications and limitations. The approach benefits from a new RSNN architecture that is new in its topology and learning law and is able to be applied in

spatio-temporal applications. The novelty of the research and its potential contribution to the field of SNNs and spatio-temporal applications make it a suitable topic for a PhD thesis.

1.5 Research Methodology

The research methodology employed in this study can be summarized in six key steps, as illustrated in Figure 1. These steps are designed to systematically explore the development and evaluation of a functional architecture for detecting, localizing, and classifying sound sources in noisy environments using spiking neural networks (SNNs).

1.5.1 Literature Review

As part of the research methodology, an extensive review of the literature will be carried out to explore the research findings of techniques for sound source localization, detection, and classification. This includes the examination of various neural network approaches such as adaptive resonance theory (ART).

The main objective of the literature review is to identify the cutting-edge methods of sound source localization, and how they tackle the challenges associated with it. The review will encompass the scrutiny and analysis of relevant studies, articles, and publications from academic databases such as ScienceDirect, CORE, and PubMed.

The primary focus of the review will be on the research findings of the studies, rather than the implementation and comparison of the techniques. The topics covered in the review will include the performance of azimuthal sound source localization under different conditions, binaural sound source localization using machine learning with spiking neural networks features extraction, sound source localization in 2D space, and deep neural network models of sound localization.

The literature review will provide a comprehensive understanding of the current state of the art sound source localization techniques and their limitations, which will inform the development of the RSNN architecture.

1.5.2 Propose a SNN Structure

The research methodology proposes a SNN architecture for detecting, localizing, and classifying sound sources in noisy environments. This architecture will address the

challenges posed by conventional techniques and considers the potential benefits of incorporating ART principles.

The proposed SNN architecture will be developed based on the findings of the literature review and will be optimized for energy efficiency, computational speed, and accuracy. The architecture will be designed to simulate the sound localization ability of the mammalian auditory pathways using the IID and ITD methods, integrated with an energy decaying model. The architecture will be developed using Python programming languages.

1.5.3 Evaluation of ART-rSNN Architecture

The performance of the SNN architecture developed in step 3 will be evaluated and compared to existing techniques, including traditional SNN models, ART-based models, and other state-of-the-art methods. The evaluation will involve measuring the accuracy, precision, recall, and F1 score of the adopted architecture and comparing these measurements to the performance metrics of existing techniques. The evaluation will be conducted using a dataset of sound sources in noisy environments.

1.5.4 Comparison with RSNN and Deep Learning Methods

The results of the evaluation will be analysed to identify the strengths and weaknesses of the SNN architecture and compare it to other state-of-the-art RSNNs and deep learning methods. The analysis will involve identifying the factors that contribute to the performance of the architecture and comparing them to the factors that contribute to the performance of the existing techniques.

1.5.5 Feature Research Directions

Based on the findings of the analysis, future research directions will be suggested for further improvement of the model, including the exploration of ART approaches. The future research directions will be based on the identified strengths and weaknesses of the architecture and the limitations of the existing techniques. The suggested future research directions will aim to improve the performance of the architecture and address the limitations of the existing techniques.

By following these methodological steps, the main contribution of the research is the design of a novel adaptive topology network that adds the necessary number of neurons

into the network according to the locations of the sound events. This approach helps to reduce the computational cost of sound event analysis and improve the efficiency of the process.

The research methodology includes a comprehensive literature review of sound source localization, detection, and classification techniques; the proposal of an SNN architecture; its development and implementation; the model's optimization and evaluation; and analysis of the results. The architecture's adaptive topology network aims to address the limitations of the existing techniques and improve the performance of sound event analysis in noisy environments by reducing the computational cost.

4. Research Contributions

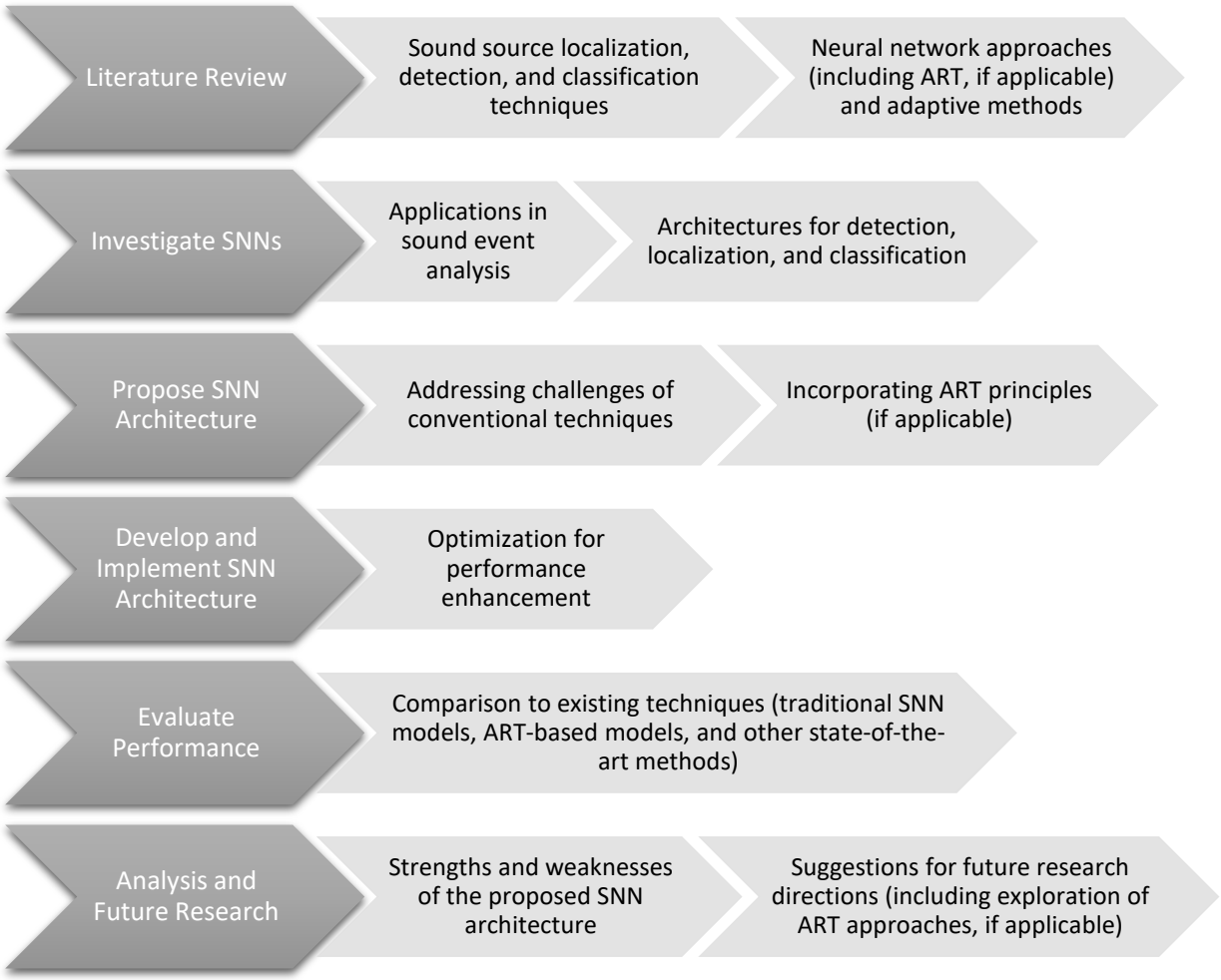
The primary goal of this research is to advance the field of sound event localization, detection, and classification framework. To this aim, an innovative approach that can adapt to various challenges posed by real-world sound environments has been developed. This approach utilizes adjusting network topology in each training iterations and online process. One can envision a system that seamlessly adjusts its neural network structure in response to changing noise levels, optimizing performance and computational costs in real time.

This research aims to achieve this vision through the following main contributions:

- Propose a novel recurrent SNN architecture that can dynamically adjust its network size and neuron arrangements to optimize performance and computational costs in real time. Unlike existing adaptive SNNs, this architecture integrates sound classifications and localizations and uses online topology adaptations. The architecture is flexible and adaptive, allowing for adjustments to changing sound environments.
- The integration of energy-based cues with traditional interaural time difference (ITD) and interaural intensity difference (IID) methods for sound localization. This combination enhances the accuracy and robustness of the sound localization process.

Figure 1

The Main Steps of the Research Methodology



- An innovative approach to sound localization that estimates the coordinates of the sound source, rather than only the azimuth angle. This method provides more precise and reliable localization results and has potential applications in robot audition and hearing aid devices.
- The ability to accurately determine the positive/negative signs of the azimuth angle, further enhancing the sound localization results.
- A pattern recognition and classification system integrated into the recurrent SNN structure. This system can recognize and classify different sound events based on their spatio-temporal patterns, with potential applications in speech recognition, environmental sound monitoring, and sound-based human-computer interaction.

- In addition to the use of SNNs in SSDLC, the PhD thesis also contributes to the field by using supervised learning and spike timing-dependent plasticity (STDP) to reinforce learning. Specifically, the thesis proposes a novel approach for sound event detection and classification using spiking neural networks (SNNs) that combines supervised learning with STDP to improve the accuracy of the network. The supervised learning is used to train the network on measured signals, while STDP is used to reinforce the connections between neurons in the network, based on the timing of their spikes.

These research contributions represent significant advancements in the field of sound event localization, detection, and classification, paving the way for more efficient and accurate systems in real-world applications.

5. Thesis Outline

The above sections introduced the research problems and objectives of this thesis. The remainder of this thesis is structured as follows:

- Chapter 2: Literature Review - This chapter explores the current state of research on sound source localization and spiking neural networks. It also delves into adaptive resonance theory and the limitations of existing approaches to sound source localization.
- Chapter 3: Theoretical Background - This chapter details the research objectives and the theoretical foundations of the dynamic structured recurrent spiking neural network. It also discusses the integration of energy-based cues with ITD and IID methods for sound source localization.
- Chapter 4: Recurrent Neural Network (RNN) for Sound Source Localization and Classification - This chapter provides an overview of recurrent neural networks (RNN) and explores their key components. Additionally, recurrent spiking networks are introduced and their traditional structures are discussed in brief detail.
- Chapter 5: Dynamic Structured Recurrent Spiking Neural Network Design, Proposed Methodology and Model Design - This chapter introduces the novel dynamic structured recurrent spiking neural network and its learning

strategies. It also explains how the architecture is designed to address the limitations of existing approaches to sound source localization.

- Chapter 6: Results and Findings - This chapter presents the results of the experiments conducted with the proposed network, comparing its performance with conventional and intelligent methods. It also discusses the effects of the dynamic architecture and the spatial data of incoming signals.
- Chapter 7: Conclusion and Future Work - This chapter summarizes the research findings and contributions, highlighting the improvements in sound source localization achieved by the proposed dynamic structured recurrent spiking neural network. It also discusses limitations and suggests areas for future research to enhance the performance of spiking neural networks in sound source localization.

2. Literature Review

2.1 Introduction

This chapter provides a detailed overview of the literature review conducted in this study, specifically focusing on recent advancements in sound processing algorithms related to sound source detection, classification, tracking, and triggering. To ensure the highest level of rigor and a systematic approach in gathering relevant research, the study followed a structured methodology. An explicit set of inclusion and exclusion criteria were defined to gauge the relevance of identified studies. These criteria served as a filter to ensure that only studies meeting specific prerequisites were included, while irrelevant or unrelated research was systematically excluded.

2.1.1 Literature Search Methodology

To pinpoint relevant studies, the research employed a methodology that incorporated the following key elements:

2.1.2 Keywords and Terminology

The study selected a set of keywords and terminology closely aligned with sound processing, encompassing sound event detection, sound source classification, sound source localization, sound localization methodologies, sound source localization and classification techniques, real-time sound detection, deep learning methods in sound source classification, spiking neural networks for sound source localization and classification, multiple sound source localization and classification, distinctive features in sound localization and classification, mel-frequency and temporal sound features, cues for sound source localization, energy-based models, interaural time differences (ITD) and interaural level differences (IID) in sound localization, optimization in spiking neural networks, and other relevant terms.

2.1.3 Inclusion and Exclusion Criteria

A set of inclusion and exclusion criteria was defined to gauge the relevance of identified studies. These criteria served as a filter to ensure that only studies meeting specific

prerequisites were included, while irrelevant or unrelated research studies were systematically excluded.

2.1.4 Search Engines

The search strategy was executed through the utilization of multiple search engines such as Google, Google Scholar, and Bing. These search engines were selected for their extensive coverage of academic and technical literature.

2.1.5 Bibliographic Databases

In addition to search engines, the research probed bibliographic databases to gain access to peer-reviewed publications, theses, and scholarly articles. The inclusion of these databases provides comprehensiveness in the literature review.

2.1.6 Citation Tracking

To augment the richness of the literature review, the research employed citation tracking as a valuable technique. This involved examining the references cited within relevant papers to identify additional sources that could provide valuable insights into the methods, advantages, and limitations of the studies.

2.1.7 Adherence to Systematic Methods

To achieve a systematic literature review, the methodology adopted principles akin to those recommended by PRISMA (preferred reporting items for systematic reviews and meta-analyses). While this study may not strictly adhere to the complete PRISMA guidelines, it utilized structured and systematic approaches to ensure the robustness and reliability of the literature review.

By adhering to this structured methodology, this study offers a comprehensive, transparent, and informative overview of recent developments in sound processing algorithms and their practical applications in sound source detection, localization and classification. This approach ensures that the review serves as a valuable resource, providing insight to underpin the research presented in this thesis.

2.2 Review of Recent Sound Source Localization Approaches

Sound source localization is the process of determining the location or origin of a detected sound in terms of direction (Desai & Mehendale, 2021), distance (Rhinehart et al., 2020), and, sometimes, velocity. According to Desai and Mehendale (2021), SSL is the problem of estimating the position of one or several sound sources relative to some arbitrary reference position, which is generally the position of the recording microphone array, based on the recorded multichannel acoustic signals. SSL is a fundamental problem in both science and engineering, with applications in robotics, signal processing, and various other fields. The mechanisms of SSL have been extensively studied, and the auditory system uses several cues for sound source localization, including time difference and level difference (or intensity difference) between the ears, and spectral information.

The human auditory system and robotics have been extensively studied in the context of sound source localization (Rascon & Meza, 2017; Desai & Mehendale, 2022; Pan et al., 2021; Pang et al., 2021). This chapter aims to provide a comprehensive review of the current state-of-the-art techniques and research in sound source localization. This section provides a review of recent sound source localization approaches, focusing on both deep learning and non-deep learning methods.

2.2.1 Deep Learning Approaches

Deep learning methods have garnered considerable attention in the field of sound source localization due to their ability to automatically learn intricate features from raw audio data. Numerous studies have delved into the utilization of deep neural networks for sound source localization. One notable survey conducted by Grumiaux et al. (2021) provided a comprehensive overview of these methods for both single and multiple sound source localization. This survey explored various techniques, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), shedding light on their performance in diverse scenarios.

In a study by Francl et al. (2022), published in *Nature Human Behavior*, deep neural network models were developed to adapt sound localization to real-world environments. These trained networks exhibited exceptional accuracy in localizing sound sources, even in the presence of disruptive noise and reverberation. This study

also investigated frequency-dependent cues such as interaural time differences (ITDs) and interaural level differences (ILDs), as well as the integration of spatial information across different frequencies.

Deep learning techniques have proven their effectiveness in automatically extracting intricate spatial details from audio signals (Alzubaidi et al., 2021). These approaches have exhibited promising outcomes in accurately pinpointing sound origins, even in challenging settings characterized by noise and reverberation. Commonly employed methods for extracting spatial information from audio signals include convolutional neural networks (CNNs) and recurrent neural networks (RNNs) (Grumiaux et al., 2022).

Furthermore, deep learning methods possess the capability to adapt to real-world conditions and effectively utilize frequency-dependent cues such as ITDs and ILDs (Bianco et al., 2019). Although these principles may not be immediately apparent in the field of machine learning, they have been substantiated by various research studies. For instance, research by Bianco et al. (2019) underscores the utilization of machine learning in acoustics, including source localization in speech processing and ocean acoustics. This study found that with ample training data, machine learning can uncover models that describe intricate acoustic phenomena such as human speech and reverberation. Another research article by Chen et al. (2021) leverages machine learning techniques to develop an innovative solution for verifying sound source localization information using a single microphone.

Additionally, research has shown that immune-based machine learning algorithms can play a role in improving accuracy and reliability, particularly in audio-visual applications (Ngo et al.,2019). For sound classification, cutting-edge techniques leverage SNN encoding and spike pattern generation. Some methodologies integrate the echo state SNN capability with CNN classification methods, resulting in significantly improved accuracy (Guo et al.,2023). Moreover, Convolutional Recurrent Neural Network (CRNN) methods, which integrate Gammatone filtering and frequency-based approaches, have demonstrated promising outcomes in sound source localization (Yiwere et al.,2019). These diverse methodologies represent the ongoing evolution of SSL techniques, highlighting their potential to deliver favorable results in various applications (Ngo et al.,2019; Guo et al.,2023; Yiwere et al.,2019).

In summary, deep learning approaches exhibit significant potential in the domains of speech processing and sound source localization. Their capacity to adapt to real-world conditions and utilize frequency-dependent cues positions them as a promising strategy for addressing real-world audio processing challenges. However, it is important to note that deep learning methods necessitate a substantial amount of labelled training data and significant computational resources for both training and inference (Mehrish et al., 2023).

2.2.2 Non-Deep Learning Approaches:

Although deep learning methods have shown promising results, non-deep learning approaches continue to play a significant role in sound source localization. Traditional methods rely heavily on signal processing techniques and feature extraction algorithms. These approaches encompass beamforming, time difference of arrival (TDOA) estimation, and sound intensity-based methods (Mehrish et al., 2023; Vera-Diaz et al., 2018).

A comprehensive survey conducted by Grumiaux et al. (2021) delves into non-deep learning methods for sound source localization, with a specific focus on indoor environments. The survey explored various algorithms and evaluated their accuracy and computational complexity. Additionally, it shed light on the challenges faced by non-deep learning methods, such as their susceptibility to noise and reverberation.

Non-deep learning approaches rely extensively on conventional signal processing techniques and feature extraction algorithms (Vera-Diaz et al., 2018). These traditional methods, which encompass techniques such as beamforming, TDOA estimation, and sound intensity-based methods, are frequently characterized by computational efficiency and the ability to yield precise sound source localization outcomes under particular conditions. Research studies in the field of acoustics have demonstrated the efficacy of these traditional signal processing methods (Desai & Mehendale, 2022).

However, these traditional methods may encounter challenges when confronted with noise and reverberation, and their effectiveness can be influenced by the complexity of the acoustic environment (Mehrish et al., 2021; Vera-Diaz et al., 2018).

A survey of deep learning methods for single and multiple sound source localization, with a focus on SSL in indoor environments, highlights that traditional methods perform poorly in difficult yet common scenarios where noise, reverberation, and several simultaneously emitting sound sources may be present (Mehrish et al., 2023). Although deep learning methods have shown great potential in this field, especially in handling noise and reverberation, traditional signal processing techniques and feature extraction algorithms play an important role in sound source localization, and their performance can be influenced by the complexity of the acoustic environment (Mehrish et al., 2021; Vera-Diaz et al., 2018).

2.2.3 Hybrid Approaches:

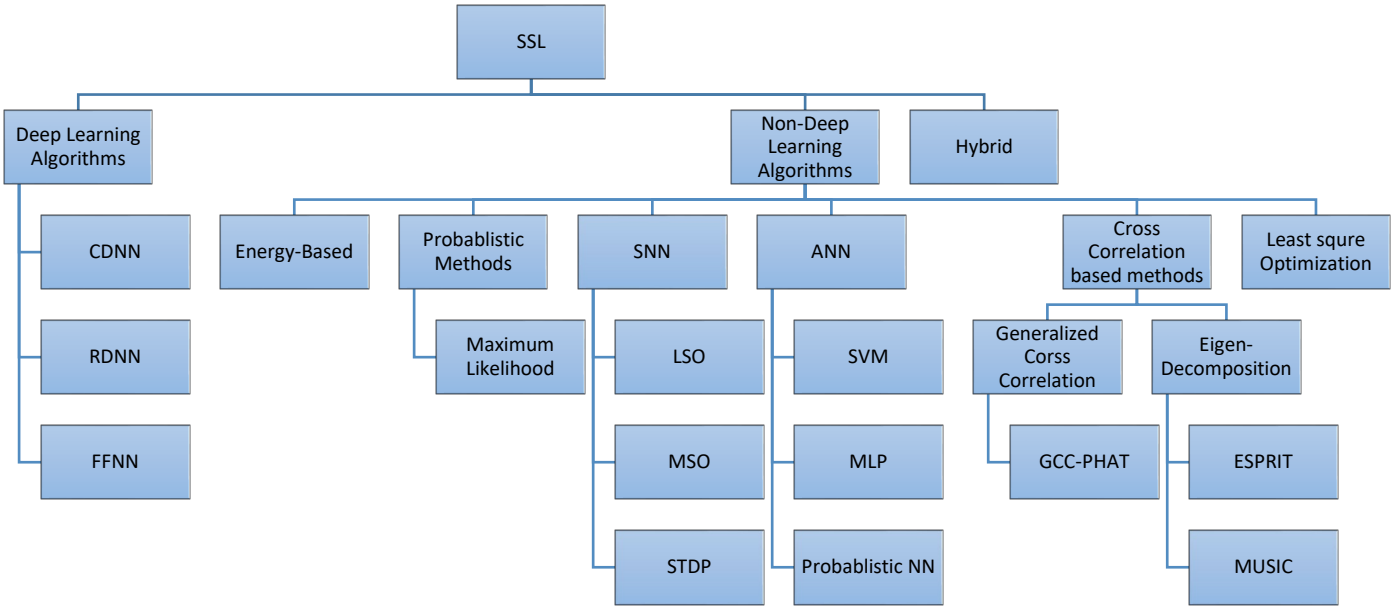
In recent years, researchers have been exploring hybrid approaches that combine deep learning and non-deep learning techniques to enhance accuracy in sound source localization. These approaches aim to capitalize on the different perspectives of both methodologies. For instance, a study by Kita et al. (2021) proposes a method that integrates a deep neural network with computer-aided engineering to estimate the position of sound sources within structures. This approach effectively utilizes accelerometer measurements and spectral analysis to achieve highly accurate localization.

In summary, sound source localization has been extensively studied using both deep learning and non-deep learning approaches. Hybrid approaches, which combine deep learning and non-deep learning methods, have shown great potential in improving sound source localization accuracy, particularly in complex scenarios (Escobar et al., 2013; Zhao et al., 2013). Hybrid approaches harness the automatic feature learning capabilities of deep learning with the robustness and efficiency of non-deep learning techniques (Escobar et al., 2013; Zhao et al., 2013). However, the success of hybrid methods often depends on the careful integration and optimization of the deep learning and non-deep learning components (Xiong et al., 2022, Zhao et al., 2013). The exploration of hybrid approaches has opened up new avenues for enhancing accuracy in sound source localization (Escobar et al., 2013; Zhao et al., 2013; Pérez-López et al., 2019). Generally, sound source localization can be categorized into deep learning, non-

deep learning, and hybrid methods. Figure 2 indicates a general classification of sound source localization algorithms.

Figure 2

SSL Method Classifications



Although there are a great number of successful studies in this area, there are some challenges and gaps. Table 1 describes some of these methods and challenges

Table 1A:

Review of Several Sound Source Localization Studies, A) MUSIC algorithms (non-deep, DOA)

Method SSL NN+SSL	Sub-method and cues	Key Approach	Relevant Studies	Challenges
Non-Deep	MUSIC algorithms DOA	Eigen value decomposition of covariance matrix Using SNR factor instead of maximum eigen value	(Ishi, 2009) (Gao, 2018)	• High computational costs due to singular value decomposition calculation. • Depends strongly on the parameters of the algorithm such as source signal frequency. • Performances are affected by noise properties, the number of microphones, and their configurations. • Accuracy is low in the case of moving sound sources.

Table 1B:

Review of Several Sound Source Localization Studies, B) GCC (non-deep, TDOA)

Method SSL NN+SSL	Sub-method and cues	Key Approach	Relevant Studies	Challenges	Method SSL NN+SSL
Non-Deep	GCC	TDOA	GCC-PHAT	(Lee, 2020) (Takahashi, 2021)	• Depends on microphone array and their distances, only DOA. • High computational costs • Cross-correlation function is more affected by low-frequency components. • GCC-PHAT may provide the TDE robust against reverberation, but it is known to be sensitive to ambient noise as the normalization emphasizes frequency components with small powers.
	Energy [5]	Energy	Solving and modelling optimization problem	(Yan, 2018) (Tang, 2022)	• The existing sound energy attenuation model is not accurate. • The ideal point (an omnidirectional sound source which does not exist in a real environment) will have an impact on decaying energy model. • The distances between the sensors and the sound source will impact accuracy. • Optimization problem should be feasible depending on optimization tools. • The prior information about the number of sound sources, is important.
	Probabilistic Method	DOA	Using maximum likelihood to associate data to the sources locations for multiple sources.	(Dang, 2019) (Masuyama, 202)	• High time costs • Measurement model is not highly accurate.

Table 1C:

Review of Several Sound Source Localization Studies, C) Deep approaches

	Method SSL NN+SSL	**Sub-method and cues**	**Key Approach**	**Relevant Studies**	**Challenges**
Deep	ANN	TDOA	MLP-GCC	(Vesperini, 2016)	• The model performance is strongly correlated to the training data size.
	SNN	ITD	STDP - ITD	(Glackin, Brendan and Wall, Julie A and McGinnity, Thomas M and Maguire, Liam P and McDaid, Liam J, 2010)	• Only calculates the horizontal angle. • Single source localization • Narrow band Frequency
	RDNN	TDOA	MUSIC algorithm based	(Adavanne, 2018)	• Single source localization requires high memory.
	CDNN	TDOA	Temporal modelling of acoustic signal transferring	(ADAVANNE, POLITIS, & VIRTANEN, 2019)	• Sensitive to reverberation • Only works for single source audio and not multisource.
	Probabilistic NN	DOA	Robust GCC-based	(Sun, 2017)	• Depends on microphone array and their distances, only DOA. • High computational costs • Cross-correlation function is more affected by low-frequency components.

Table 1 offers a comprehensive overview of both deep and non-deep sound source localization (SSL) methods, highlighting their distinctive features and associated challenges. By examining a range of approaches and their limitations, this table identifies the common problems encountered in SSL. While the table primarily presents a list of key features and challenges associated with each method, it also serves as a foundation for identifying shared issues and obstacles in SSL. In the subsequent analysis, we will delve deeper into the common challenges that cut across these methods, thus providing a more cohesive understanding of the overarching issues faced in SSL research.

These methods are also categorized in binaural and monaural methods. Binaural methods utilize at least two sensors to localize the sound source. In these methods, the received signals are compared in a manner that is similar to mammalian ears.

Having studied the auditory perception of space literature, Escudero et al., (2018) indicate that the mammalian auditory systems utilize two major localization cues – interaural time difference (ITD) and interaural intensity difference (IID), which as a whole constitute the duplex theory of sound localization (Goodman, 2009; Escudero et al., 2018).

Corresponding to the “duplex theory of sound localization”, some researchers have integrated both cues together. In this regard, high-frequency sounds are localized in the lateral superior olive (LSO) by calculating IID (Tollin, 2003), while low-frequency sounds are localized in the medial superior olive (MSO) by ITD (Grothe, 1993). With the assumption of a sound source located on the left side of the listener, the emitted sound waves will travel to the right ear with a time delay to the left one. In addition, due to the shadowing effect of the head, the right ear hears a significantly lower intensity than the left ear. ITD and IID feature in sound emission, providing the information for inferring the azimuth angle of the sound source in the horizontal plane, 0 degrees and 180 degrees. In the context of sound localization, azimuth is the angle between the north vector and the sound source's vector on the horizontal plane.

2.3 Review of Sound Source Classification Approaches

Sound classification is a crucial task in various applications, including surveillance, noise mitigation, and context-aware computing. Recent advancements in this field have led to the development of innovative techniques aimed at improving accuracy and robustness. This section provides an overview of recent sound source classification techniques and categorizes them into deep and non-deep approaches. It also delves into the challenges associated with each category.

2.3.1 Non-Deep Approaches

Non-deep or, in a more traditional approach, classic methods often depend on manually crafted features and may face difficulties in capturing the intricate spectral and temporal attributes of sound sources (Grumiaux et al., 2022). To effectively examine an auditory scene and faithfully represent the diverse array of sounds present in the world, auditory neurons must possess the capability to adapt their response characteristics continuously. This adaptation is necessary due to the dynamic alterations in acoustic energy across both spectral and temporal dimensions, which are prevalent in most natural sounds, including human speech (Barroso et al., 2023). For the effective analysis of an auditory scene and the precise representation of the extensive variety of sounds encountered in the world, auditory neurons need the capacity to constantly adjust their response characteristics to match the prevailing acoustic conditions (Grumiaux et al., 2022; Barroso et al., 2023). However, non-deep methods can be effective in certain scenarios, such as when the number of classes is small, and the dataset is limited. Various existing methods are studied and factors influencing the choice of method are discussed in Bansal et al. (2022) and Liaquat et al. (2021).

Non-deep approaches have been widely employed in the field of sound source classification. These approaches often utilize handcrafted features to represent audio signals and make classification decisions (Bianco et al., 2019). While non-deep methods may face challenges in capturing intricate spectral and temporal characteristics of sound sources, they have their own advantages in specific contexts. For instance, in scenarios where the number of distinct sound classes is small and the available dataset is limited, non-deep methods can demonstrate efficacy (Abayomi-Alli et al., 2022).

2.3.2 Deep Approaches

Deep learning approaches have gained significant traction in the localization of sound sources, not simply due to their capacity to acquire intricate features from data, but also due to their capability to address specific objectives that go beyond feature acquisition (Grumiaux et al., 2022). Although numerous machine learning techniques have the ability to acquire complex features from data, researchers in deep learning harness these approaches for more nuanced goals within the context of sound source localization. These approaches leverage neural networks, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to classify sound sources (Grumiaux, et al., 2021).

Deep learning algorithms has been extensively grasped the researchers' attentions in the field of sound source localization for various reasons. One of the main reasons is the potential to improve accuracy and precision in identifying the source of sound in various environments (Grumiaux et al., 2022; Lee et al., 2021). Deep learning techniques, especially CNNs and RNNs, have demonstrated remarkable performance in various fields, including computer vision and natural language processing (Grumiaux et al., 2022; Lee et al., 2021; Yiwere & Rhee, 2019; Tan et al., 2021). Researchers have explored their application in sound source localization to potentially improve accuracy and precision in identifying sound sources in various environments. Deep learning models have the ability to capture intricate patterns and features in audio data, which can be useful for distinguishing between multiple sound sources and localizing them accurately.

As an example, a review article on sound source localization with deep learning methods provides an overview of the application of deep learning techniques for single and multiple sound source localization. Another survey paper focuses on deep learning methods for sound source localization, particularly in indoor environments (Grumiaux et al., 2021).

In terms of specific studies, Tan et al. (2021) propose a CNN-based approach for sound source distance estimation using an image classification approach. Another study presents a regional localization method for indoor sound sources based on CNNs (Grumiaux et al., 2021). These studies demonstrate the effectiveness of deep learning methods in sound source localization tasks.

Deep learning methods offer advantages in handling challenging scenarios, such as noisy and reverberant environments, by automatically learning complex features from data (Mehrish et al., 2023). However, deep learning models depend on extensive and diverse datasets for training, which is of utmost importance for their effectiveness (Sarker, 2021). Although data are essential for most machine learning (ML) methods, deep learning places great emphasis on it. The significance of large datasets in deep learning is attributed to its capacity to autonomously acquire intricate hierarchical characteristics and representations from unprocessed data. Deep neural networks are made up of multiple interconnected layers of nodes, which allow for the detection of patterns and characteristics in the data. This ability to extract sophisticated representations distinguishes deep learning from conventional ML techniques (Shorten & Khoshgoftaar, 2019). Deep learning utilizes extensive datasets to independently uncover and refine characteristics, allowing it to address complex tasks such as image recognition, natural language processing, and speech comprehension with remarkable precision (Munappy et al., 2022). Although data are vital for many ML approaches, the depth and complexity of deep learning networks make them particularly reliant on substantial datasets to unveil intricate patterns and achieve exceptional performance (Sun & Scanlon, 2019). In conclusion, extensive and varied datasets are vital for deep learning models, enhancing their capacity to acquire sophisticated characteristics and representations, distinguishing them from conventional ML approaches. This capacity empowers deep learning to excel in numerous challenging tasks.

In summary, deep learning methods, including CNNs and RNNs, have shown promise in sound source localization tasks (Takeda & Komatani, 2016; Wu et al., 2021). They have the ability to learn complex features directly from data, making them effective in challenging environments. Research papers and surveys provide valuable insights into the application and performance of deep learning methods in sound source localization.

2.3.3 Environmental Sound Classification

Environmental sound classification (ESC) is a specialized domain within sound classification that focuses on categorizing sounds originating from natural environments (Nasiri & Hu, 2021; Meedeniya, et al., 2023). This field has gained significant attention due to its practical applications in environmental monitoring, wildlife conservation, and soundscape analysis. ESC plays a crucial role in discerning a wide range of acoustic

phenomena found in natural settings, such as the sounds of birds, water streams, wind rustling through trees, and various animal vocalizations. Researchers and practitioners in ESC employ advanced machine learning techniques, including deep neural networks and transfer learning, to distinguish between different sound classes effectively (Meedeniya, et al., 2023). This not only aids in understanding the acoustic characteristics of ecosystems, BUT ESC also contributes to environmental research and conservation efforts by automating the identification of specific sound events, ultimately helping scientists and conservationists better comprehend and protect natural environments.

One of the most comprehensive reviews of this field is presented by Nogueira et al. (2022). This survey encompasses various crucial facets of ESC in a survey, including the diverse range of datasets employed for training and assessment purposes, the preprocessing techniques applied to refine raw sound data, the extraction of pertinent features from the audio data, and the deployment of diverse classifiers for effective sound classification (Nogueira et al., 2022).

2.3.4 Urban Sound Classification

Urban sound classification is another pivotal subfield of sound classification, uniquely concentrated on the categorization of sounds sourced from urban environments. The cornerstone of this subfield is a meticulous literature review that systematically examines the methodologies underpinning classification and audio segmentation techniques. The intention of this review is to gain insights into the mechanisms that drive effective classification within urban soundscapes (Das et al., 2020).

Recent advancements in this domain have shown promising progress, emphasizing the integration of machine learning techniques (Lezhenin et al., 2019; Mushtaq et al., 2020; Massoudi et al., 2021). One notable approach involves the utilization of CNNs for urban sound classification. Studies have explored models that utilize both audio and contextual information, enhancing the accuracy of sound classification systems (Massoudi et al., 2021). Additionally, the exploration of SNNs offers a novel perspective in this field. SNNs, owing to their event-driven nature, can efficiently process temporal data, making them suitable for urban sound classification tasks (Arnault & Hanssens, 2020). Arnault and Hanssens (2020) introduced a solution in the DCASE 2020 challenge, emphasizing the importance of urban noise pollution monitoring (Arnault et al., 2020).

2.3.5 Machine Learning Models

In the quest for highly precise categorization of auditory sources, machine learning has made significant strides. ML methods are instrumental in extracting essential acoustic features during the training phase, allowing the confident identification of previously unheard auditory patterns. Key ML frameworks used for auditory classification encompass neural networks, decision trees, and support vector machines (SVMs) (Bianco et al., 2019). Notably, neural networks, known for their ability to grasp intricate patterns, excel at capturing complex associations within acoustic data.

Beyond these established frameworks, alternative ML approaches are available for auditory classification. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have made substantial contributions to this field (Tsalera et al., 2021). Moreover, transfer learning, involving the adaptation of pre-trained models fine-tuned for specific tasks, has proven valuable in enhancing accuracy and efficiency in auditory classification tasks (Ahmed et al., 2023). The choice of a ML framework or methodology depends on the specific characteristics of the acoustic data and the complexity of the classification task, highlighting the adaptability of ML in auditory classification.

Decision trees provide a transparent framework for sound classification, employing sequential decision pathways to differentiate between sound categories. Support vector machines (SVMs), with their expertise in defining decision boundaries in high-dimensional spaces, hold the potential for distinguishing intricate sound variations. Collectively, these machine learning models constitute a versatile toolkit for enhancing the precision of sound classification algorithms (Aggarwal et al., 2020; Nogueira et al., 2022).

2.3.6 Feature Extraction

In the landscape of sound classification, feature extraction emerges as a pivotal undertaking. This process serves as a bridge between raw audio signals and the classification algorithms, extracting pertinent information that unveils the inherent characteristics of sound sources. This endeavor holds immense importance as the chosen features significantly impact the discriminatory power of the classification model. Among the array of techniques available for feature extraction, a few stand out as widely embraced in the pursuit of precision (Abayomi-Alli et al., 2022).

Mel-frequency cepstral coefficients (MFCCs): MFCCs, inspired by the human auditory system, distill spectral information from audio signals while reducing the impact of noise and irrelevant variations. They capture spectral characteristics, such as pitch, timbre, and spectral shape, making them suitable for sound source discrimination (Abdusalomov et al., 2022).

Spectrograms: Spectrograms offer a visual representation of signal frequencies over time, providing insights into both temporal and spectral attributes and allowing the model to detect variations and patterns in sound sources (Mesaros et al., 2021).

Wavelet transforms: Wavelet transforms enable multi-resolution analysis, which is beneficial for extracting transient events and frequency components that might remain hidden with other techniques. They capture time-frequency information and can identify abrupt changes or subtle features in the sound data (Tzanetakis et al., 2001).

In the context of sound source localization and classification using spiking neural networks (SNNs), the feature extraction process plays a crucial role in revealing the inherent characteristics of sound sources. The chosen features significantly influence the effectiveness of the classification model. Integrating features and data augmentation can be one of the approaches to enhance accuracy in classification, briefly described in the next section.

2.3.7 Data Augmentation

In the realm of sound source localization and classification using SNNs, the feature extraction process is pivotal, shaping the model's efficacy. In this research study, the focus is on localization through ITD, IID, and energy-based methods, emphasizing their robustness. While data augmentation techniques, such as time stretching and pitch shifting, enhance the diversity and adaptability of the training dataset, the system designed in this study doesn't incorporate them in classification. However, their potential in enriching the model's adaptability across diverse scenarios cannot be overlooked.

In addition, sound classification often grapples with the challenge of limited data availability. Addressing this concern, data augmentation surfaces as a strategic solution to expand the size of the dataset. This technique endeavors to create new samples from

existing ones, infusing diversity and richness into the training pool (Aggarwal et al., 2020). Various augmentation methodologies, such as time stretching, pitch shifting, and noise addition, infuse the dataset with variations that mirror real-world conditions. This augmentation not only counteracts the scarcity of data but also contributes to the model's adaptability, enabling it to generalize better across different scenarios.

2.3.8 Transfer Learning

In the quest for refining sound classification performance, transfer learning emerges as a formidable technique. Capitalizing on the knowledge amassed from one task and applying it to another, transfer learning empowers the utilization of pre-trained models on vast datasets, subsequently fine-tuning them on smaller, specialized datasets (Nogueira et al., 2022). This strategy proves particularly beneficial in scenarios where data volume is limited. By harnessing the capacity of pre-existing knowledge, transfer learning augments the model's proficiency, enabling it to achieve commendable accuracy even with constrained resources.

In the context of our research, focused on designing a SNN-based localizer and classifier for sound events, transfer learning emerges as a valuable technique. Transfer learning bridges the gap between neuroscience and machine learning, using biologically realistic models of neurons, such as SNNs, to enhance the real-time performance of sound event detection (Nogueira et al., 2022).

Transfer learning with SNNs involves leveraging pre-existing knowledge from pre-trained models and adapting them to the specific task of sound event classification (Tsalera et al., 2021). The process entails fine-tuning the SNN model on smaller, specialized datasets tailored for sound event recognition. This approach proves particularly beneficial in scenarios where data volume is limited, which is often the case in real-time sound event detection (Ekpezu et al., 2021).

By harnessing the capacity of pre-existing knowledge, transfer learning augments the proficiency of SNNs. This augmentation is crucial for achieving commendable accuracy, especially in resource-constrained real-time applications where timely and accurate sound event classification is paramount (Li et al., 2018). Additionally, transfer learning allows SNNs to benefit from features extracted from diverse datasets, enabling the network to capture complex patterns in sound events (Lamrini, et al., 2023).

2.3.9 Challenges in Deep and Non-Deep Approaches

Deep learning techniques, celebrated for their potency, introduce their own set of challenges. Training deep neural networks necessitates substantial annotated datasets, a resource-intensive process that can impede progress (Yun et al., 2018). Moreover, the opacity of deep models complicates the understanding of their decision-making process, raising concerns about their interpretability (Piczak, 2015; Presannakumar et al., 2023).

Non-deep approaches, while adept, confront their own hurdles. These methods can struggle to encapsulate the intricate spectral and temporal characteristics of diverse sound sources (Umapathy et al., 2010). They can also be susceptible to environmental conditions like background noise and reverberation, potentially compromising their robustness (Picou et al., 2016). Embedding domain knowledge into these approaches emerges as an imperative strategy to bolster their efficacy (Nogueira et al., 2022).

Recent advancements in the field of sound source classification represent a diverse spectrum of approaches. These encompass a collaborative journey of deep learning and traditional methods, all contributing to the enhancement of accuracy and the reliability of sound source classification. Notably, deep learning plays a central role in this evolution by its innate ability to directly extract intricate features from data. Nevertheless, it is essential to acknowledge persistent challenges associated with deep learning, including the scarcity of labelled data and the interpretability of models. Concurrently, non-deep methods, which rely on the artful engineering of features and domain-specific knowledge, continue to hold a vital place in the progression of sound source classification. They offer a pragmatic alternative for addressing complex acoustic scenarios. In essence, the synergy between deep and non-deep paradigms is shaping the future landscape of sound source classification, harmonizing the strengths of both to navigate the evolving challenges in this domain.

Our study delves into the multifaceted challenges of sound source localization, a domain that requires innovative solutions. We explore the potential of adaptive and dynamic methods in addressing these challenges. Adaptive methods, which can adjust to varying conditions and sources, are critical in a world where sound sources are often in motion. Dynamic approaches, on the other hand, offer flexibility in handling multiple sources

simultaneously. This aligns with the ongoing shift in sound source localization towards real-world scenarios.

2.4 Adaptive Sound Source Localization Methods

Adaptive sound source localization methods are widely used to estimate the position of one or multiple sound sources in relation to a reference point based on recorded acoustic signals (Chung et al., 2022; Grumiaux et al., 2022). These methods have gained prominence in various applications, from robotics to hearing aids, where precise knowledge of sound source localization is crucial. In this section, we delve deeper into the realm of adaptive audio source localization, with a particular emphasis on single sound sources and the significant developments that have revolutionized this field.

Single sound source localization has been a long-standing focus of sound source localization research due to its importance in real-world applications. Recent advancements in this field have ushered in a new era of accuracy and robustness, primarily through the integration of sophisticated methodologies such as microphone arrays, machine learning, and binaural approaches**.**

Microphone arrays: Adaptive microphone arrays equipped with multiple microphones are an essential technology in the field of sound source localization (Adel et al., 2012). These arrays capture sound from various angles, providing valuable spatial information for identifying the origin of a sound. In this context, adaptive beamforming techniques, such as the minimum variance distortionless response (MVDR) method, are utilized to amplify the signal from the desired source while effectively suppressing undesired interference and noise (Adel et al., 2012; Marques et al., 2022). This technology is of utmost importance in domains like autonomous vehicles, where precise localization of sirens or horns is critical for avoiding collisions (Marques et al., 2022).

Adaptive beamforming is an indispensable signal processing technique employed with sensor arrays to enable directional signal transmission or reception (Nemade et al., 2014). It accomplishes this by detecting and estimating the signal of interest at the output of a sensor array, utilizing optimal spatial filtering and interference rejection (Prather et al., 2023). Techniques such as MVDR are commonly used in this context to

enhance the desired signal while mitigating the impact of undesired interference and noise (Boughaba et al., 2022).

Phased arrays are particularly practical in the higher frequency regions of the radio spectrum, such as UHF and microwave bands, where the wavelengths are conveniently small (Brown, 2021). Radio stations, especially AM broadcast stations, frequently utilize phased arrays to amplify signal strength within their licensed areas while minimizing interference in adjacent regions (Brown, 2021). This versatility extends to applications in radio and sound waves, as well as various domains including radar, sonar, seismology, wireless communications, radio astronomy, acoustics, and biomedicine (Pradhan, 2021).

One specific application of adaptive beamforming is the construction of beamforming microphone arrays, which can display heightened sensitivity to sounds from specific directions while attenuating sounds from other angles (Lashi et al., 2018). The delay-and-sum beamforming method, utilizing multiple microphones, is a prevalent technique for achieving this directional sensitivity (Perrot et al., 2021).

However, a limitation of this approach is that the array's geometry may require adjustments to accommodate varying sound source locations (Lashi et al., 2018). To surmount this limitation, current research efforts focus on refining microphone arrays for delay-and-sum beamforming using genetic algorithms (Lashi et al., 2018). These algorithms optimize the array's geometry and parameters, enhancing sound source localization and adaptability. This technique holds promise in various domains, including speech enhancement, noise cancellation, sound source localization, and environmental monitoring (Perrot et al., 2021). Those interested in implementing delay-and-sum beamforming can access resources that are readily available, including MATLAB code, which aids in understanding and prototyping this technique (Kurc et al., 2013). Additionally, researchers have explored a 3D impulsive sound-source localization approach using a 2D MEMS microphone array through the delay-and-sum beamforming technique, thereby further extending its applicability (Seo et al., 2017).

In this thesis, the aim of combining interaural intensity difference (IID), interaural time difference (ITD), and energy-based approaches is to reduce the dependency on a fixed microphone array and increase adaptability in sound source localization. By integrating

multiple cues such as intensity, time, and energy, the system becomes less reliant on specific microphone array configurations and can adapt to various sound source locations and environments effectively. This approach enhances the versatility and accuracy of sound source localization systems.

Machine learning: The integration of machine learning methodologies, particularly deep learning, has led to significant progress in single sound source localization. CNNs and RNNs have played a pivotal role in the transformation of adaptive audio source localization. These neural network architectures are designed to process audio data, effectively learning the intricate relationship between acoustic signals and the position of the sound source. This enables them to adapt seamlessly to varying environments, providing remarkable generalization capabilities. Deep learning-based methods excel in scenarios where the sound source characteristics or acoustic environment may change over time, such as in smart audio devices, offering enhanced accuracy and adaptability.

Binaural approaches: Inspired by the complex workings of the human auditory system, binaural methods have been developed for single sound source localization. These methods use a pair of microphones, replicating the human ears, to approximate the source's location based on interaural time and level differences. By mimicking the way humans perceive sound directionality, binaural approaches provide reliable localization results, even in scenarios where a single source is moving. This capability is particularly valuable in applications where tracking the position of a mobile sound source is crucial, such as robotics and surveillance systems (Rascon & Meza, 2017; Grumiaux et al., 2022).

The binaural approach will receive additional attention for further review due to a major objective of this study is to design a spiking neural network (SNN) structure that incorporates binaural principles. This SNN-based approach demonstrates the potential to enhance sound source localization through the integration of binaural cues, making it a noteworthy development in the field. The review aims to explore the innovative aspects and implications of a SNN-based method, shedding light on the significance of combining binaural techniques with neural network structures for improved sound source localization.

2.5 Dynamic Methods for Sound Source Localization

Traditional techniques for identifying acoustic origins have predominantly focused on fixed or slowly moving sources, which are insufficient for scenarios where sources are constantly changing position (Nakamura et al., 2009). To detect dynamic acoustic sources, various approaches such as Kalman filtering, particle filtering, and deep learning-based strategies are necessary to simulate the motion dynamics of the source and track it in real time. Additionally, sensor array configurations such as microphone arrays and distributed sensor networks are also employed to improve localization accuracy (Chiariotti et al., 2019; Chung et al., 2022).

The challenges encountered in detecting dynamic acoustic sources encompass achieving low-latency tracking, enhancing localization robustness in diverse acoustic environments, and integrating data from various sensor modalities (Nakamura et al., 2009; Grohn et al., 2002). The detection of moving sound origins has practical consequences in multiple areas, such as self-governed automation, interaction between human and robot, safeguard and monitoring systems, enhanced reality, and virtual reality (Chiariotti et al., 2019; Chung et al., 2022). Consequently, the comprehension and advancement of dynamic approaches are crucial in effectively addressing the challenges of dynamic environments (Chiariotti et al., 2019).

2.5.1 Dynamic Sound Source Models

Kalman filtration is a fundamental technique for dynamic acoustic origin determination (Portello et al., 2011). It approximates sound source positions by exemplifying the kinetics of the source's movement (Zhang et al., 2021). Kalman filters find widespread usage in acoustic origin determination due to their capacity to manage noisy measurements and non-linear systems (Gehrig et al., 2005).

Kalman filters operate by recursively approximating the state of a system on the basis of an array of noisy measurements (Gehrig et al., 2005). The principles underpinning Kalman filters entail predicting the state of the system at every time step, updating the prediction with a new measurement, and amalgamating the prediction and measurement to acquire an improved assessment of the system state (Gehrig et al., 2005).

The benefits of Kalman filters include their ability to cope with noisy measurements, their computational efficiency, and their ability to furnish estimates of the system's uncertainty (Portello et al., 2011). Nonetheless, Kalman filters possess limitations, such as their dependence on a linear model of the system and their susceptibility to model errors and incorrect assumptions about the system (Portello et al., 2011). Recent research has investigated the utilization of Kalman filters in conjunction with neural networks to enhance the accuracy of acoustic origin determination (Zhang et al., 2021).

2.5.2 Deep Learning-Based Approaches

Recent advancements in the realm of deep learning have resulted in the creation of neural network architectures that specifically cater to the task of dynamic sound source localization (Grumiaux et al., 2022; Yalta et al., 2017; Wang & Cavallaro, 2022). These cutting-edge methodologies prioritize their ability to adapt to various environmental conditions and simultaneously track multiple sources (Grumiaux et al., 2022; Wang & Cavallaro, 2022).

Deep learning-based techniques utilize neural networks to comprehend the correlation between acoustic signals and sound source locations, enabling them to handle intricate and non-linear systems (Wang & Cavallaro, 2022). Such approaches are also equipped to manage noisy measurements and enhance localization accuracy in reverberant environments (Yalta et al., 2017). Deep learning-based techniques have been employed for diverse applications, such as indoor acoustic source localization with microphone arrays (Wang & Cavallaro, 2022), sound source localization from a flying drone (Wang & Cavallaro, 2022), and real-time tracking of multiple sound sources (Grumiaux et al., 2022).

Deep learning-based methodologies do present certain limitations, such as the necessity for copious amounts of training data and computational resources, as well as the difficulty of deciphering the outcomes of the neural network (Grumiaux et al., 2022). Future research in this area may concentrate on enhancing the intelligibility of deep learning-based approaches and developing more effective training methods (Grumiaux et al., 2022).

2.6 Sensor Array Configurations

2.6.1 Microphone Arrays

The implementation of microphone arrays is a ubiquitous selection for the purpose of acoustical source localization (Chung et al., 2022). Microphone arrays are comprised of multiple microphones arranged systematically to capture sound signals emanating from distinct directions. The design and implementation of microphone arrays hinge on the specific application and the requisite precision of sound source localization.

Microphone arrays offer a plenitude of advantages, including high spatial resolution, discernment between multiple sources, and the ability to suppress noise and reverberation (Praveen et al., 2022). However, they also encounter challenges when confronted with dynamic sources, such as the need for precise calibration, sensitivity to microphone placement and orientation, and the complexity of handling moving sources (Chung et al., 2022; Praveen et al., 2022).

2.6.2 Distributed Sensor Networks

Distributed sensor networks present robust solutions for dynamic sound source localization (Praveen et al., 2022). These networks consist of various sensor types and communication mechanisms, including microphones, cameras, and wireless networks. Distributed sensor networks can effectively manage large and intricate environments, and they offer scalability and redundancy. Furthermore, they can provide multiple modalities of data, such as audio and video, which can enhance localization accuracy and robustness (Praveen et al., 2022).

Distributed sensor networks also face obstacles, including the necessity for synchronization and coordination between sensors, the intricacy of handling data fusion from multiple modalities, and the potential for communication delays and failures (Zhang et al., 2021). Recent research has explored the utilization of neural networks and Kalman filtering, in conjunction with microphone arrays and distributed sensor networks, to enhance sound source localization accuracy (Zhang et al., 2021).

2.7 Challenges in Dynamic Sound Source Localization

Dynamic sound source localization faces several challenges that need to be addressed to achieve accurate and reliable results.

2.7.1 Real-time Tracking

The dynamic nature of sound sources necessitates real-time tracking capabilities (Nakamura et al., 2009). Achieving low-latency tracking is a challenge due to the need for fast processing of large amounts of data. Strategies to mitigate this challenge include optimizing the sensor array configuration, using efficient algorithms such as Kalman filtering and particle filtering, and leveraging parallel processing and hardware acceleration (Nakamura et al., 2009; Gehrig et al., 2005).

2.7.2 Robustness in Complex Environments

Dynamic sources often operate in diverse acoustic environments, from quiet rooms to noisy urban settings. Improving localization robustness in such varied conditions is a challenge due to the need to handle noise, reverberation, and interference from other sources. Strategies to mitigate this challenge include using sensor arrays with high spatial resolution, developing robust algorithms that can handle noise and reverberation, and using distributed sensor networks that can provide multiple modalities of data (Huang et al., 2019; Zhang et al., 2021).

2.7.3 Multimodal Sensor Integration

Combining data from different sensor modalities, such as audio and video, can enhance sound source localization accuracy (Gehrig et al., 2005). However, multimodal sensor integration presents challenges such as data synchronization, data fusion, and the need for specialized hardware and software. Strategies to mitigate these challenges include developing efficient algorithms for data fusion, using specialized hardware such as synchronized cameras and microphones, and leveraging machine learning techniques to learn the relationship between different modalities of data (Gehrig et al., 2005; Zhang et al., 2021).

2.8 Spiking Neural Networks in Sound Localization

Sound localization with spiking neural networks (SNNs) is emerging as a promising avenue for achieving accurate and biologically inspired sound source localization (Goodman & Brette et al., 2010).

SNNs are a type of neural network that closely resembles actual neural networks in the brain (Pietrzak et al., 2023). They have the potential to be more energy efficient than

artificial neural networks (ANNs) on event-driven neuromorphic hardware. SNNs operate based on the timing and frequency of spikes, enabling them to process temporal information more effectively.

In the context of using SNNs for sound localization applications, several noteworthy studies have demonstrated encouraging results, leveraging neural mechanisms akin to those in the auditory system. One of the key advantages of SNNs is their biologically plausible nature. They are designed to mimic the behavior of real neurons, allowing for a more accurate representation of neural processing. This makes SNNs a valuable tool for studying the neural mechanisms underlying sound localization (Li et al., 2023).

For instance, Wall (2008) and Glackin (2010) undertook two studies focused on the localization of noise-free single-frequency tones (5 kHz, 15 kHz, and 25 kHz). They employed a spiking neural network model for sound localization based on interaural intensity difference (IID) cues, utilizing head-related transfer function (HRTF) acoustic data from adult domestic cats (Wall, 2012). The model was trained to learn spike patterns using the ReSuMe rule (Ponulak, 2006), with IID cues encoded as spike rates of lateral superior olive (LSO) neurons. Impressively, the achieved error rates of approximately 10%, with an error tolerance of ±10°, indicate the potential of SNNs in achieving accurate sound localization.

Other studies have explored the encoding of interaural time difference (ITD) cues for sound localization. Voutsas (2007) introduced a binaural sound source lateralization neural network (BiSoLaNN), drawing inspiration from the Jeffress model. The network employed a single delay line to encode ITD cues, and while it achieved a robust localization accuracy of 72.5% within the range of ±45°, it primarily focused on low-frequency pure tones between 440 Hz and 1,240 Hz.

Further advancements in SNN-based sound localization are found in Glackin's (2010) work, which concentrated on sound frequencies ranging from 270 Hz to 1.5 kHz. This study utilized the spike timing dependent plasticity learning rule (STDP) and incorporated experimentally observed HRTF data from adult domestic cats. The research achieved accuracy of 91.82% with an error tolerance of ±10°, highlighting the potential of STDP-based SNNs for accurate sound localization.

Moreover, Pan et al. (2021) expanded the horizon of SNN-based sound source localization to multiple microphone array processing. They introduced an innovative ITD-SNN model that utilized multi-tone phase coding for ITD cues. This novel approach yielded a remarkable 5-degree-angle resolution of sound source direction, showcasing the ability of SNNs to advance sound localization methodologies in complex scenarios.

In summary, the exploration of SNNs for sound localization applications has yielded promising outcomes. From utilizing IID cues with the ReSuMe rule to STDP-based models and multi-tone phase coding approaches, these studies collectively illuminate the potential of SNNs to achieve biologically inspired and accurate sound source localization. These advancements signify a promising trajectory in the realm of auditory perception research and signal processing.

2.9 Spiking Neural Network in Sound Source Classifications

As indicated in the previous section, SNNs have been used successfully to locate sound sources. SNNs are generally less frequently used to simultaneously classify the type of sound source and its location. However, SNNs have been successful in classifying patterns.

SNNs are a class of neural networks motivated by event-based computation. Many temporal learning rules have been proposed to train SNNs to perform temporal pattern classification tasks. Depending on how the error function is formulated, it can be classified as spike time-based (Ponulak, 2010; Mostafa, 2017) or membrane potential-based (Han, 2020). The main purpose of spike time-based learning rules is to minimize the time difference between the actual and desired output spike patterns by means of updating the synaptic weights. In contrast, membrane potential-based learning rules use the voltage difference between the actual membrane potential and the firing threshold to drive synaptic weight updates.

SNNs manifest high potentiality in various pattern recognition problems (Orchard, 2015; Zhao, 2014). Recent studies have further demonstrated the computational power of spike neurons and perceptrons for processing spatio-temporal patterns, demonstrating computational advantages in processing both visual and auditory sensory signals (Yu, 2015). Although these methods are mainly focused on image-related tasks, their

applications in the field of speech recognition are still largely unexplored. Sound detection in noisy environments is a challenging task in audio processing.

Early sound event detection methods borrowed speech recognition and music information retrieval techniques. As a result, these methods were based on traditional model classification techniques such as Gaussian mixture models (GMMs) and hidden Markov models (HMM). However, these models are most useful in modeling speech and music through specific techniques capable of modeling fundamental units of speech or music, such as phoneme state binding or time evolution. Phonemes and notes events often do not contain speech-like base units, making GMMs and HMMs less relevant for sound event detection (SED). In contrast to speech signals, sound events have more pronounced time-frequency characteristics that can be processed by the human auditory system, which has a lower signal-to-noise ratio (SNR).

Accordingly, SNNs play an important role in pattern recognition. Recent studies have shown that neural networks using the temporal encoding of audio signals can be used to classify sound events (Zhao, 2014). Yu (2015) presents a general structure for pattern recognition, using the temporal coding.

Generally, SNN-based classifiers utilize an encoding method to convert the auditory signals into spike train patterns. A great number of self-organized SNNs, equipped with STDP-based concurrent learning rules, have been developed to detect entire repetitive patterns by sequential firing (Masquelier, 2009). STDP-based learning rules are used in more complex feedforward networks for classification tasks with promising results (Kheradpisheh, 2018; Diehl, 2015). Thus, STDP-based learning rules offer an attractive alternative for training deep SNNs.

Meanwhile, optimal tuning parameters for accuracy, execution time, and learning strategy, as well as network structure properties such as neuron population size and density of connectivity, are major challenges in using SNNs for sound source localization and classification.

2.10 Summary

In this chapter, we have undertaken an extensive exploration of sound localization and classification approaches. These methods have been categorized into three groups: deep, non-deep, and hybrid methods. Our discussion has illuminated the benefits of hybrid methods, which leverage both classic and deep techniques to enhance their performance.

We have emphasized the practical utility of adaptive methods, particularly in real-time applications. We have also highlighted the significance of a dynamic approach in refining the modeling process for tracking, localization, and classification tasks.

We have investigated the role of spiking neural networks (SNNs) in sound classification and localization, recognizing their potential in these domains. Notably, our research seeks to amalgamate binaural methods, such as ITD and IID, with monaural methods, alongside an energy-based approach. This integration aims to reduce reliance on microphone arrays and enhance accuracy through data augmentations.

We have addressed the dynamic and adaptive approaches to sound localization and classification, emphasizing their pivotal role in augmenting accuracy and robustness within various environmental conditions.

A central inquiry in our exploration is determining the optimal and efficient SNN structure. To address this, the next chapter examines background theories pertaining to dynamic network structures, which are underutilized in SNNs, but widely adopted in classification applications. Subsequently, we investigate SNN architecture, different learning techniques, and encoding approaches to obtain the necessary knowledge and tools for the design of our new adaptive SNN structure, catering to real-time and multi-sound source detection, localization, and classification.

3. Theoretical Background

3.1 Introduction

In this chapter, we delve into the theoretical underpinnings that form the bedrock of our research, creating a bridge between our previous exploration of ART theory, SNN architecture, and sound processing methods, and the critical research questions that define the core of this work.

The amalgamation of ART theory and computational neuroscience, particularly in the form of spiking neural networks (SNNs), serves as a unique backdrop for our investigation. We have laid the groundwork for understanding how artistic principles can be intertwined with cutting-edge technology, paving the way for innovative applications in the field of sound processing.

Our exploration encompasses a multifaceted approach to sound, beginning with the fundamentals of sound source localization and classification. In this chapter, we elucidate the essential concepts that underpin our research questions. These questions not only arise from the knowledge acquired in the preceding chapters, but also address the complexities and nuances of implementing SNNs in the realm of sound processing.

As we transition from theory to practical application, we will address key questions such as improving sound localization accuracy by synergizing energy-based cues with traditional interaural time difference (ITD) and interaural intensity difference (IID) methods in SNN structures. We will also explore the factors influencing the determination of the azimuth sign when locating a sound, and the optimal arrangement and connection of neurons within the SNN structure to effectively process spatio-temporal data from incoming signals.

Furthermore, our inquiry extends to sound event triggering and classification, where we define effective methods for encoding audio data into electrical impulses suitable for SNNs in noisy environments and establish the appropriate thresholds for initiating neuronal activity. These questions culminate in a comprehensive system evaluation, evaluating the dynamic spatio-temporal rSNN architecture's performance in sound localization compared to conventional approaches and understanding the impact of

adaptive size increases and dynamically assigned neuron positions on the learning process and overall model performance.

3.2 Sound Source Localization

Sound source localization is a crucial aspect of auditory perception, allowing us to determine the position of a sound source in three dimensions: azimuth, height, and distance (Risoud et al., 2018).

The auditory system utilizes spatial cues provided by the interaction of sound with the head and external ears to derive the location of a sound source. These cues are analysed in specific brainstem pathways and integrated as cortical representations of locations. The localization of sound sources is most accurate in the horizontal and vertical dimensions, with cues such as ITDs and ILDs playing a significant role. However, localization in distance is less precise and can be influenced by factors such as the acoustics of the surroundings and familiarity of source spectra and levels (Middlebrooks, 2015).

Various methods have been developed to achieve SSL. These methods generally comprise two main steps, illustrated in Figure.3.

Figure 3

Common Main Steps of Sound Source Localization Methodology

In Figure 3, deep learning methodologies do not obviate the need for preliminary processing stages; instead, they frequently necessitate specific preparatory measures to optimize the training of the model. Although deep learning models possess the capacity to autonomously acquire features from raw data to some extent, in numerous instances preprocessing remains imperative to guarantee that the data is in an appropriate format and of sufficient quality for training.

Preprocessing in the realm of deep learning can encompass a multitude of tasks, including:

1. Data sanitization: Eliminating extraneous or noisy data points that have the potential to detrimentally impact the performance of the model.

2. Data standardization: Rescaling the data to conform to a standardized range in order to prevent any particular feature from exerting disproportionate influence.

3. Feature manipulation: Generating novel features or modifying existing ones to more accurately represent the underlying patterns present within the data.

4. Data encoding: Converting categorical data into a numerical format that can be readily utilized by deep learning models.

5. Data expansion: Augmenting the dataset through the generation of variations of preexisting data, a technique frequently employed in the context of image data.

6. Treatment of missing data: Addressing the issue of missing values through the process of either imputation or elimination.

While deep learning models are capable of autonomously extracting certain features via the utilization of neural network layers, the appropriate application of preprocessing techniques can enhance the efficiency and effectiveness of the learning process. The extent to which preprocessing is required is contingent upon the specific problem at hand, the characteristics of the data, and the deep learning architecture that is being employed.

The mammalian auditory system uses various cues for sound source localization, including time difference, level difference, and spectral information. The brain utilizes subtle differences in intensity, spectral, and timing cues to allow mammals to locate sound sources. Localization can be described in terms of a three-dimensional position: azimuth, elevation, and distance (for static sounds) or velocity (for moving sounds).

In practical applications, such as locating sound sources using multiple sound sensors, techniques such as triangulation can be employed to estimate the location of the sound

source. However, distinguishing the sound source from background noise can be challenging.

Overall, sound source localization is a complex process that relies on the integration of spatial cues and the analysis of various auditory cues. Understanding the theories and mechanisms behind sound source localization is essential for developing accurate and reliable localization systems.

This study focuses on preprocessed signal utilization, but ART-rSNN has the potential to be modified so that it obtains the ability to analyze raw signals by making the structure robust, especially in coding the input signal procedure (Meng & Xiao, 2017).

Generally, there are five popular main methods for feature extraction, briefly described below (Liaquat, 2021).

3.2.1 Energy-based localization methods

Energy-based localization methods are a popular approach for sound source localization. These methods utilize the energy ratios of the sensors and the target, which are restricted to a hypersphere (Meng & Xiao, 2017).

By analyzing the energy attenuation, which is inversely proportional to the distance from the sound source to the acoustic sensors, these methods estimate the position of the sound source. One advantage of energy-based localization methods is that they do not require synchronization among multiple sensors, making them more practical and efficient. Additionally, these methods do not necessarily require multiple microphones for each node, reducing the complexity and cost of the system (Meng & Xiao, 2017).

The localization accuracy of energy-based methods can be improved by increasing the number of sensors, as this increases the number of hyperspheres and allows for more precise intersection points, which correspond to the location of the sound source (Li et al., 2023).

Energy-based localization methods have been studied in various research studies. For example, a maximum likelihood approach has been proposed for acoustic source position estimation, taking into account noisy and correlated environments (Meng et al., 2017). Another study introduced an energy-efficient scheme for target recognition and

localization in resource-constrained sensor nodes (Algobail et al., 2019). These methods have shown promising results in sound source localization, providing a practical and efficient solution for estimating the position of sound sources. Energy-based localization methods are particularly suitable for applications in wireless sensor networks, where energy efficiency and resource constraints are important considerations.

3.2.2 Time-Dependent Localization Methods (TD)

These methods work with the time difference between the signals, based on different measurement methods. Well-known examples of this technique are defined as follows:

Time-of-arrival (TOA) depends on time instants, in which the microphones detect the received signal, by making use of sensors cooperating together, to estimate the signal propagation time. In case of non-availability of the cooperation, TOA is not sufficient to determine the propagation time of the solitary signal (Liaquat, 2021). The propagation time can be expressed mathematically as:

$$t_{prop} = \frac{d}{c} \qquad (3\text{-}1)$$

where t_{prop} is the propagation time, d is the distance between the microphones and the target, and c is the speed of sound.

Time-of-flight (TOF) is an approach that approximates the distance between microphones and the target. The distance can be calculated as follows:

$$d = c \times t_{prop} \qquad (3\text{-}2)$$

where d is the distance, c is the speed of sound, and t_{prop} is the propagation time.

Time difference of arrival (TDOA) is a method that works based on the time difference between the signals. In another words, the time difference between zero level crossings or between the onset times of both signals is measured. To implement this method, there is a need to choose a reference node to nullify the noise factors and ease the synchronization needs. Hence, the accuracy of sound source localization greatly depends on the choice of this reference. The time difference can be described by:

$$\Delta t = t_2 - t_1 \qquad (3\text{-}3)$$

where Δt is the time difference, and t_1 and t_2 are the times at which the signals are detected by the two microphones.

3.2.3 Direction of Arrival (DOA) Methods

Each node in this approach estimates the direction of arrival (DOA) of the sources and transmits the results to the center. Since each node makes individual estimates, synchronization is not essential as long as the motion of the source is very low. This method uses triangulation of the points in localizing the source. DOA needs more computational power and multiple microphones (Shaukat, 2021; Liaquat, 2021).

The MUSIC algorithm is one of the most widely used DOA estimation algorithms. It uses the eigenvalues and eigenvectors of the covariance matrix of the received signal to estimate the DOA of the sources. The MUSIC algorithm can be expressed as follows:

$$\hat{\theta} = \arg\max_{\theta} \frac{1}{\mathbf{a}^H(\theta)\mathbf{R}_n\mathbf{a}(\theta)} \qquad (3\text{-}4)$$

where $\hat{\theta}$ is the estimated DOA, θ is the DOA angle, $\mathbf{a}(\theta)$ is the steering vector, $\mathbf{R}_n$ is the covariance matrix of the noise, and $\arg\max$ is the argument that maximizes the expression.

The ESPRIT algorithm is another popular DOA estimation algorithm that uses the eigen structure of the received signal to estimate the DOA of the sources.

The adaptive directional time-frequency distribution (ADTFD) algorithm is another DOA estimation algorithm that uses spatial averaging of TFDs and multi-component analysis.

3.2.4 Beamforming

Beamforming is a signal processing technique used to enhance a desired signal from a specific direction while reducing interference from other directions. It is commonly used in planar microphone array systems and has a wide range of applications in various fields such as radar, sonar, seismology, wireless communications, radio astronomy, acoustics, and biomedicine.

The process of beamforming involves filtering the microphone signals and combining their outputs to extract the desired signal while suppressing interfering signals based on their spatial location. By leveraging the spatial information captured by the microphone array, beamforming can effectively separate sources with overlapping frequency content originating from different locations (Adel et al., 2012).

One of the key advantages of beamforming is its ability to remove noise and reverberation from speech signals by taking advantage of the spatial information captured by the microphone array. Techniques like Wiener beamforming have been developed to model the microphone array and enhance speech signals by reducing unwanted noise and reverberation.

Beamforming is a powerful tool for multichannel signal processing, enabling the extraction of desired signals while mitigating interference. It offers significant advantages in terms of spatial filtering and can be applied to various scenarios where directional signal transmission or reception is required (Ma et al., 2020).

Minimum Variance Distortionless Response (MVDR) Beamformer

One of the commonly used mathematical formulas for beamforming is the minimum variance distortionless response (MVDR) beamformer, which is also known as the Capon beamformer. It minimizes the output power subject to the constraint that the desired signal is preserved. The MVDR beamformer output, $y(n)$, is given by:

$$y(n) = w^H x(n) \qquad (3\text{-}5)$$

where w^H is the conjugate transpose (Hermitian) of the weight vector w, and $x(n)$ is the vector of microphone signals. $y(n)$ is the output of the beamformer.

The weight vector w that minimizes interference and preserves the desired signal can be calculated as:

$$w = \Gamma^{(-1)} d / (d^H \Gamma^{(-1)} d) \qquad (3\text{-}6)$$

where d is the steering vector that characterizes the spatial relationship between the microphones and the desired signal direction, and Γ is the covariance matrix of the microphone signals. The given mathematical formula is the fundamental expression for MVDR beamforming, where the weight vector is optimized to increase the signal-to-

interference-plus-noise ratio (SINR) in the direction of the desired signal while reducing interference from other directions.

3.2.5 Inter-Microphone Intensity Difference (IMID)

This method measures the difference of energy between the signals at any instance. The obtained signal helps to determine whether the source is at the right, left, or front of the microphone. To increase the resolution, a greater number of microphones can be utilized. In the case of using frequency domain approaches, the method turned to inter-microphone level difference (IMLD) techniques, and the spectrum level of the signal is investigated.

Although the above methods are developed in different areas of SSL, still there are some challenges such as the reduction of estimation accuracy in high reverberant environments and the possibility of misunderstanding the spectral characteristics of undesired background noise with the source signals. Machine learning has been suggested as a powerful tool to deal with complex circumstances and uncertain environments in SSL (Liu. et al. 2021; Pan et al., 2020; Zhang et al., 2021).

Recently, sound event detection and localization (SEDL) has gained attention due to the benefits of event trigger-based technologies, which reduce energy consumption and optimize communication bandwidth (Mesaros et al., 2021). SEDL is an audio processing task that involves jointly detecting and localizing temporally targeted classes of sound events in space.

To address the challenges of complex circumstances and uncertain environments, several studies have suggested to utilize spiking neural networks (SNNs) as bio-inspired machine learning approaches (Wu et al., 2018; Wall et al., 2012). SNNs are biologically plausible and have strong performance in pattern recognition, making them a suitable tool for sound source localization. However, there is a tendency in the research area to focus on both static and dynamic environments, with the latter being more representative of real-world scenarios. In dynamic environments, sound sources are in motion, and most studies focus on locating and quantifying distributed noise sources, which require many microphones to produce a satisfactory map (McLoughlin et al., 2017). Localizing multiple moving sound sources is a more challenging problem, as even the smallest overlap of sources can disrupt the localization of the original source. To

address this issue, some sound localizing strategies employ a temporal tracking model, and most of them use parametric methods that require manual tuning of multiple parameters (Mesaros et al., 2021). Several ongoing research and development efforts aim to improve the accuracy of dynamic sound source localization. Some of the approaches use audio cueing and audio spatialization, which can enhance localization accuracy in immersive environments (Planinec et al., 2023; Gronh et al., n.d.). Other studies focus on developing systems with new functions, such as localization with generalized eigenvalue decomposition of correlation matrices for noise robustness (Substance Abuse and Mental Health Services Administration, 2019). Additionally, the use of individual head-related transfer functions (HRTF) in binaural synthesis, instead of generic HRTF, can lead to improved accuracy and quality of dynamic sound source localization and recognition ability (Planinec et al., 2023). Finally, dynamic sound localization during rapid eye-head gaze shifts has also been studied (Vliegen et al.,2004).

In the case of localizing moving sound sources, a few sound-localizing strategies employ a temporal tracking model, and most of them use parametric methods in which manual tuning of multiple parameters is important (Adavanne et al., 2019). For example, Adavanne et al. (2019) developed a tracking model using recurrent neural networks and deep learning methods to improve the performance of multiple moving sound source localization. However, this approach triggers high computational costs as well as higher angular error. Moreover, although sound event detection and localization (SEDL) systems have been recently developed using convolutional recurrent neural networks (CRNNs), the recurrent nature of CRNNs poses a significant challenge in implementing them efficiently on embedded hardware due to strenuous computations, high memory bandwidth requirements, and large memory buffers (Adavanne et al., 2019).

There is a need to design a new framework, able to accurately detect the sound events and classify the target locations, as well as to track and localize them. This framework needs to benefit from lower computational cost as well as preserve the accuracy of sound source localization.

3.3 Dynamic Sound Source Classification

The world of sound source classification is undergoing a remarkable transformation, driven by the rapid advances in deep learning techniques. Dynamic sound source

classification provides invaluable insights into the far-reaching impact of sound across diverse contexts.

Within the framework of this thesis, we embark on an exploration of dynamic sound source classification, a domain that holds a prominent position in contemporary research. The inclusion of this review acknowledges the pivotal role that adaptability and precision in sound source classification play in a multitude of applications. From the analysis of urban soundscapes to the recognition of environmental acoustics, the importance of dynamic sound source classification is undeniable.

Dynamic audio source classification is an essential facet of auditory signal processing and environmental sound analysis. It involves distinguishing and classifying audio signals that change with time, such as human conversation, animal vocalizations, and environmental noises. Diverse classification methodologies – including machine learning algorithms, deep learning networks, and feature extraction – are utilized to discriminate among distinct sound groups (Nogueira et al., 2022; Zhang et al., 2023).

While human conversation, animal vocalizations, and environmental noises all have their unique characteristics, they often occur together and interlink with each other in real-world applications. In scenarios like an urban sound scape or a natural monitoring station, the sounds produced by humans, animals, and environmental sources are interacting and meshed together. Designing such a technology to differentiate between these sounds within one model offers advantages in unifying the approach, leading to accurate detection and classification in mixed and dynamic environments.

This is very useful in integrated real-time applications for which security monitoring, wildlife tracking, and smart city management require fast detection of the predominant type of sound to trigger an action. Besides, with one multi-class model, there is saving of computation and avoidance of complexity in parallel deployment for different models of each category of sound.

Dynamic sound source classification can be framed as a supervised learning problem. Given a dataset of sound signals and their corresponding labels, the goal is to learn a function (f) that maps the input sound signals (X) to their respective classes (Y).

$$Y = f(X) + \varepsilon \quad (3\text{-}7)$$

where *(Y)* is the true label, *(X)* is the input sound signal, *(f)* is the classification function, and ε represents noise or error.

3.3.1 Deep Learning Architecture

Researchers have investigated various deep learning architectures, encompassing CNNs and RNNs, to enhance performance in dynamic sound source classification tasks (Grumiaux et al., 2022; Purwins et al., 2019). These deep learning models possess multiple layers, empowering them to acquire intricate hierarchical representations of sound data. Deep learning architectures have demonstrated superior performance on a broad spectrum of sound source classification challenges, comprising speech recognition, music information retrieval, and environmental sound detection and localization (Grumiaux et al., 2022; Purwins et al., 2019). For instance, a recent investigation proposed a deep-learning-based sound source classification model for concrete pouring work monitoring at a construction site, attaining high accuracy in categorizing distinct types of sounds (Kim et al., 2023). Another exploration explored the employment of data augmentation in a CNN sound source classification mechanism, demonstrating enhanced performance in sound source classification tasks (Chu et al., 2023). These investigations evince the enduring interest and advancement in the utilization of deep learning architectures for dynamic sound source classification.

3.3.2 Convolutional Recurrent Neural Networks (CRNNs)

Convolutional recurrent neural networks (CRNNs) are a powerful deep learning architecture that combines the strengths of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CRNNs have gained popularity due to their ability to handle both spatial and temporal aspects of sound data, making them effective for dynamic sound source classification tasks (Choi et al., 2017; Kwak & Chung, 2020).

3.3.3 Main Steps for Dynamic Sound Source Classification in CRNNs

Input data: CRNNs receive as input a sequence of spectrograms or mel-spectrograms. Each spectrogram represents the sound signal at a specific time, effectively capturing the time-frequency information of the audio.

The model needs re-training for every new environment in order to optimize the parameters according to the unique situation, such as acoustic conditions, background noise, and specific sound characteristics that may highly influence performance. This is because models that are trained on one dataset may depict poor performance upon application to various environments due to domain shifts including variations of recording conditions and ambient settings.

Convolutional layers: The initial convolutional layers in the CRNN are accountable for feature extraction. They utilize two-dimensional convolutions to the spectrogram, acquiring spatial patterns that depict significant acoustic features.

Recurrent layers: The recurrent layers, commonly employing long short-term memory (LSTM) or gated recurrent unit (GRU) cells, process the temporal aspects of the audio. This permits the network to apprehend the sequential dependencies in the sound signal.

Combining spatial and temporal features: CRNNs excel because they can proficiently merge spatial features extracted by the convolutional layers with temporal information captured by the recurrent layers. This amalgamation enables them to comprehend both the content and the context of the audio.

Classification output: The final layers of the CRNN often consist of fully connected layers that generate the classification output. The network learns to anticipate the class labels or probabilities based on the extracted features.

3.3.4 Training a CRNN for Dynamic Sound Source Classification

The process of training a CRNN for the classification of dynamic sound sources entails a series of distinct stages.

First, it is necessary to gather a dataset that is appropriately labelled, consisting of sound signals accompanied by their corresponding class labels. These sound signals can be obtained through the use of microphones in the surrounding environment, or they can be generated by employing sound simulators.

Second, the sound signals need to be pre-processed in order to enhance their overall quality and eliminate any unwanted noise. Common techniques utilized for this purpose include noise reduction and signal normalization.

Third, relevant features need to be extracted from the sound signals. This entails selecting and isolating key characteristics that are conducive to sound analysis. Popular choices in this regard include mel-frequency cepstral coefficients (MFCCs) or log-mel spectrograms, both of which are specialized representations optimized for sound analysis.

Finally, the CRNN must be trained using a supervised learning approach. This involves defining a suitable loss function and optimizer. Through this training process, the network effectively learns to map the extracted features to their respective class labels.

Researchers continuously explore various manifestations and improvements to enhance the effectiveness of CRNNs in different domains, such as the identification and localization of auditory events (Choi et al., 2017; Vesperini et al., 2020). For example, a recent investigation postulated an algorithm for the detection of snoring sounds founded on CRNNs and augmentation of acoustic data (Vesperini et al., 2020). Another research endeavour analysed the utilization of secondary attributes in deep neural networks for the identification of auditory occurrences (Kwak & Chung, 2020).

For CRNNs, effective feature extraction and selection are essential for achieving optimal performance. While these networks naturally utilize convolutional layers to extract spatial (spectral) features and recurrent layers to capture temporal dependencies, incorporating unsupervised techniques can enhance the model's ability to identify the most relevant features from raw data, especially in diverse and complex sound environments.

Unsupervised methods like auto encoders, clustering algorithms, or techniques such as PCA (Principal Component Analysis) can be employed to pre-process the data, reducing its dimensionality while preserving the key features that are most significant for classification tasks. These approaches can help pinpoint and retain the most discriminative features, potentially making the CRNN more resilient to variations within the dataset.

The process of feature selection in CRNNs is indeed influenced by the environment and the types of sounds being localized. For instance, in noisy or reverberant settings, spectral features (like Mel-frequency cepstral coefficients, or MFCCs) may be vital, whereas temporal features might be more important for capturing the dynamics of transient sounds. The specific environment and sound type will dictate which features are the most informative, and CRNNs can adjust to these variations by learning the most relevant features during training.

3.4 The Azimuth Sign Ambiguity in Binaural Approaches: Biological and Physical Perspectives

The azimuth sign ambiguity presents a notable challenge in binaural sound localization systems. This problem arises from the symmetry of ITDs for sound sources that are symmetrically placed along the azimuth axis, making it tough to tell whether sounds are coming from the front or back when using just two microphones (Willert et al., 2006). To tackle this issue, several strategies have been developed:

3.4.1 Biological Inspiration

Human auditory systems manage this ambiguity through various mechanisms:

Pinna shape: The outer ear filters sound differently depending on its angle of arrival, offering spectral cues that assist in distinguishing between front and back sources (Musicant & Butler,1984).

Head-Related Transfer Function (HRTF): This concept explains how sound is modified by the head, torso, and outer ear before it reaches the eardrum, providing essential spectral cues for localization (Denk, F., 2020).

Head movements: Slight shifts of the head can change the received signals, helping to clarify ambiguities (McAnally, & Martin, 2014).

3.4.1.1 Technical Solutions

Binaural localization systems can employ different techniques to reduce azimuth sign ambiguity:

HRTF modelling: Integrating HRTF models into the system can mimic how human ears filter and process sound, delivering vital spectral cues for resolving ambiguities (Xie, .2013).

Directional microphones: Utilizing directional microphones can assist in distinguishing sound source orientation based on intensity and directionality (Reid, 2017).

Dynamic movement: Introducing controlled movement of the microphone array can modify the received signals, helping to clarify the direction of incoming sounds(Grondin & Michaud, 2019).

Multi-modal approaches: Merging auditory data with visual information or data from additional sensors like gyroscopes can yield more accurate spatial orientation information(Chen.et al, 2021).

3.4.2 Advanced Techniques

Recent studies have investigated more advanced methods to enhance binaural sound localization:

Deep learning: Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have shown potential in localizing sounds in noisy environments by learning to extract relevant features(Atito et al,2024; Pang. et al,2019).

3.5 Adaptive Resonance Theory (ART)

In pursuit of optimizing computational efficiency and focusing calculations on sound events in this study, the implementation of adaptive resonance theory (ART) can be an efficient approach. The term "resonance" in the context of ART refers to the process by which the network adjusts its learning based on incoming stimuli. Specifically, resonance occurs when the input pattern matches an existing category in the network closely enough to activate that category's representation.

ART, a cognitive and neural theory, elucidates how the brain autonomously learns to attend, categorize, recognize, and predict objects and events within an ever-changing environment (Grossberg, 2013; Carpenter et al., 2012). This theory emphasizes

resonance as a regulator of learning within neural networks, transcending its role as a mere architectural framework (Carpenter et al., 2016). ART offers a considerable pathway to achieve the objectives of designing a dynamic arrangement for the spiking neural network (SNN) sound classifier and localizations in this research.

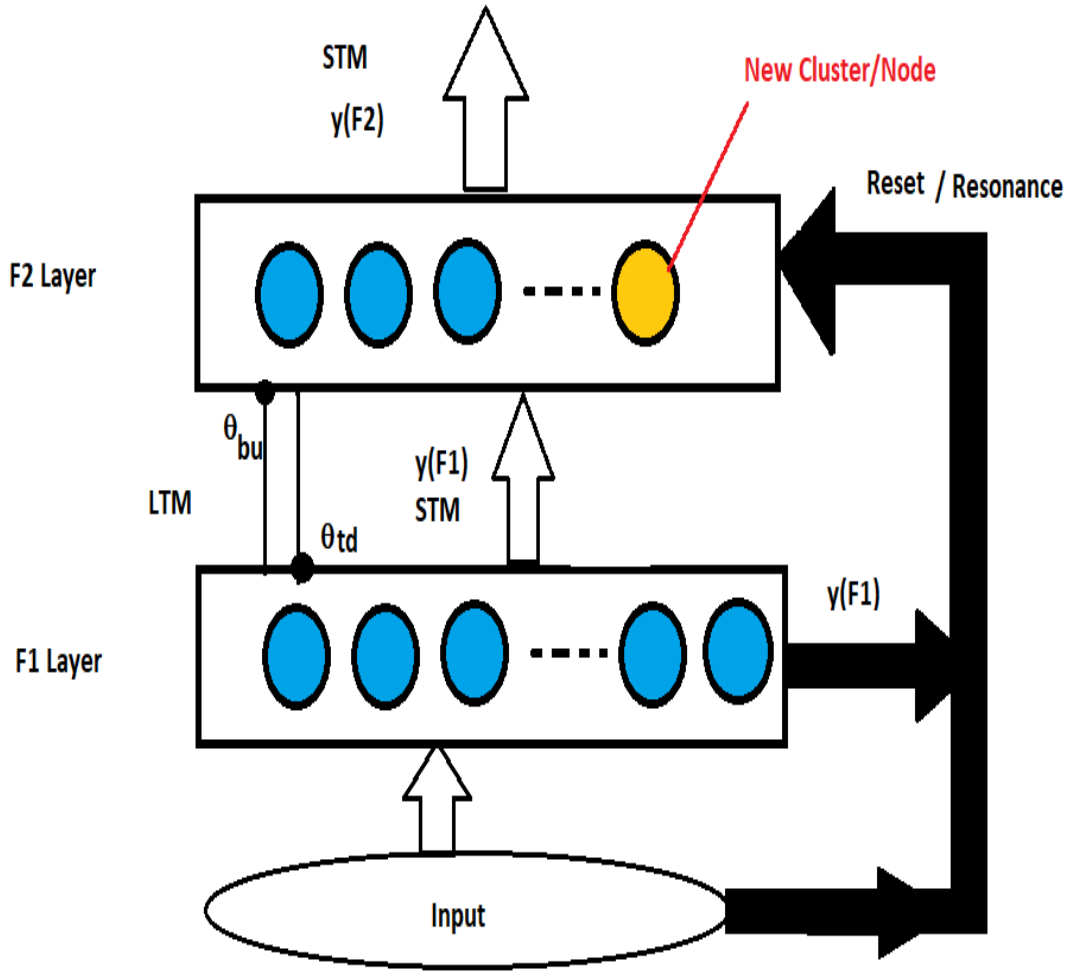
One key feature of ART is its ability to dynamically create recognition categories for encoding distinct input samples. This allows an ART module to self-adjust the scale of its recognition field, which refers to the number of committed nodes, based on the complexity of the problem domain. The fast commitment mechanism and moderate learning speed of ART contribute to its high efficiency (Grossberg, 2013).

The scale of the ART recognition field, specifically the number of output clusters, depends on a global threshold parameter known as vigilance. Fine-tuning the vigilance parameter can control the recognition representation of ART. However, suggesting an appropriate vigilance value in practice requires prior knowledge of the scale and distribution of the problem dataset, which may not be readily available (Grossberg, 2013).

ART has inspired various neural network architectures that possess attractive properties for applications in science and engineering, such as speed, configurability, explainability, parallelization, and hardware implementation. However, implementing ART models efficiently on embedded hardware presents challenges due to their computational and memory requirements. The elementary ART models are predominantly used for unsupervised learning applications. They also lay the foundation to build complex ART-based systems capable of performing all three machine learning modalities. An elementary ART neural network model (Figure 4) usually consists of two fully connected layers as well as a system responsible for its decision-making capabilities.

Figure 4

Elementary ART Model Structure



The orienting subsystem uses the vigilance threshold to regulate whether ART can go into resonance or if it must be reset.

For clarity, Table 2 summarizes the common notation used in the following subsections.

Table 2

Some Important Notations Used in this Section

Notation	*Description*
x	input sample according to the clustering problem
d	original data dimensionality
F_1	feature representation field
F_2	category representation field
N	number of categories
$y(F_1)$	F1 output
$y(F_2)$	F2 output
c	a category
θ	category parameter
M	match function
J	chosen category index
ρ	vigilance parameter
V_R	vigilance region
LTM	long-term memory
STM	short-term memory

Feature representation field F1: This is the input layer. In feedforward mode, the output y(F1) of this layer, or short-term memory (STM), simply propagates the input samples $x \in R^d$ to the F2 layer via the bottom-up long-term memory units (LTMs) _ θ^{bu} . In feedback mode, the F1 layer works as a comparator, in which x and the F2's expectation (in the form of a top-down LTM _td) are compared and the outcome y(F1) is sent to the orienting subsystem. Hence, F1 is also known as the comparison layer, and θ is category parameter.

Category representation field F2: This layer yields the network output y(F2) (STM). It is also known as the recognition or competitive layer. Neurons, prototypes, categories and templates are used interchangeably when referring to the F2 nodes. The LTM associated with a category j is $\theta = \left\{\theta_j^{bu}, \theta_j^{td}\right\}$, j = 1, ..., N. Note that not all elementary ART models discussed in this survey have independent bottom-up and top-down LTM parts; however, θ is always used to indicate the LTM (or set of adaptive parameters) of a given category.

Orienting subsystem: This is a system that regulates both the search and learning mechanisms by inhibiting or allowing categories to resonate.

Researchers have developed different types of ART learning algorithms. In the next section we briefly introduce hypersphere ART method as a kind of fuzzy ART.

3.5.1 Hypersphere ART

The hypersphere ART architecture was designed as a successor for fuzzy ART that inherits fuzzy ART's advantageous qualities while utilizing fewer categories and having a more efficient internal knowledge representation.

Fuzzy ART represents the expansion of the classic model of ART by further use of fuzzy logic for the purpose of treatment of uncertainty in data in a more effective way. While in traditional ART binary membership is used - either an object belongs to a category or it does not - in Fuzzy ART, the membership in a category is represented by continuous range [0, 1]. This is a more nuanced classification process, where an object could partially belong to several categories, harking back to a far more flexible and adaptive model.

Probably the most significant advantage that comes with Fuzzy ART is the way uncertainty and noise are treated-characteristics common to most real-world data. Thanks to fuzzy logic, Fuzzy ART is able to maintain its performance in such cases of imprecise or ambiguous data, providing system robustness that traditional ART cannot match. This kind of capability makes it particularly effective in applications where data may not fit perfectly into one class or where subtle variations in the data need to be captured.

Besides, Fuzzy ART has complemented coding that enhances efficiency by coding both positive and negative features of the data. It also allows adaptation dynamically to continuously update the classification rules with each new coming datum. These features make Fuzzy ART a privileged approach to several machine learning and pattern recognition tasks, in so far as it leads to the realization of a more flexible, adaptive, and resistant model when facing any input of a complex and noisy dataset.

LTM: Each category is represented as $\theta = \{R, m\}$, in which $R_j \in R$ is the radius and $m_j \in R^d$ is the centroid of hemisphere, d+1 memory per category is utilized.

Activation: The activation function T_j for each F_2 category j is calculated as:

$$T_j = \begin{cases} \dfrac{\overline{R} - \max\left(R_j \quad , \quad \left\|x - m_j\right\|_2\right)}{\overline{R} - R_j + \alpha} & if \quad j \quad is \quad comitted \\ \dfrac{\overline{R}}{2\overline{R} + \alpha} & if \quad j \quad is \quad uncomitted \end{cases} \qquad (3\text{-}8)$$

where $\| \ \|_2$ is the L2 norm, $\alpha \in (0, +\infty)$ is the choice parameter, and $\overline{R} \in [R_{\max}, +\infty)$ is the radial extend parameter, which controls the maximum possible category size achieved during training. The lower-bound R_{max} is defined as:

$$R_{\max} = \frac{1}{2} \max_{p,q} \left\|x_p - x_q\right\|_2 \qquad (3\text{-}9)$$

When a sample x is presented, a winner-takes-all (WTA) competition takes place over its categories at the output layer F2. Then, the neuron j that optimizes that model's activation (or choice) function T across the nodes is chosen (e.g., the neuron that maximizes some similarity measure to the presented sample):

$$J = \arg\max_j \quad (T_j) \qquad (3\text{-}10)$$

Match and resonance: The winning category J is selected using WTA competition and the match function is calculated as:

$$m_j = \begin{cases} 1 - \dfrac{\max\left(R_j \quad , \quad \left\|x - m_j\right\|_2\right)}{\overline{R}} & if \quad j \quad is \quad comitted \\ 1 & if \quad j \quad is \quad uncomitted \end{cases} \qquad (3\text{-}11)$$

where the vigilance criterion is $m_j > \rho$ ($\rho \in (0,1)$ is the vigilance parameter.

Learning: *If the winning category satisfies the vigilance test, then resonance occurs, and the radius* R_j *and centroid* m_j *of the winning node are updated as follows:*

$$R_j^{new} = R_j^{old} + \frac{\beta}{2}\left[\max\left(R_j^{old}, \left\|x - m_j^{old}\right\|\right) - R_j^{old}\right] \quad (3\text{-}12)$$

$$m_j^{new} = m_j^{old} + \frac{\beta}{2}\left(x - m_j^{old}\right)\left[1 - \frac{\min\left(R_j^{old}, \left\|x - m_j^{old}\right\|_2\right)}{\left\|x - m_j^{old}\right\|_2}\right] \quad (3\text{-}13)$$

where, $\beta \in [0,1)$ is the learning rate parameter. If the winning category fails the vigilance test, it is reset, and the process is repeated.

Eventually, either a category succeeds or a new one is created with its radius and centroid initialized as $R_{N+1} = 0$ and $m_{N+1} = x$, respectively.

As it is denoted in this section, this learning algorithm can be applied with the aim of clustering and classification, but in the proposed structure, we suggest using this concept to update the activated zones where the source sound is located. In other words, we have a tendency to add finite new hidden neurons on that area, to increase the precision.

3.6 Sound Energy Attenuation Model

While sound travels through the air, acoustic energy is emitted omnidirectionally from the sound source. The strength of a sound source diminishes at a rate inversely proportional to the square of the distance(Deng et al., 2017). The traditional formulation of this algorithm is given as follows(Deng et al., 2017):

$$\psi_i(t) = \zeta_i \frac{S(t)}{\left|r_i - r(t)\right|^{\alpha}} + \varepsilon_i(t) \quad (3\text{-}14)$$

where $\psi_i(t)$ indicates the signal energy measured on the i_{th} sensor, ζ_i is the gain factor of the i_{th} acoustic sensor, and S(t) is the sound energy 1 m from the sound source (we can think of this as the energy of the sound source). r_i (sensor location) and r(t) (unknown location) indicate the coordinates of i_{th} sensor node and sound source at time t. Each variable is a vector with two additional variables (when in a two-dimensional (2-D) plane).

When there are m sensor nodes, the value of α can be obtained as:

$$\alpha = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \alpha_{ij} \qquad (3\text{-}15)$$

where α_{ij} is the path loss exponent obtained by the i_{th} and j_{th} sensor nodes ($\alpha_{ij} = \alpha_{ji}$). Regarding both deterministic and metaheuristic algorithms, all observations from the multiple sensors are aggregated as an estimator of r(t), where the solution of the localization problem is the argument (pair of coordinates) that minimizes the expression

$$\hat{r}(t) = \arg \min_{r} \sum_{i=1}^{m} \frac{1}{\sigma_{\xi_i}^2} \left(\psi_i - \zeta_i \frac{S(t)}{|r_i - r(t)|^{\alpha}} \right)^2 \qquad (3\text{-}16)$$

where $\frac{1}{\sigma_{\xi_i}^2}$ is the variance of acoustic gain factor.

The estimator in expression (3-15) is highly nonconvex, with singularities in each sensor's coordinates, several suboptimal solutions, and saddle regions. All of the enumerated features make the problem very challenging in the field of numerical optimization, making it a good candidate in the context of regression and artificial neural networks (ANNs).

Approximating the location based on energy in the model of Sound Energy Attenuation consists of a nonconvex optimization problem, which is very time-consuming. Usually, this type of problem would have to be solved iteratively using such techniques as numerical optimization methods or such models as RNNs. Given this, we decided to use an rSNN because rSNNs are particularly suited to optimization problems and to tasks where the temporal dynamics of the input is important for continuous update, as in sound localization.

3.7 Spiking Neural Network

In the language of machine learning and artificial intelligence, a neural network can be defined as a network of neurons capable of performing calculations and solving problems. Neural network learning has changed a lot since its introduction in the late 1950s. The first generation of neural networks consisted of neurons that computed a weighted sum of inputs, and if the sum exceeds a predetermined threshold, the output is one. These neurons are also called perceptron.

As a model of computational dynamics inspired by the brain, ultra-fine neural networks lead to the encoding and processing of neural information through completely random paths. Spiked neural networks consist of spiking neurons, which are suitable tools for processing complex spatial and temporal information. However, due to their implicit and discontinuous nonlinear mechanisms, formulating efficient supervised learning algorithms to motivate neural networks is challenging and has become an important issue in this field (Kasabov, 2014).

SNNs are the third generation of neural networks whose level of realism has increased. In addition to having synaptic states, these networks have the ability to store temporal information simultaneously. Therefore, the neurons will not fire in each propagation loop, and their firing depends on the membrane potential reaching the threshold. SNNs are superior to non-spiking neural networks due to the ability of temporal coding in their signals, but they require different biologically acceptable rules of synaptic plasticity. In SNNs, the level of activation flow, which is described as differential equations, is considered as neuron states. Like previous common neural networks, SNNs are used for information processing. The spiking neuron model that is utilized in this work is defined below.

3.7.1 Spiking Neuron Model

This study utilizes a leaky integrate-and-fire neuron. The membrane potential V evolves according to the equation (Dutta et al., 2017):

$$\frac{dV_i(t)}{dt} = \frac{1}{\tau_m}\left(-V_i(t) + I_i(t)\right) \qquad (3\text{-}17)$$

where τ_m denotes the membrane decaying time constant and $I_i(t)$ is the synaptic current.

3.8 Encoding Information for Neural Computing

A fundamental question is discussed in this section. Specifically, considering the codes used by neurons to transmit information, how is input data encoded in neural networks? Traditionally, there are two main theories of neural coding – pulse codes and rate codes – and we will provide explanations about both theories.

3.8.1 Pulse Encoding

The first type of neural encoding is neural firing or pulse encoding. In this cryptography, the precise timing of neural firing is assumed as the carrier of information between neurons. Evidence for temporal correlation between neuronal firings has been shown through computer simulations using the (I&F) neuron model (Legenstein, 2005), biological experiments such as electrophysiological wave recording and colorimetric methods (Nawrot, 2009). As another approach, vivo measurements has been shown in the space-time patterns, utilized neural activities of the behavioural response of rats (Villa, 1999).

A pulse code based on the timing of the first shot following a reference signal is discussed by (Thorpe & Trehub, 1989). This encoding is known as first-shot time. Each cell can emit only a small number of nerve impulses that can contribute to the whole process of a stimulation. Additionally, a new stimulus is processed between 20 and 50 milliseconds (ms) after its onset. Therefore, early neural firings carry the most information that is hidden in the stimulation (Thorpe, 2001).

Other pulse codes consider correlation and similarity as important parameters. Neurons that show similar function, target, or classification are marked by identical firing (Von Der Malsburg, 1994). In general, each precise spatio-temporal pulse sample is potentially important and may encode specific information. Neurons can fire with a specific relative time delay that can indicate a specific stimulus.

3.8.2 Rate Encoding

In rate encoding theory, it is assumed that the average output rate of a cell carries most, but not all, of the transmitted information. These codes are called rate codes and they are derived from classical perceptron methods. The average firing rate v is usually considered as the ratio of the average number of neuronal firings (spikes) n_{sp} observed in a given time interval T.

$$v = \frac{n_{sp}}{T} \qquad (3\text{-}18)$$

This concept has been particularly successful in the field of nervous and motor systems. Since each neuron must integrate the firing activity of the presynaptic neurons at least

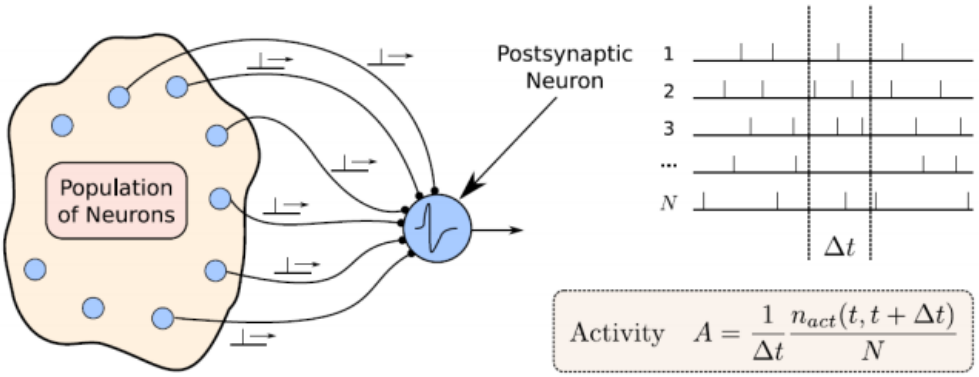
in the time interval T, the main issue is the comparable slow transmission of information from one neuron to another. In particular, the very short temporal responses of the brain to a particular stimulus cannot be expressed by a temporal averaging of neural firings. For example, Thorpe et al. (1996) report that the human brain could recognize a visual stimulus in approximately 150 ms. Since the appropriate number of neural layers are faced with the process of visual stimulation, if each layer is forced to wait for a period of time T to receive information from the previous layer, the detection time will be greatly increased. In other words, the average neural firing rate is defined as the average neural activity in a population of neurons, whose rules are shown in Figure 5. A postsynaptic neuron receives excitatory inputs in the form of output neural firings from a population of presynaptic neurons. This population produces a neuronal firing activity A, defined as the number of neurons that were active in a short time interval *[t, t+Δt]* divided by the population size N and the time period Δt (Gerstner, 2002).

$$A = \frac{1}{\Delta t} \times \frac{n_{act}(t.t+\Delta t)}{N} \qquad (3\text{-}19)$$

A neuron receives impulses from a population of presynaptic neurons that produce a certain (level of) activity (A). Figure 5 shows how the postsynaptic neuron is stimulated due to the stimulation of presynaptic neurons.

Figure 5

Post Synaptic Neurons



where n_{act} *(t, t+Δt)* is the number of active neurons in the time interval *[t, t+Δt]* and *N* is the total number of neurons in the population. The activity of a population may change very quickly, which allows neurons to respond quickly to changes in stimulation (Gerstner, 2000).

3.9 Learning Rules

In this section, some common methods of learning rules related to SNNs are introduced in the field of firing neurons. Various problems have hampered the development of learning methods for SNNs. The precise time dependence (reliance) leads to the heterogeneity of the input information, which often requires the use of complex hardware or software before the neural network can operate. The geometry of recurrent networks that are commonly used in SNNs prevents the creation of a simple learning method such as the error backpropagation method in multilayer perceptron. Similar to traditional neural networks, three different learning models can be distinguished in SNN, which are unsupervised, reinforcement, and supervised learning methods.

Reinforcement learning methods are less common in SNNs than other methods. Some algorithms have been successful in the field of robotics (Florian, 2005) and these methods have also been analysed theoretically in some studies (Seung, 2003; Florian, 2007).

Unsupervised learning in the form of Hebbin learning is the closest method to the real biological learning method (Cooper, 2005).

The term synaptic time dependent plasticity (STDP) belongs to this category. Supervised methods impose a specific input-output mapping on the network, which is essential for practical SNN applications. STDP is described next.

3.9.1 Synaptic Time Dependent Plasticity

Learning law (STDP) uses Hebbian plasticity (unsupervised) in the form of long-term potentiation and depression (LTP & LTD). The synaptic effect is strengthened or weakened based on the timing of the postsynaptic action potential in relation to the presynaptic nerve firing. Figure 6 indicates both mentioned learning laws.

Figure 6

Synaptic Changes in a Neuron by STDP (Egger, 1999)a) Hebbian Learning b)LTP&LTD

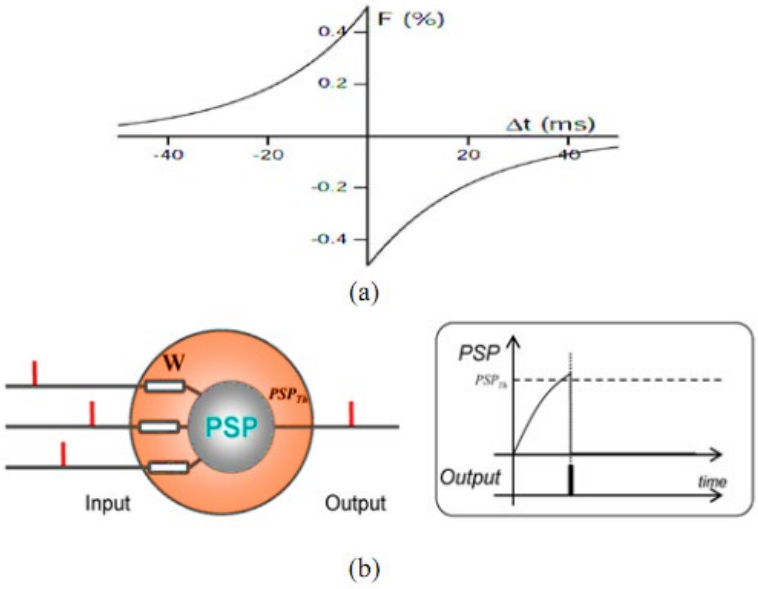


Figure 6 describes the rule behind STDP, highlighting its potentiation and attenuation procedures. STDP is a learning rule that changes the strength of the connection depending on the timing difference in spikes between a pre-synaptic neuron—source and a post-synaptic neuron—target.

Figure 6 illustrates potentiation; more specifically, it shows that when a presynaptic spike is just before a postsynaptic spike, the synaptic connection is potentiated.

The strong response of the postsynaptic neuron immediately after receiving the input illustrates a strong relationship between the two neurons. In contrast, attenuation on the right side means that if the postsynaptic spike occurs much earlier or much later than the presynaptic spike, then the synaptic connection weakens. This implies a weaker or less important relationship in their firing patterns. The timing window for such changes is very critical: the small timing differences induce either potentiation or attenuation of synaptic strength. This temporal sensitivity in SNNs allows adaptation and learning in a complex temporal pattern, just as biological learning processes in the brain do. This dynamic change of connections due to spike timing in STDP adds up to the ability of the network to learn from input data in a biologically plausible manner.

STDP is described by the function $W(t_{pre} - t_{post})$, which describes the partial changes of synaptic weights as the difference between the arrival time (t_{pre}) of a presynaptic

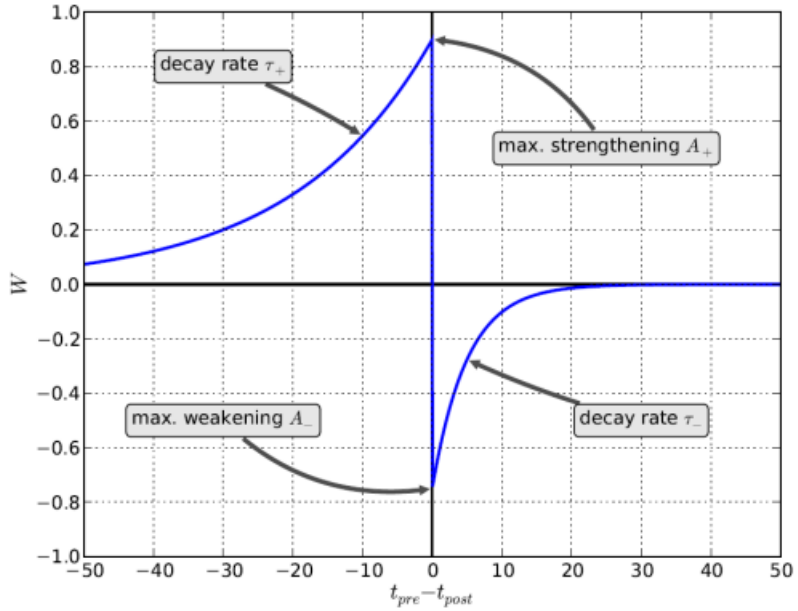
neuron firing and the time of an activity emitted from a neuron (t_{post}). Usually, W is expressed as follows:

$$W(t_{pre} - t_{post}) = \begin{cases} A_+ \, exp\left(\frac{t_{pre}-t_{post}}{\tau_+}\right) if \; t_{pre} < t_{post} \\ A_- \, exp\left(-\frac{t_{pre}-t_{post}}{\tau_-}\right) if \; t_{pre} > t_{post} \end{cases} \quad (3\text{-}20)$$

where τ_+ and τ_- determine the time interval between pre-synaptic and post-synaptic and A_+ and A_- denote maximum partial variation in small $|t_{pre} - t_{post}|$. According to Eq.3-20, learning synaptic weight (W) window is shown in figure 7.

Figure 7

Synaptic Weight Learning Law by STDP (Egger, 1999)



More information about STDP can be found in Lim (2021) and Aljadeff (2021).

As well as unsupervised learning strategy, there are supervised methods. Spike-back propagation (spike-prop) a well-known supervised learning strategy, as described in the next section.

3.9.2 Spike-Prop

Traditional neural networks, such as multilayer perceptron, usually use gradient descent-based learning methods such as error backpropagation to modify synaptic weights. The results of this operation are shown as an input-output network. However, the recursive geometry of SNN networks and their obvious time dependence will not

allow a simple evaluation of the gradient in the network. For this reason, certain assumptions can be used before applying error backpropagation in SNNs.

In the error backpropagation supervised learning algorithm (Nuntalid, 2011), SNNs learn an arbitrary set of firing times of all *j* output neurons (t_j^d for a given input pattern. This method minimizes *E*, which is the sum of squared errors of firing times in output neurons from arbitrary times (t_j^{out}).

$$E = \frac{1}{2}\sum_j (t_j^{out} - t_j^d)^2 \qquad (3\text{-}21)$$

Accordingly, the synaptic weight connecting neuron *i* to neuron *j* at *k*th sample time is updated as follows:

$$\Delta W_{ij}^k = -\eta \frac{dE}{dw_{ij}^k} \qquad (3\text{-}22)$$

where η is the learning rate of updating in each step. The limitation of this method is that each neuron is allowed to fire only once. Due to the dependence of this method on the difference between the actual firing time and the desired firing time, it will only be suitable for the time-to-first-spike encoding method.

3.10 Summary

In this section, we have examined the essential theoretical groundwork vital for our thesis. We conducted a concise exploration of the central concepts encompassing sound source localization and sound source classification, particularly through deep learning techniques. We delineated the fundamental stages involved in these approaches, elucidating their overarching principles.

Additionally, we introduced adaptive resonance theory (ART) and prominent methodologies, shedding light on ART's significance in our research. Furthermore, we delved into the intricacies of spiking neural networks, focusing on spike encoding and the development of learning rules, which are pivotal components of our research methodology. This section lays the foundation for the subsequent exploration and implementation of these concepts in our research.

Worth mentioning, the ART clustering approach is related to the spatial features of neurons. This fact enhances the capability of the network, as it grows and concentrates its computation costs on locations we estimate the sound source is at that position.

4. Recurrent Neural Networks in Sound Source Localization and Classification

4.1 Introduction

Recurrent neural networks (RNNs) play a pivotal role in sound source localization and classification. RNNs are a class of neural networks widely recognized for their adeptness in handling sequential data and temporal associations. This chapter examines the importance of RNNs in the realm of sound classification and localization, shedding light on their indispensable role.

Within the framework of this thesis, we implement a specific RNN structure known as the recurrent spiking neural network (rSNN). Understanding this structure is pivotal as it constitutes a subtype of RNNs that forms the foundational core of our model for sound source localization and classification. Hence, this chapter examines both the role of RNNs and the primary phases and intricacies of the rSNN structure, establishing a comprehensive basis for our research.

RNNs are a category of neural networks that have been devised to manage sequential and temporal data (Grossberg et al., 2013; DiPietro et al., 2020; Salehinejad et al., 2017). Unlike feedforward neural networks, RNNs possess connections that spiral back on themselves, allowing for the maintenance of a concealed state or memory of prior inputs. This capability to capture temporal interdependencies means that RNNs are well-suited for tasks related to sequential data, such as audio processing for sound source localization and classification (Grossberg et al., 2013; DiPietro et al., 2020; Salehinejad et al., 2017).

The conceptual constituents of RNNs encompass a concealed state (memory), recurrent connections, and training with backpropagation through time (BPTT) (Lillicrap & Santoro, 2019; Grossberg et al., 2013; DiPietro et al., 2020; Salehinejad et al., 2017). The concealed state captures information pertaining to past inputs and functions as a form of memory, allowing the network to consider a previous context when processing the current input. Recurrent connections are an essential tool for keeping the network updated with information from past time steps and enabling it to update its concealed state based on the current input. Training often involves BPTT, which is an extension of

the standard backpropagation algorithm adapted for the sequential nature of RNNs (Lillicrap & Santoro, 2019; Grossberg et al., 2013; DiPietro et al., 2020; Salehinejad et al., 2017).

RNNs are employed in diverse applications, such as speech recognition, natural language processing, and stock market prediction (Grossberg et al., 2013; DiPietro et al., 2020). They are particularly potent when context is crucial to predicting an outcome, and they are distinguishable from other types of synthetic neural networks due to their employment of feedback loops to process a sequence of data that informs the ultimate output (Lillicrap & Santoro, 2019; Grossberg et al., 2013). The capability of RNNs to manage sequential data and capture temporal dependencies renders them a potent tool for addressing a broad spectrum of problems.

While recurrent neural networks are potent tools for managing sequential data, they do possess certain constraints when they are used to process noisy or unfinished data (Arbel, 2018; Salehinejad, 2017). One of the primary limitations is the vanishing gradient predicament, which arises when the gradients utilized for updating the network's weights become too trivial to be efficacious, culminating in sluggish or ineffectual training (Zhang et al., 2018; Arbel, 2018). Another constraint is their arduousness in dealing with longer sequences owing to the limitations of short-term memory (Ma et al., 2021). Moreover, RNNs may encounter difficulties in coping with noisy or incomplete data, which can impede their performance (Orta Alemán & Horne, 2020).

However, there are methods that can be employed to alleviate these limitations, such as injecting noise into concealed states during training to enhance resilience to noisy data (Orta Alemán & Horne, 2020). Other approaches comprise utilizing long short-term memory (LSTM) networks, which are a variety of RNNs that manage long-term dependencies and mitigate the vanishing gradient problem (Elsaraiti & Merabet, 2021).

4.2 Theoretical Component of RNNs

Recently, reservoir computing, which is a category of recurrent spiking networks, has gained considerable attention due to its ability to carry out intricate computations while maintaining low computational expenses and power consumption (Hamedani, 2020). In this type of computing, an untrained stochastic recurrent network, known as the

reservoir, is utilized, and a simple classification or regression approach is employed to process the output of the reservoir (Schrauwen et al., 2007).

The reservoir functions as a high-dimensional dynamic system and can be achieved using various neuron types, such as spiking neurons, rate neurons, or binary neurons (Lukoševičius & Jaeger, 2009). Typically, the reservoir is constructed with a large quantity of neurons to allow it to capture complex temporal patterns and dynamics within the input data. The input data are fed into the reservoir, where the neurons engage with one another through synaptic connections. Afterwards, the readout layer processes the output from the reservoir, acting as a simple linear or nonlinear classifier/regressor and mapping the output of the reservoir to the desired output (Melandri, 2004).

Reservoir computing has demonstrated its effectiveness in various tasks such as speech recognition, image classification, and time series prediction (Hamedani, 2020). Moreover, it has demonstrated energy efficiency, making it a promising approach for implementing low-power machine learning systems (Reynolds, 2019).

The reservoir can be trained through different techniques such as Hebbian learning, spike-time-dependent plasticity (STDP), and intrinsic plasticity (IP) which improves its capacity to distinguish speech data from real-world and enables neurons to transmit information more efficiently (Lukoševičius& Jaeger,2009). Additionally, the reservoir can be combined with other recurrent neural network architectures such as long short-term memory (LSTM) and gated recurrent unit (GRU) to form more complex networks. These structures aid the reservoir in capturing long-term dependencies and temporal patterns in the input data (Kholkin et al.,2023).

Figure 8 depicts different structures of RNNs – standard RNN, long short-term memory (LSTM), and gated recurrent unit (GRU) – as per Wei et al. (2021).

Figure 8

Overall Architecture of RNNs Units

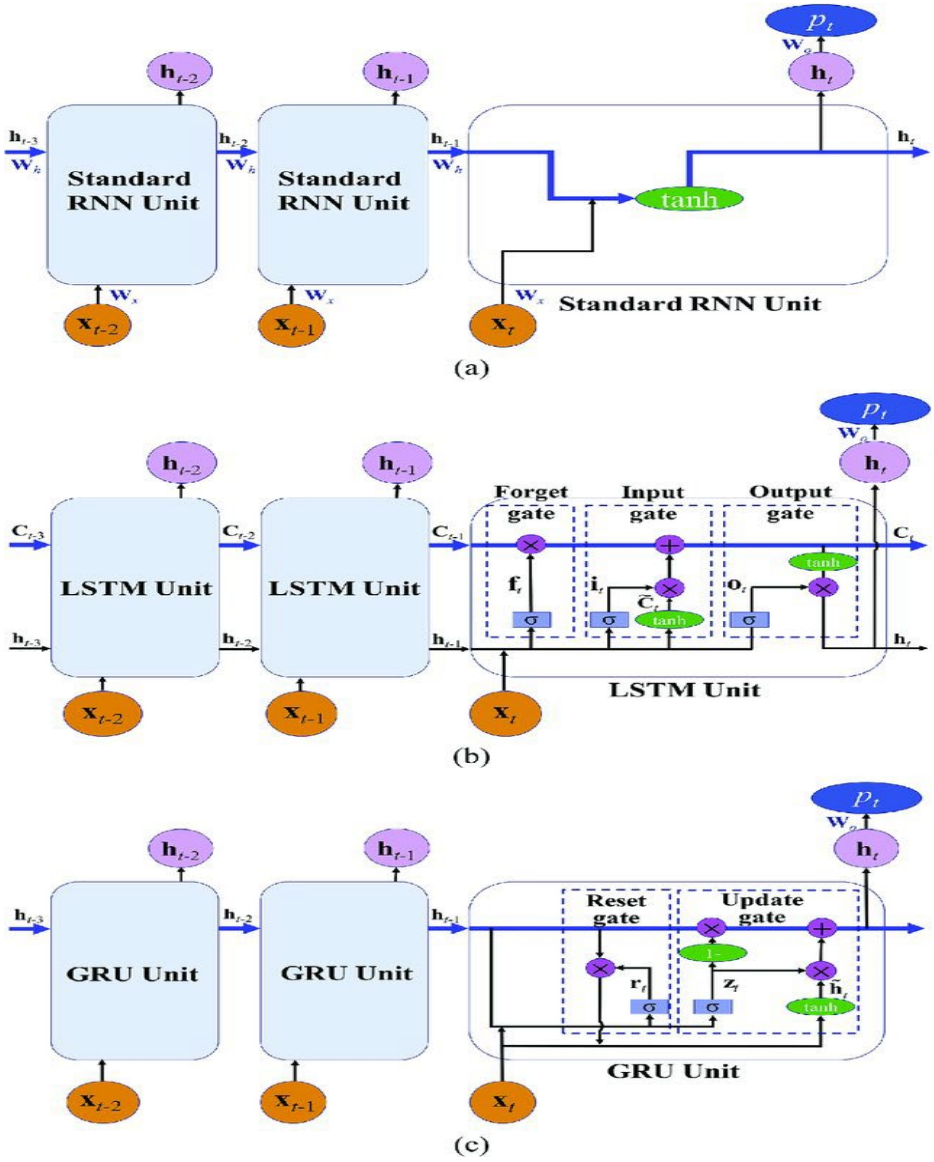


Figure 8 depicts a detailed review of different RNN Units architectures in use for processing sequences, stating their respective strengths and weaknesses in complex environments.

Traditional RNNs show excellent performance in relatively simple sequential tasks, thanks to their ability to maintain connections across different time steps (Elman, 1990).

However, they suffer from issues related to the vanishing gradient phenomenon, which limits their effectiveness in learning long-term dependencies (Hochreiter, 1991).

LSTMs deal with the vanishing gradient problem effectively, allowing the network to preserve long-term dependencies necessary for complex tasks involving time (Hochreiter & Schmidhuber, 1997). However, increased complexity means greater computational resources and longer training times are required (Greff et al., 2017).

Gated Recurrent Units (GRUs) are a mid-point between LSTMs and standard RNNs in that they retain many of the benefits of LSTMs but with simpler architecture, resulting in faster training times and lower computational costs, which proves helpful in situations where the model needs to be able to react quickly to new inputs (Cho et al., 2014).

SNNs are characterized by high efficiency, contributed to by their event-driven properties that allow information processing only in response to spike occurrences. This feature makes them especially appropriate for applications that require high temporal precision and saving of energy (Maass, 1997). Training and deploying SNNs, however, is not straightforward because they are inherently non-differentiable (Lee et al., 2016).

Adaptive Resonance Theory (ART) is extremely effective at learning in dynamic environments, elegantly incorporating new information without sacrificing any previously learned data (Carpenter & Grossberg, 1987). However, the power of ART comes at the price of its elaborate mechanisms and requirement for careful parameter tuning.

The proposed approach combines the strengths of ART and SNNs in an attempt to design an adaptive, computationally efficient model capable of addressing complex real-time environments. This hybrid model gives better results than traditional RNN-based approaches concerning adaptability and computational cost, making it very suitable for

application in dynamic and resource-constrained environments (Chung et al., 2014; Tavanaei et al., 2019; Carpenter & Grossberg, 1991).

4.3 Recurrent Spiking Neural Networks

Recurrent spiking networks is a class of brain-inspired recurrent algorithms that reduce the computational complexity and cost of training machine learning models (Soures & Kudithipudi, 2019). These networks are inspired by the way neurons in the brain communicate with each other through electrical impulses or spikes. In this section, we introduce recurrent spiking networks and then focus on reservoir computing, which is a type of recurrent spiking network.

Reservoir computing has become increasingly popular in recent times, owing to its ability to conduct intricate computations with minimal computational cost and power consumption (Hamedani, 2020). In reservoir computing, an untrained random recurrent network, commonly known as the reservoir, is employed and a straightforward classification/regression method is utilized to handle the output of the reservoir (Schrauwen et al., 2007).

The reservoir is a high-dimensional dynamical system that can be actualized utilizing various forms of neurons, including spiking neurons, rate neurons, or binary neurons (Lukoševičius & Jaeger, 2009). The reservoir is typically structured to possess a substantial number of neurons, enabling it to capture sophisticated temporal dynamics and patterns in the input data.

The input data are introduced into the reservoir, and the neurons within the reservoir interact with each other through synaptic connections. Subsequently, a readout layer processes the output of the reservoir, which functions as a simple linear or nonlinear classifier/regressor that maps the output of the reservoir to the desired output (Melandri, 2004).

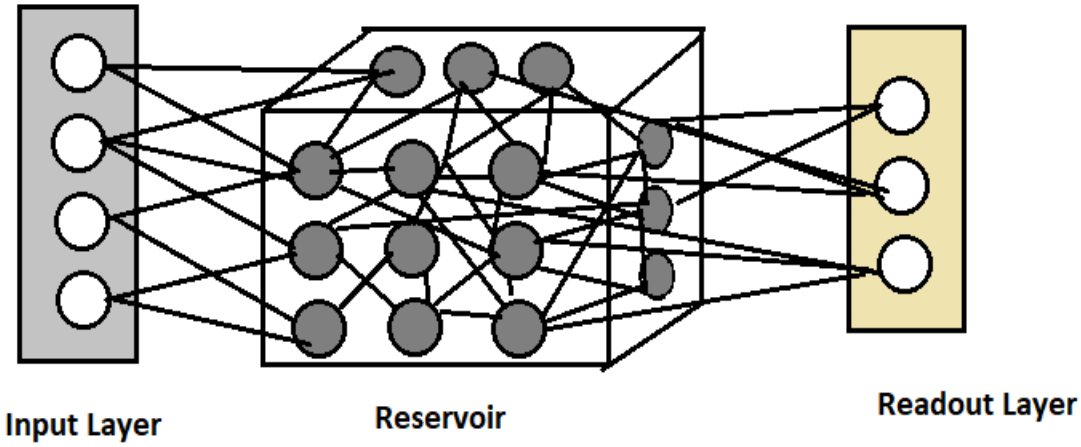
Reservoir computing has been successfully employed in various fields, including speech recognition, image classification, and time-series prediction (Hamedani, 2020).

Moreover, it has been demonstrated to be energy efficient, which makes it a promising method for implementing low-power machine learning systems (Reynolds, 2019).

The general structure of reservoir calculations is shown in Figure 9. In this structure, an input signal enters a static dynamic system and transforms the dynamics of the input reservoir to a higher state (higher dimension). A direct readout function is then trained to convert the state response (higher dimensions) to the desired output. The main advantage is that the learning approach takes place only in the reading phase. The two main types of reservoir computing are liquid state machines (LSM) and echo state networks. Non-correlation backpropagation methods and temporal recurrent networks are also other types of reservoir computing.

Figure 9

Reservoir Network Structure



The theoretical results of liquid state machines are quite general and are formulated with mathematical frameworks from dynamical systems theory and filtering theory. These theories are also applied to echo-mode networks, but since the primary goal of developing LSMs is to provide an acceptable biological model for general computation in the microcircuits of brain neurons – due to its application to circuits composed of firing neurons with internal noise – many biological characteristics have been explored. In contrast, echo mode networks are used for engineering applications with well

performance in different tasks and no internal noise (Jaeger, 2007). Due to the nature of audio signals that are accompanied by noise, the liquid state machine method has been used in this thesis.

In addition, while traditional engineering approaches like echo-based methods have their merits—especially in controlled environments with predictable conditions—biologically realistic networks like Liquid State Machines (LSMs) offer significant advantages in terms of adaptability, robustness, temporal processing capabilities, and computational efficiency. These qualities make LSMs particularly well-suited for complex tasks such as sound source localization in dynamic and noisy environments. Incorporating this understanding into your thesis will provide a comprehensive perspective on the strengths of biologically inspired models compared to conventional engineering solutions.

4.4 Neuron Arrangement in Reservoir Spiking Neural Networks

Reservoir spiking neural networks entail a significant paradigm referred to as reservoir computing, which falls under the purview of recurrent neural networks. Within reservoir computing, spiking neural networks (SNNs) have emerged as a highly promising approach. The configuration of neurons within a reservoir SNN plays a pivotal role in determining its performance, functionality, and computational efficiency. The optimization of the neuron arrangement and topology constitutes a fundamental objective of this doctoral thesis, aiming to enable the effective reception and processing of spatio-temporal input data.

Furthermore, computational cost is important in real time application, especially in SECL, so reducing unnecessary computational costs can speed up the process without losing accuracy. In the other words, optimized connectivity in neural networks significantly enhances sound event localization and classification by improving information flow and feature extraction while increasing robustness against noise. By streamlining connections, the network can efficiently process auditory signals, prioritize relevant features, and filter out background noise, which is crucial in dynamic environments with overlapping sounds. This adaptability allows for accurate identification and localization of sound events, even in challenging conditions. Furthermore, optimized connectivity supports scalability and generalization, enabling

the model to maintain performance across different environments and new sound types without extensive retraining. Overall, these advantages make optimized connectivity essential for effective sound localization tasks.

4.4.1 Strategies for Neuron Arrangement

Neuron arrangement and, consequently, network architecture can significantly impact the efficiency of a neural network. In this context, various techniques have been introduced. We will provide a brief review of some well-known approaches in the following sections.

Stochastic connectivity: A distinctive characteristic of reservoir SNNs is the utilization of stochastic connectivity within the reservoir. In this approach, neurons are randomly interconnected, thereby engendering a network that exhibits chaotic dynamics. The inherent randomness inherent in this connectivity fosters the exploration of a wide array of dynamic behaviours by the reservoir, thereby augmenting its capacity to capture intricate spatio-temporal patterns. Additionally, the adoption of random connectivity simplifies network initialization, obviating the need for reservoir training (Kholkin et al., 2023; Tanaka et al., 2019). In this approach, neurons are randomly interconnected, resulting in a connectivity matrix, W, with random connection strengths. This can be mathematically represented as:

$$w_{ij} \sim \text{Uniform}(a,b) \qquad \textit{(4-1)}$$

where w_{ij} represents the connection strength between neuron *i* and neuron *j and* $Uniform(a,b)$ is a uniform distribution with parameters *a* and *b*, representing the range of possible connection strengths.

Optimized connectivity: Despite the potency of stochastic connectivity, researchers have embarked on an exploration of optimized connectivity structures to bolster reservoir performance. Techniques such as echo state networks (ESN) employ optimized connections, frequently generated through methodologies like spectral radius scaling and weight regularization. The spectral radius scaling and weight regularization methods can be mathematically represented as:

Spectral radius scaling

$$W_{\text{scaled}} = \frac{W}{\rho_{\max}} \quad (4\text{-}2)$$

where W_{scaled} is the scaled connectivity matrix, W is the original connectivity matrix, and $\rho_{\max}$ is the desired maximum spectral radius (usually less than 1 for stability).

Weight regularization

$$L_{\text{total}} = L_{\text{task}} + \lambda \sum_{i,j} w_{ij}^2 \quad (4\text{-}3)$$

where L_{total} is the total loss function, L_{task} is the task-specific loss, λ is the regularization strength, and $\sum_{i,j} w_{ij}^2$ represents the sum of squared weights.

These methods fine-tune the reservoir's behaviour for specific tasks, aiming for an optimal arrangement of neurons and network topology (Kholkin et al., 2023; Wang & Yan, 2015).

The integration of optimized connectivity enables fine tuning of the reservoir's behaviour, rendering it more amenable to specific tasks. The overarching objective of this optimization lies in the identification of the most efficacious arrangement of neurons and network topology to facilitate the efficient processing of spatiotemporal data (Kholkin et al., 2023; Wang & Yan, 2015).

4.5 Architecture and Topology Effects

This section focuses on various topology methods that significantly influence the network's behaviour and efficiency. We will explore the impact of different architectural configurations, highlighting key insights into how the topology of a neural network can be harnessed to optimize its functionality and enhance overall performance. To this aim, we express critical aspects influencing the performance of reservoir spiking neural networks (SNNs), focusing on the nuanced considerations of neuronal classifications, spatial configurations, and stratified architectures.

Neuronal classifications: The selection of neuronal classifications within the reservoir has a profound impact on its performance. Spiking neurons, rate neurons, and binary

neurons each exhibit distinct computational characteristics. The choice of neuronal classifications must align with the specific task the reservoir is designed for, particularly when processing spatio-temporal input data. Diligent consideration of neuronal classifications can optimize the effectiveness of the network (Gaurav et al., 2023; Kholkin et al., 2023).

Spatial configurations: The spatial configuration of neurons within the reservoir can also influence its performance. Small-world and scale-free topologies, inspired by intricate networks, have been examined(Kholkin et al., 2023; Rathi et al., 2023) to augment the network's capability to capture global and local spatiotemporal patterns concurrently. These spatial configurations can result in improved effectiveness by leveraging the network structure, especially in the context of spatio-temporal data (Kholkin et al., 2023; Rathi et al., 2023).

Stratified architectures: Some reservoir SNNs employ stratified architectures where neurons are organized into multiple hierarchical layers. This arrangement allows for the extraction of hierarchical spatiotemporal features in the input data, rendering it suitable for tasks such as hierarchical pattern recognition in intricate data streams. Stratified architectures can amplify both performance and effectiveness by capturing intricate relationships within spatio-temporal data (Kholkin et al., 2023).

In the next chapter of the thesis, we put forward a novel methodology to optimize the effectiveness of a reservoir SNN for the localization of sound sources. In contrast to the strategy of random connectivity, we eradicate redundant connections and incorporate neurons into the framework at the estimated position of the sound. This methodology amplifies the focus on specific spatial locations, enhancing precision in the analysis of spatio-temporal data.

We also employ an adaptable topology to amplify the significance of the region from which the sound is presumed to emanate and diminish computations in other regions. The selection of neuron classifications and the spatial configuration of neurons within the reservoir will be deliberated, in order to optimize the performance of the network.

Additionally, we exploit small-world and scale-free topologies to concurrently capture global and local spatio-temporal patterns. Furthermore, a stratified architecture will be

employed to extract hierarchical spatio-temporal characteristics from the input data, thereby rendering it suitable for the localization of sound sources in intricate acoustic environments.

The ultimate aim of this investigation is to formulate a reservoir SNN that can proficiently process spatio-temporal input data from acoustic signals for sound source localization, thus paving the way for innovative applications in fields such as audio processing and classifications.

4.6 Summary

In this chapter, we explained the structure and theoretical components of recurrent neural networks (RNNs) and highlighted the growing popularity of reservoir computing, a subtype of recurrent spiking networks. This approach has gained prominence due to its capability to perform complex computations while maintaining low computational cost and energy consumption (Hamedani, 2020).

Our primary objective in this endeavor is to create a real-time sound localization and classification system. To achieve this, we harness the power of reservoir spiking neural networks (SNNs), which offer minimal computation time while maintaining acceptable accuracy levels for real-time applications.

Recognizing that the arrangement, topology, and connections within the RNN profoundly impact its effectiveness, the next chapter focuses on an innovative approach. We intend to integrate adaptive resonance theory (ART) with reservoir SNNs. This integration will enhance the performance of the reservoir SNN and optimize computational costs and neuron arrangements by leveraging ART's capabilities.

In conclusion, our research is at the forefront of combining cutting-edge neural network structures to advance the field of sound localization and classification, with a keen focus on real-time applications and efficiency.

5. Dynamic Structured Recurrent Spiking Neural Network Design

5.1 Introduction

In the previous chapter, we examined the intricacies of RNNs and highlighted the exceptional capabilities of reservoir spiking neural networks (rSNNs) in precise modeling, classification, and localization of sound sources. One of the central objectives of this thesis is the real-time localization and classification of sound events. To achieve this, we proposed rSNNs as a key component of our strategy. Moreover, mindful of the imperative to reduce computational costs, we investigated strategies to optimize the arrangement of neurons. We explored the spatial-temporal dimensions and proposed a refined approach that focuses computational efforts on specific positions. This fundamental shift led to a reconfiguration of the network's arrangement and topology.

In alignment with our quest for efficiency and innovation, we will integrate rSNN with adaptive resonance theory. This novel combination allows us to adapt the network's connections and augment the neuron population precisely in areas where sound events are detected, ensuring synergy with the environment. Notably, the burgeoning efficiency of spiking neural networks (SNNs) in pattern recognition, as exemplified by Wu (2021), and their significantly reduced computational demands, in contrast to artificial neural networks (ANN) (Srinivasan, 2019), fuels our motivation to craft a customized SNN structure, tailored for online detection and tracking of moving sound sources. To this end, we introduced the reservoir SNN (rSNN) structure and delved into the exploration of diverse interconnection parameters, investigating their impact on the performance of our sound event recognizer in the previous chapter.

One significant feature of the suggested network is a novel energy-based connectivity of neurons in 2D or 3D spaces, where only a subset of the neurons' connectivity weights will update based on received energy signals from the microphone arrays. This feature reduces the computational cost of the SNN.

This section underlines the real-time localization importance in this thesis. The realization of real-time localization by our method involves extensive testing and experimentation to validate the system's ability in continuously processing the auditory

data without experiencing lag. Our training method is tailored in such a way that it processes one sample at a time and, therefore, satisfies conditions embedded in applications that are real-time. In particular, each data sample of an epoch must have at least double the sampling rate to process it efficiently. Thus, the processing and the localization of the data occur within a time frame that admits online applications.

5.2 The Main Idea of Designing a New Architecture of rSNN

Due to the linearity of the physical laws of sound, the received sound is a linearly filtered version of the audio, corresponding to the location of the sensors, the location of the source, and the acoustical environment. In the proposed structure, inspired by binaural hearing, there should be at least two sensors in the environment to specify the synchrony patterns, by means of a pair of location-specific filters related to the sensors. The overall structure of the proposed architecture is depicted in Figure 10.

Figure 10

Overall Structure of the Proposed Method

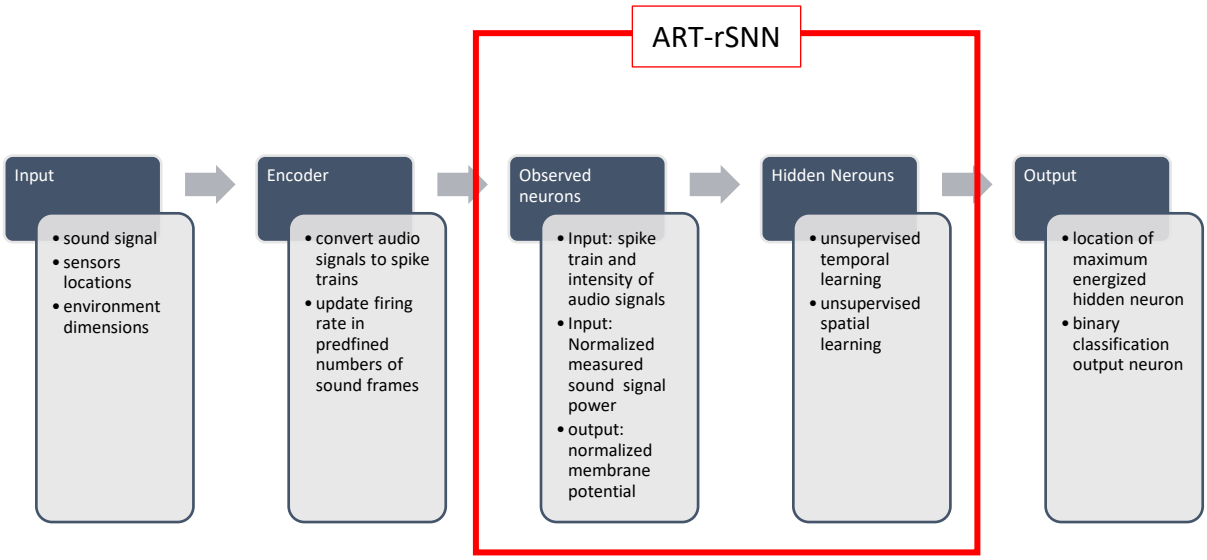


Figure 10 depicts that sound is first encoded based on a desired spike detection algorithm. Then both audio signals and the spike sequence code are input data to the proposed structure, which is composed of two groups of observed and hidden neurons. The observed neurons receive data directly from the environment. The hidden neurons

do not receive direct input from the outside, instead they are trained based on the observed activities of the neurons. In this structure, the location of the maximum-energy hidden neuron estimates the location of the sound source.

As shown in Figure 11, the first network size is directly relevant to the number of sensors in the under-study environment. Here we consider the minimum possible sensor quantity is two. In the proposed recurrent network, the position of each neuron in the initial arrangement is matched to the locations of the sensors. Clarifying the issue, Figure 11 indicates how the network grows, and the new neurons are generated.

Figure 11

Changing the Network Size Approach

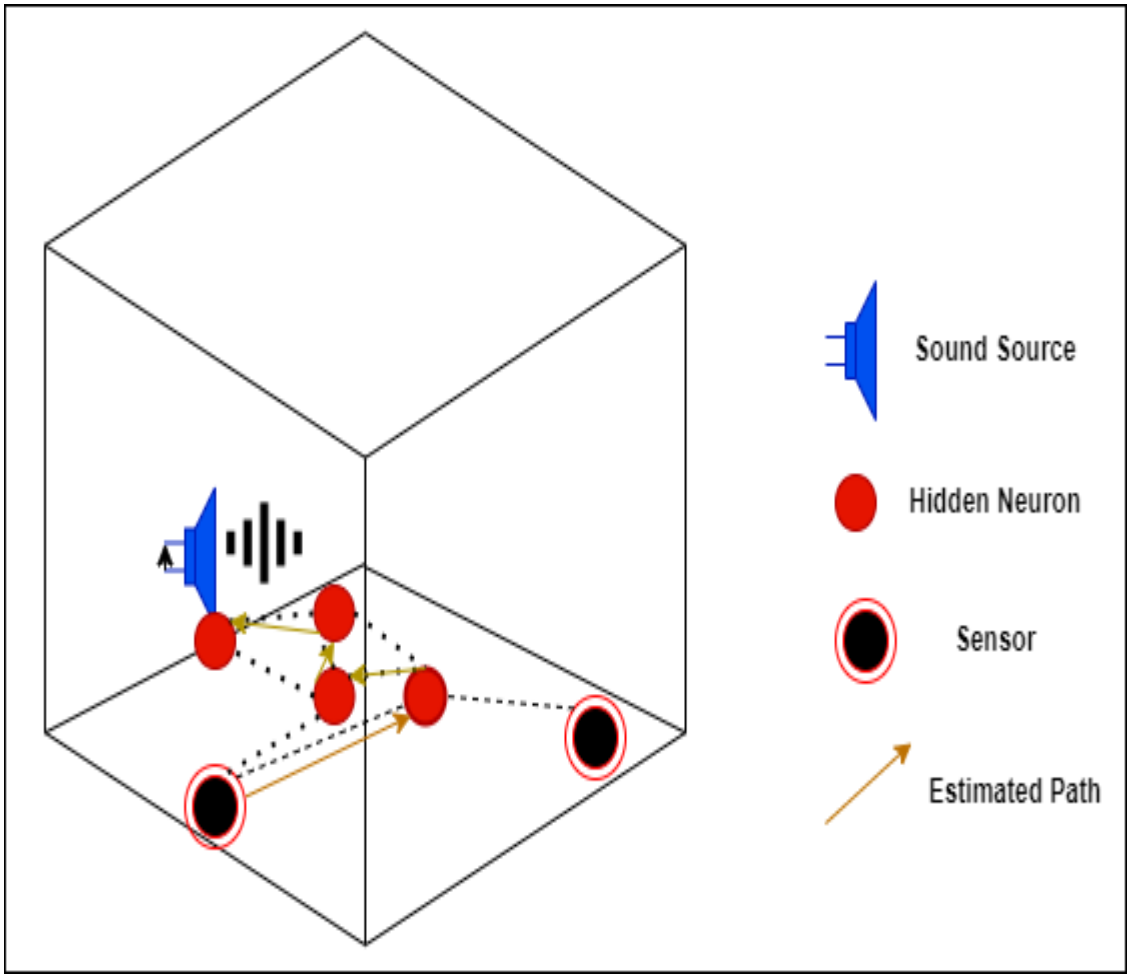


Figure 11 indicates that, at the initial state, the architecture embarks on its work by the number of observed neurons being the same as the number of sound sensors. The number of hidden neurons can be considered as the minimum possible number, for example, zero. Then, in each estimation epoch by the rSNN, a new hidden neuron will generate according to the estimated location of the sound source. Then, by considering the small-world technique, the role of newly generated neurons improves the learning quality of the proposed structure.

The "small-world technique" denotes a network connectivity paradigm distinguished by a pronounced clustering coefficient and a comparatively brief average path length among nodes. This configuration facilitates effective communication and information dissemination within the network, rendering it particularly appropriate for practical applications. The small-world technique is fundamentally distinct from alternative connectivity frameworks, including entirely random networks or connectivity configurations wherein less robust nodes are selectively eliminated based on weight magnitudes or eigenvalue analyses. While these alternatives possess their own advantages, the small-world methodology is favored in our context owing to its ability to sustain and enhance efficient communication among densely interconnected nodes.

This methodology is particularly pertinent for the management of spatial data and the integration of networks founded on location-based interrelations. Its implementation in our model enables proficient administration of clusters, as exemplified in our ART approach, thereby promoting superior clustering and data representation. By cultivating resilient connectivity through a harmonious blend of local and global links, small-world networks augment adaptability and uphold effective information flow, rendering them beneficial in applications such as sound localization.

In the new configuration, a threshold of a minimum required energy is considered to eliminate the neurons that receive low power signals. This threshold limits the connectivity of the networks. The active neighborhood area is detected based on the neurons' interactions criterion. In this regard, this work proposes a new dynamical structure inspired by the resonance theory in neural networks.

To extend the above idea in a mathematical model, this thesis uses the concept of adaptive resonance theory (ART) as a biologically plausible theory of how a brain learns to consciously attend, learn, and recognize patterns in a constantly changing

environment. The theory states that resonance regulates learning in neural networks with feedback (recurrence). Thus, it is more than a neural network architecture or even a family of architectures.

Through dynamic creation of recognition categories for encoding distinct input samples, an ART module is capable of self-adjusting the scale of its recognition field, in terms of the number of committed nodes, with respect to the complexity of the problem domain. Its fast commitment mechanism and capability of learning at a moderate rate of speed guarantee high efficiency. However, given a dataset, the scale of the ART recognition field (i.e., the number of output clusters) depends on a global threshold parameter called vigilance. While in principle, one could control ART's recognition representation by fine tuning the vigilance parameter, in practice, suggesting an appropriate vigilance value requires prior knowledge of the scale and the distribution of the problem dataset, which is unlikely to be available (Carpenter et al., 2016).

Additionally, while sound travels through the air, acoustic energy is emitted omni-directionally from the sound source. The strength of a sound source diminishes at a rate inversely proportional to the square of the distance. The traditional formulation of this algorithm is given as follows (Chen et al., 2021):

$$y_i(t) = \zeta_i \frac{S(t)}{|r_i - r(t)|^{\alpha}} + \varepsilon_i(t) \qquad (5\text{-}1)$$

Equation 5-1 describes the relation of measured signal $y_i(t)$ on the ith sensor with $S(t)$ as the actual sound energy, recorded from a 1-meter distance from the sound source. ζ_i is the gain factor of the ith acoustic sensor. r_i (sensor location) and $r(t)$ (unknown location) indicate the coordinate of ith sensor node and sound source at time t. Each variable is a vector with two additional variables when in a two-dimensional (2-D) plane.

While this model provides a useful approximation for modelling sound attenuation, it does indeed have limitations. Specifically, the model does not explicitly account for reverberation and other complex acoustic phenomena.

When there are m sensor nodes, the value of α can be obtained as:

$$\alpha = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \alpha_{ij} \qquad (5\text{-}2)$$

where α_{ij} is the path loss exponent obtained by the ith and jth sensor nodes ($\alpha_{ij} = \alpha_{ji}$). To simplify the problem, $\alpha=2$ is considered. Regarding both deterministic and metaheuristic algorithms, all observations from the multiple sensors are aggregated as an estimator of $r(t)$, where the solution of the localization problem is the argument (pair of coordinates) that minimizes the expression.

$$\hat{r}(t) = \arg \min_{r} \sum_{i=1}^{m} \frac{1}{\sigma_{\xi_i}^2} \left(y_i - \zeta_i \frac{S(t)}{|r_i - r(t)|^{\alpha}} \right)^2 \qquad (5\text{-}3)$$

The estimator in Equation 5-3 is highly nonconvex, with singularities in each sensor's coordinates, several suboptimal solutions, and saddle regions. All the enumerated features make the problem very challenging in the field of numerical optimization, making it a good candidate in the context of regression and ANNs.

In this study, a leaky integrate-and-fire neuron is utilized. The membrane potential V evolves according to the equation (Goodman & Brette,2010) :

$$\frac{dV_i(t)}{dt} = \frac{1}{\tau_m}\left(-V_i(t) + I_i(t)\right) \qquad (5\text{-}4)$$

where τ_m denotes the membrane decaying time constant and $I_i(t)$ is synaptic current.

5.3 Proposed ART-rSNN Method

This study presents a new structure of a reservoir-spiking neural network, able to generate new hidden neurons. This structure uses the small-word connection strategy. In the initial states, only m observable neurons, which receive input signals (measured signal of microphones/sensors) are regarded. The main goal is estimating the real energy of the signal by approximating y as the neuron output value. Consider the cost function as follows:

$$J = \frac{1}{2} E^T E \qquad E = \left(\tanh(y_{s_{N\times 1}}) - \tanh(V_{o_{N\times 1}})\right)_{N\times 1} \qquad Vo: Observed \quad Neuron \qquad (5\text{-}5)$$

Equation 5-5 describes the square error on normalized energy estimation. Here, we integrate the concept of energy-based methods to ITD via this formula:

$$\text{suppose}: \quad V = \exp(-c\Delta t_s / \tau) \quad \rightarrow I = \frac{1-c}{\tau}\exp(-c\Delta t_s / \tau) \rightarrow I = W_{ij}\exp(-c\Delta t_s / \tau)$$

$$c : sound\ speed$$

$$\Delta t_s = \quad input\ spike \quad time \quad - neuron \quad spike \quad time$$

(5-6)

where W_{ij} is the synaptic weight between neuron i and j and updated based on spike time dependent plasticity (STDP) laws for hidden neurons:

$$STDP: \ w_{ij} = \begin{cases} Ae^{\frac{\Delta t}{\tau}} & \Delta t \geq 0, \quad i, j \in Ni \\ 0 & \Delta t < 0, \quad or \quad i, j \notin Ni \end{cases}, \qquad \Delta t = t_i - t_j \qquad (5\text{-}7)$$

where t_i and t_j are the spike times of the i_{th} and j_{th} neurons, respectively. *Ni* is the neighbourhood of the neurons in the small word connections, A is maximum synaptic weight, $\Delta t = t_i - t_j$ is spike time difference, and τ is time constant. *A* is the tuning parameter of the weight and τ is the time constant of the synaptic plasticity laws.

Each neuron is located in a special coordinate, corresponding to the sensor's location. Intending $\psi_i = \zeta_i \frac{S(t)}{|r_i - r(t)|}$, the following cost function is defined, based on the minimum estimation of the energy error:

$$J = \frac{1}{2}\sum_i \left(y_i - \psi_i\right)^2 = \frac{1}{2}E^T \times E \qquad (5\text{-}8)$$

Where The observed neuron updating synaptic weight law is calculated as follows:

$$\vec{y}_o = \tanh([y_1, \ldots y_{N_{ob}}]^T), \qquad N_{ob} : \text{Senesors Numbers}$$
$$\vec{\Psi}_o = \tanh([V_1, \ldots V_{N_{ob}}]^T)$$

$$E = \vec{y}_o - \vec{\Psi}_o$$
$$\Delta W_{oh} = -0.5\frac{\partial J}{\partial W_{oh}} = \qquad (5\text{-}9)$$
$$-E^T \times \left(\frac{\partial J}{\partial W_{oh}}\right) = -\left(E^T \times \left(\vec{1}_{N\times N} - \tanh(\vec{V}_o \times \vec{V}_h)\right).e^{-c\frac{\Delta t_s}{\tau}}\right)^T$$

Where $\vec{V}_h = \begin{bmatrix} V_1 & \ldots & V_N \end{bmatrix}^T$ and $\vec{V}_o = \begin{bmatrix} V_1 & \ldots & V_{N_{ob}} \end{bmatrix}^T$, *N* is the neuron numbers, including both observed and hidden neurons.

$$\Delta\Phi = -\left(E^T \times \left(\vec{1}_{N\times N} - \tanh(\vec{V}_o \times \vec{V}_h^{\,T}) \right) . \left(\frac{1}{\tau} \times I_{indentity} \right) \right)^T \qquad (5\text{-}10)$$

For a mere localization of a sound source, we can utilize the concept of STDP in hidden neurons, updating the law shown in Equation 3-19. However, in order to jointly event triggering and classifications, the STDP rule will be modified in the new proposed structure, which is presented in the following sections and equipped with the classification output module. We utilize the ART strategy to develop our SNN structure as follows:

$$\Delta t_{sij} = t_{si} - t_{sj} \qquad (5\text{-}11)$$

$$\kappa = \frac{V_i}{V_j + (\mathrm{sgn}(V_j) + 0.5)\varepsilon} \qquad V_i = \max \quad V \; and \quad V_i \geq V_j > \quad V_k, \varepsilon > 0 \quad k \neq i, j$$

(5-12)

$$d = \begin{cases} \dfrac{c\kappa\Delta t_{sij}}{1-\kappa} & if\ V_i \neq V_j \quad d_{ij} = \left\| PosV_i - PosV_j \right\|_2 \\ \dfrac{d_{ij}}{2} & else \end{cases} \qquad (5\text{-}13)$$

$$PosV_{new} = PosV_i + \begin{bmatrix} d\cos(\theta) \\ d\sin(\theta) \end{bmatrix} \qquad (5\text{-}14)$$

$$f(\varphi) = \alpha_l(\varphi + \sin(\varphi)) - \alpha_l(\mathrm{sgn}(\varphi)\pi - 2\varphi)\left|\beta\sin(\varphi)\right|, \quad \alpha_l = \frac{d_{ij}}{2c} \quad , 0 < \beta < 1 \qquad (5\text{-}15)$$

$$\theta = \begin{cases} \cos^{-1}\left(\dfrac{c\Delta t_{sij}}{d.d_{ij}}\right) & if\ \left|\Delta t_{sij} - f(\varphi)\right| < \left|\Delta t_{sij} - f(-\varphi)\right| \quad , \varphi = \cos^{-1}\left(\dfrac{c\Delta t_{sij}}{d.d_{ij}}\right) \\ -\cos^{-1}\left(\dfrac{c\Delta t_{sij}}{d.d_{ij}}\right) & else \end{cases} \qquad (5\text{-}16)$$

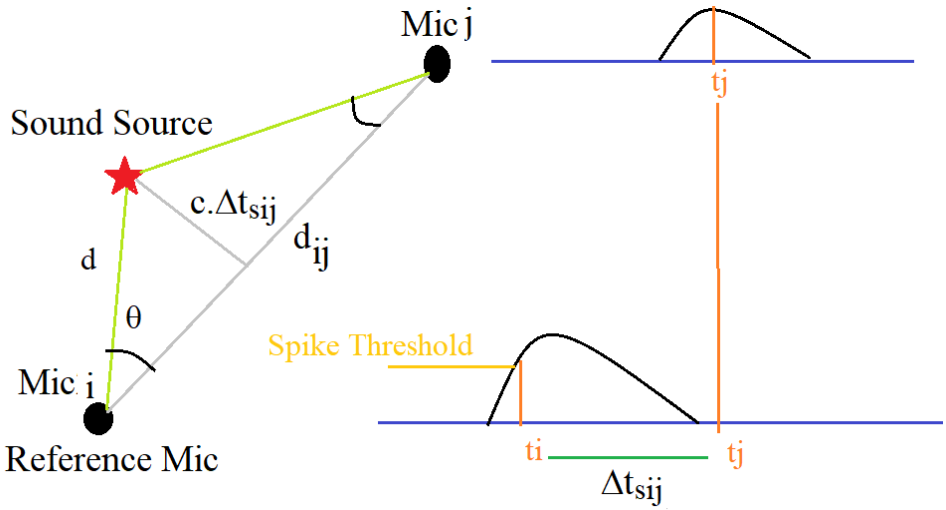
where Δt_{sij} is the difference of spike time in neuron *i* and neuron *j* and ε is constant parameter. β is 0.001.

Equations 5-12 to 5-15 describe how a new neuron is generated. κ is the ratio of neuron potential i and j. This parameter indicates that the received energy by neuron i is how much stronger than jth neuron. *d* denotes to the sound source distance from the

reference neuron. Equation 5-14 calculates the new position of neuron, and $f(\varphi)$ in Equation 5-15 indicates if the source is located at the front or back. Equation 5-16 describes the azimuth angle calculation formula. Figure 12 indicates the graphical abstract of the proposed method. So, the new generated neuron is the estimation of sound source location, in each epoch.

Figure 12

Sinusoidal Relation between Parameters Based on Time Delay



According to Equation 5-17, the power of the signals is considered directly relevant to the membrane potentials of the neurons in the proposed structure; namely:

$$power = \frac{V^2}{Z} \rightarrow power \propto V^2 \propto \exp\left(\frac{-t}{\tau_{new}}\right) \qquad (5\text{-}17)$$

In the next section, sound event trigger classification ability is synergized in the new structure by modifying the STDP rule and threshold of the membrane potential, following the method presented in Wu (2019).

5.4 Classification Module in the New Structure

In this section, we propose to modify the STDP learning strategy based on homeostatic plasticity law through adaptive threshold (Wu, 2019). The plasticity of homeostatic by restricting the synaptic weight and firing rate, which is essential for the stability of the

neural circuit, complements the STDP rule (Watt & Desai, 2010). Cook (2015) discusses the inhomogeneity of input and lateral input in conjunction with the various firing rates. As a result, neurons are increasingly being upgraded, whereby a small portion of the neurons with a high firing rate dominates the synaptic plasticity, according to the STDP rule, and the rest of the neurons do not manifest considerable features. Therefore, it is desirable to maintain a balanced firing rate among the neurons, so that a variety of fields of receiving fields are more diverse to form in the feature extraction layer. To achieve this goal, this study adopts an adaptive threshold mechanism (Wu, 2019), whereby the optimal firing rate of f_d for neurons in the feature extraction layer is predefined. The firing rate of the f_t is calculated after presenting a determined number of audio frames and the rate of neuron membrane potential threshold v_{th} is updated as follows:

$$\frac{dv_{th}}{dt} = \lambda_h (f_d - f_t) \qquad (5\text{-}18)$$

where λ_h is a constant learning rate. This can be set to 0.5, in order to match quickly in thousands of frames.

5.4.1 Supervised Temporal Classifier

Tempotron (G¨utig & Sompolinsky, 2006) is a temporal learning rule that works based on the membrane potential of the neurons to train binary classifiers. In this learning law, a trained neuron fires whenever it observes patterns from its own class. Otherwise, its membrane potential remains under v_{th}. In this classification module, each class is assigned to a specific output neuron. In other words, if the input pattern belongs to the *i*th class, the *i*th output neuron is a mere neuron which fires, and the rest of the neurons are silent. The STDP-tempotron for the *i*th output neuron is described by:

$$\Delta w_i = \begin{cases} \lambda_0 \sum_{t_i < t_{\max}} K(t_{\max} - t_i) & if \quad v_{th} > v_{t_{\max}} \\ -\lambda_0 \sum_{t_i < t_{\max}} K(t_{\max} - t_i) & v_{th} < v_{t_{\max}} \\ 0 & else \end{cases} \qquad (5\text{-}19)$$

where λ_0 is a constant learning rate, $t_{\max}$ is the time sample when the neuron potential peak is maximized, and t_i is the last spike time of neuron *i*.

$K(.)$ is the kernel function, which is calculated by:

$$K(t-t_i) = K_0\left(\left(\exp\left(-\frac{t-t_i}{\tau_m}\right)\right)-\left(\exp\left(-\frac{t-t_i}{\tau_s}\right)\right)\right) \quad (5\text{-}20)$$

where K_0 is a normalization factor that guarantees that the maximum amplitude of the kernel ($K(t-t_i)$) is 1. τ_m and τ_s are, respectively, the membrane and synaptic decay time constants to determine the shape of the kernel function. $K(t-t_i)$ regards only spikes before time *t*. The membrane potential of the output neurons can be updated by Equation 5-9 and is based on the STDP-tempotron learning law.

5.5 Sound Event-Triggering Capability

Event-based applications of new architectures need to eliminate the effects of noise on the received signal. Therefore, the input signal should be encoded in such a way that the effects of noise can be removed. In this regard, the mean of the input signal can be updated and changed by the following equation:

$$X_{mean}(k+1) = X(k) + \frac{k-1}{k} X_{mean}(k) \quad (5\text{-}21)$$

where $X_{mean}(k)$ is the spike detection threshold, $X(k)$ is the input signal, and k is the sample time epoch. The estimated covariance of the signal is calculated by (Bruce,1969):

$$\sigma^2(k) = \sigma^2(k-1) + (X(k) - \frac{X_{mean}(k-1)+X(k)}{k})^2 + \frac{k-1}{k}(\frac{X_{mean}(k-1)}{k-1} - \frac{X_{mean}(k)}{k})^2 \quad (5\text{-}22)$$

The input signals are suggested to be encoded as follows:

$$R_k = \exp(-\frac{\left(X(k) - X_{mean}(k)\right)^2}{2\sigma^2(k)}) \quad (5\text{-}23)$$

$$t_{sp} = (1-R_k)k\Delta t \quad (5\text{-}24)$$

where R_k is the normalized input, and t_{sp} is the spike time. By the new proposed structure, we present the adaptive physic-informed ART-rSNN architecture to sound event triggering and classification for a single sound source.

5.6 Adaptive Physic-informed ART-rSNN

In order to embed the new adaptive real-time encoding approach to the newly suggested structure, we define the error of the power estimation to match the physical concepts of the electrical power definition for a spike neuron circuit model with the measured power of input audio signal:

$$E = (X_{in} - V_{ob}(k) I_{oh}(k)) \qquad (5\text{-}25)$$

where $V_{ob}(k)$ is the observed neuron membrane potential and $I_{oh}(k)$ is the input synaptic current to the observed neurons, which is calculated by:

$$I_{oh}(k) = W_{oh}(k) V(k) \qquad (5\text{-}26)$$

where $W_{oh}(k)$ is the weight of observed neuron and $V(k)$ is the vector of membrane potential of all neurons. In this regard, Equations 5-7 and 5-8 are modified and integrated as follows:

$$\Delta W_{oh}(k) = \eta (X_{in}(k) - V_{ob}(k).I_{oh}(k)).(V_{ob}(k) \times V(k)^T) \qquad (5\text{-}27)$$

In this structure, STDP-tempotron is utilized for rSNN and the output binary layer (Equations 5-11 - 5-24). To intend the spatial-temporal STDP-tempotron learning law for the hidden neurons, we suggest modifying the membrane potential of the hidden neurons as follows:

$$V_j = \sum_{i}^{h+o} W_{ji} (v_0 \left(\exp(-\frac{t - t_i - \frac{d_i}{c}}{\tau_m}) - \exp(-\frac{t - t_i - \frac{d_i}{c}}{\tau_s}) \right) - v_{th} \exp(-\frac{t - t_j - \frac{d_j}{c}}{\tau_m})) \qquad (5\text{-}28)$$

where V_j is the hidden neuron *j* and *i* denotes to *i*th neuron among all neurons. c is the acoustic wave speed, here *c*=343 (m/s). t_i and t_j are, respectively, spike times of neurons *i* and j. d_i and d_j are distance of the neurons *i* and j from the origin or reference node.

In the ART-rSNN structure, we have considered LIF neuron with *I* inputs, and its postsynaptic membrane potential represented by *Vj (t)* stays at the resting potential $V_{rest} = 0$ when no spikes are received. Keeping abreast of a spike produced at a pre-synaptic neuron i, a postsynaptic potential (PSP) is induced in the LIF neuron. By integrating the

PSPs resulting from a number of the spikes inputs, the LIF neuron fires a spike when its membrane potential V (t) reaches the firing threshold v_{th}. Therefore, Equation 5-28 results, based on the dynamics of the neuron postsynaptic membrane *V (t).*

The position of each hidden neuron is updated according to the following criteria:

$$\Delta d = -\frac{\partial J}{\partial d} = \frac{1}{c.\tau_m}(X_{in}(k) - V_o(k)^T W_{oh} \exp(-\frac{t - t_j - \frac{d_h}{c}}{\tau_m})).\left(V_o(k)^T W_{oh} \exp(-\frac{t - t_j - \frac{d_h}{c}}{\tau_m}) \right) \qquad (5\text{-}29)$$

Equation 5-29 indicates that position of the neurons, namely, distance from the reference node, is updated so that energy estimation is minimized. In this situation, we can determine how much of the sound source is far away from the origin.

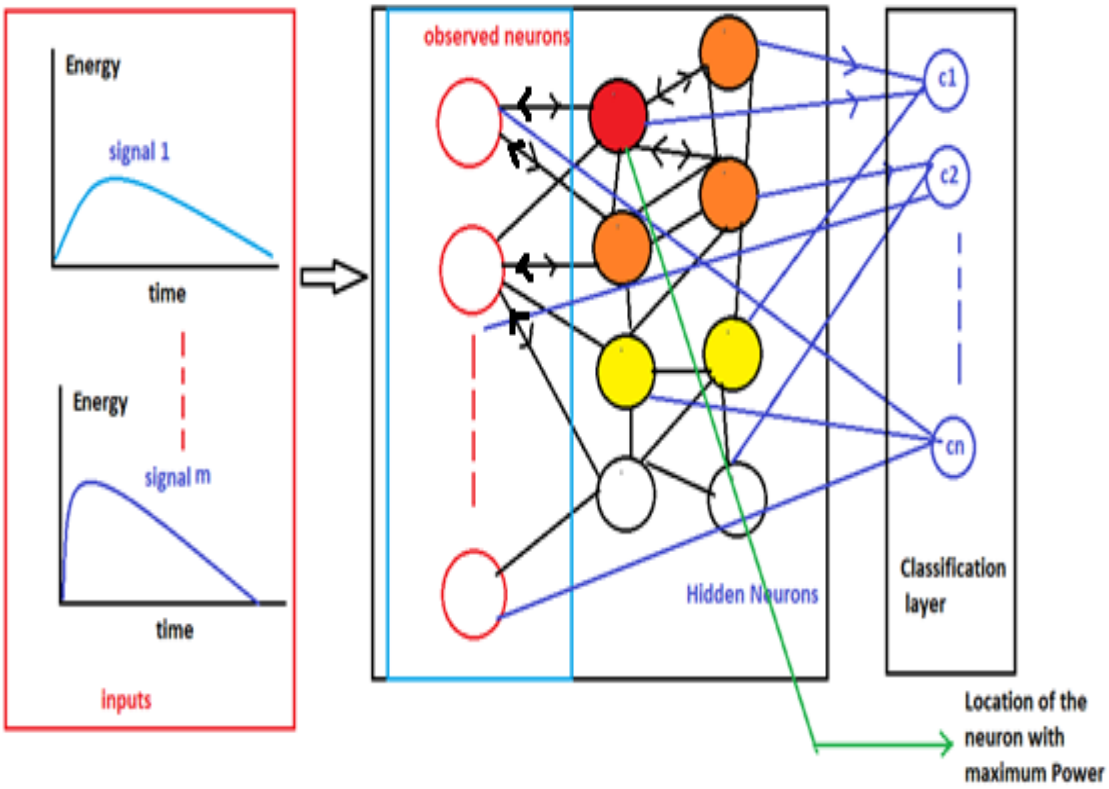
To calculate azimuth angle, Equations 5-15 and 5-16 are still held for this algorithm. The synaptic weights of hidden neurons are updated by Equation 5-17.

The output layer of the proposed structure is the classification layer as well as the indicator of maximum-power hidden neuron location as the final estimation of sound source position. In fact, this structure is designed to handle the audio tracking problem with only one sound source, however the case of multiple sound sources is also investigated in the next chapter.

The overall structure of the adaptive physic-informed ART-rSNN is shown in Figure 13.

Figure 13

Overall Structure of Adaptive Physic-informed ART-rSNN

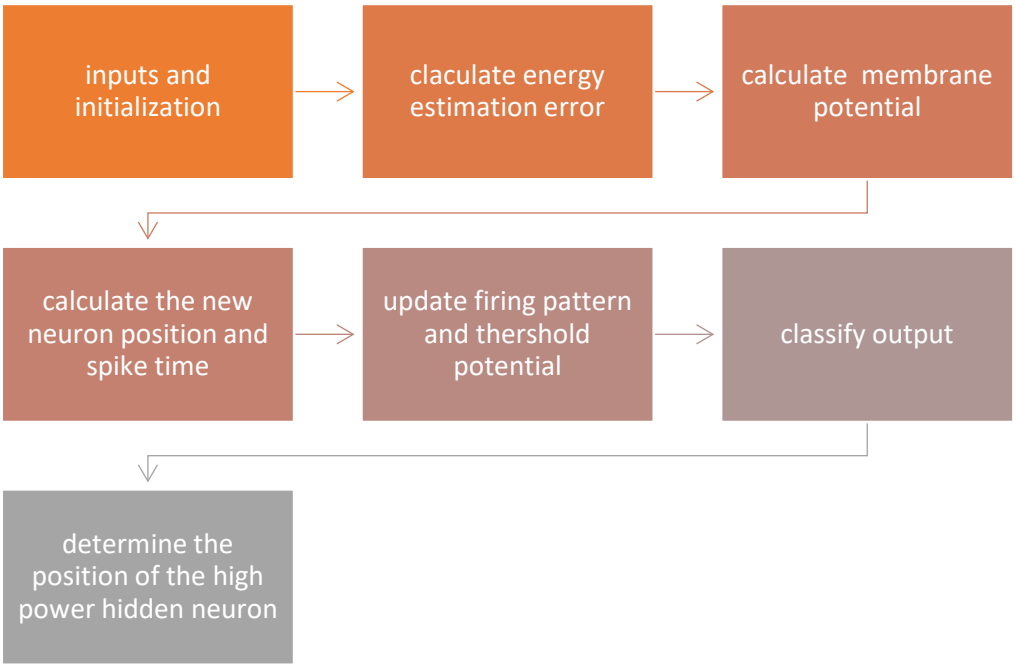


In Figure 13, *C1,…,Cn* is the n event binary class indicator, which works based on the firing rate criterion. The positions of the hidden neurons and their quantity are updated based on the ITDs and energy-based cues sound localization method in each epoch.

Figure 14 shows the main steps of the of the adaptive physic-informed ART-rSNN algorithm.

Figure 14

Major Steps of the Adaptive Physic-informed ART-rSNN



The main steps of the designed structure are defined as follows:

Step 1: Initialize weights, *V* and *I.*

Step 2: Calculate estimation error (Equation 5-25), synaptic current and membrane potential update (Equations 5-26 - 5-28).

Step 3: Calculate spike sequence and spike time.

Step 4: Update positions of the neurons and new generated neuron as per Equations 5-25, 5-26, and 5-28 - 5-32.

Step 5: Calculate firing rate based on computing average firing numbers of determined numbers of windows.

Step 6: Classify the output by calculating output neuron membrane potential, based on the hidden neuron model and updating law (Equation 5-28), but we consider $d_i=0$ for the output layer.

Step 7: Find maximum power hidden neuron and its location.

In the case of multiple high power hidden neurons, we calculate the centroid of the high-power neurons' neighborhood in the case of multiple sound sources or under available disturbances.

5.7 Summary

In this chapter, we have devised advanced localization techniques by harnessing the potential of reservoir spiking neural networks (rSNNs) in conjunction with minimizing sound energy decay modelling error. This optimization was achieved through the application of a cost function within the learning process, seamlessly integrating insights from energy-based methods and interaural time difference (ITD) and interaural intensity difference (IID) cues.

Within this context, the combined usage of ITD and IID methods, structured in a triangular approach within the adaptive resonance theory (ART), led to the expansion of neuron sizes, with a primary focus on the most recent sound event locations. Subsequently, the highest potential neuron, post-training, serves as the indicator of the estimated sound source location.

In the next chapter, our focus will shift towards enhancing this algorithm by introducing a physics-informed relationship between neuron power and acoustic sound events. Furthermore, we will delve into the development of a classification and sound detection module, thereby expanding the capabilities of our localization approach.

6. Results and Findings

6.1 Introduction

In this chapter, we evaluate the new designed structures in localization, classification, and jointly classification and localization tasks.

The code utilized in this thesis is developed and simulated using Python 3.10, a robust platform for our sound classification and localization models. We employed the librosa library for comprehensive sound signal processing, including feature extraction tasks, such as MFCCs and other transformations. The simulations for models like the CRNN, MLP and RNN were implemented with the TensorFlow library, ensuring effective training and evaluation of these deep learning models. Our method was custom-developed using NumPy and SciPy, employing the algebraic formulations (and principles) detailed in this thesis to create a unique, physics-informed, Art-rSNN model. Given our plans for future enterprise and commercialization, our code will remain private; it will not be shared publicly. This approach aligns with our vision of leveraging our innovative method in potential applications and further development initiatives. However, we recognize the importance of collaboration and may consider partnerships in the future.

6.2 Datasets

In this work, we have utilized the following datasets:

- Real recorded signals by two different mobile microphones: This dataset, which is described in the last section, includes a single moving source of hand clapping sound, in the frequency band of 400Hz to 2400 Hz, with environment noise.
- L3DAS22 Task 2: This dataset is split into three subsets: training, validation, and test. The training subset consists of 600 30-second-long audio recordings, while the validation and test subsets each consist of 150 30-second-long audio recordings. The dataset includes 14 types of sound events selected from the FSD50K dataset, with one overlapping sound event and four classes of sound events utilized: writing, knock, drawer open and close, and cupboard open and

close. The room impulse response (RIR) is sampled in an office room with dimensions around 6 m (length) by 5 m (width) by 3 m (height), and FOA microphone arrays are placed in the centre of the room. The position of the FOA microphone arrays is set to be the origin of the coordinates.

- TAU-NIGENS spatial sound events 2020: A spatial sound dataset that consists of several spatial sound scene recordings that contain different classes of sound events embedded in various acoustic spaces and various directions and distances of sound sources seen from the recording locations. All sound events are based on filtering through real spatial impulse responses (RIRs) recorded in various rooms of different shapes, sizes, and sound absorption characteristics. Additionally, each scene recording is provided in two spatial recording formats – microphone array (microphone) and first order ambisonics (FOA) sound events – that are static or moving sound sources in a room that uses time-varying RIRs. Each audio event in an audio scene is associated with its entry path to the save point, as well as a start time and time offset The individual tapes used to build the audio scene come from the NIGENS public audio event database (Politis, 2020).

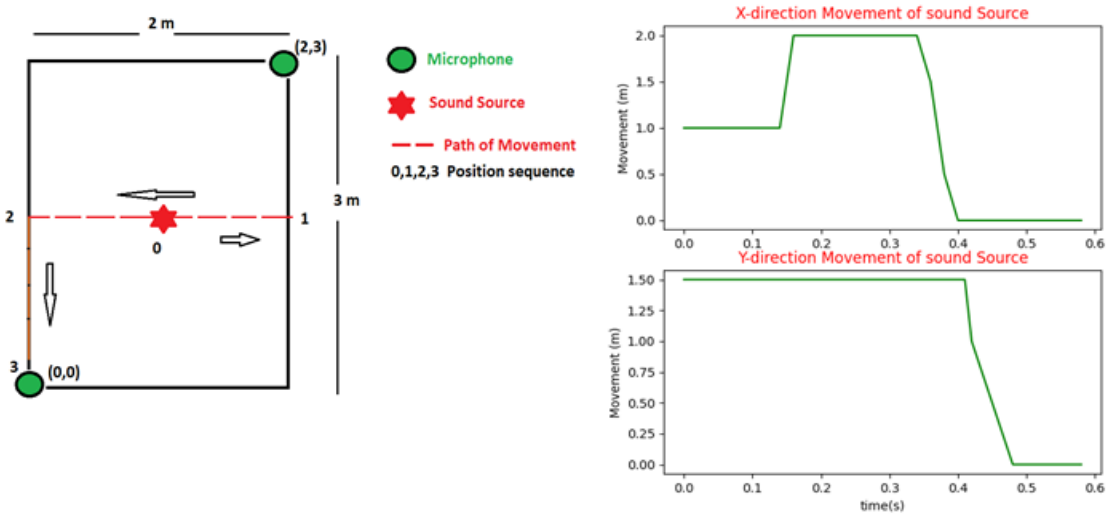
The first dataset is comprised of two recorded sounds, including periodic noisy clapping sounds. We have used two omnidirectional microphones to record the audio and the positions of the microphones are shown in Figure 15. The noises are mainly generated by vehicle movements, approximately lower than 30 db. The under-study coordinate is 2*3 m, which is located in a larger 12 m2 area equipped with furniture and negligible reverberant environment. The microphones' Z-axis is zero. A sound source is considered in this record, which moves linearly along a 2D environment. Environment arrangement and the sound source x and y movement paths are indicated in Figure 15.

The use of 2 microphones in this study is primarily due to the constraints of the external dataset employed in the research. This dataset was designed with a two-microphone setup, which aligns with many practical applications and common experimental configurations in sound localization tasks.

However, it's important to note that the proposed model and methodology are not inherently limited to just two microphones. The approach can be extended to accommodate more than two sensors if desired or required by different experimental setups or real-world applications.

Figure 15

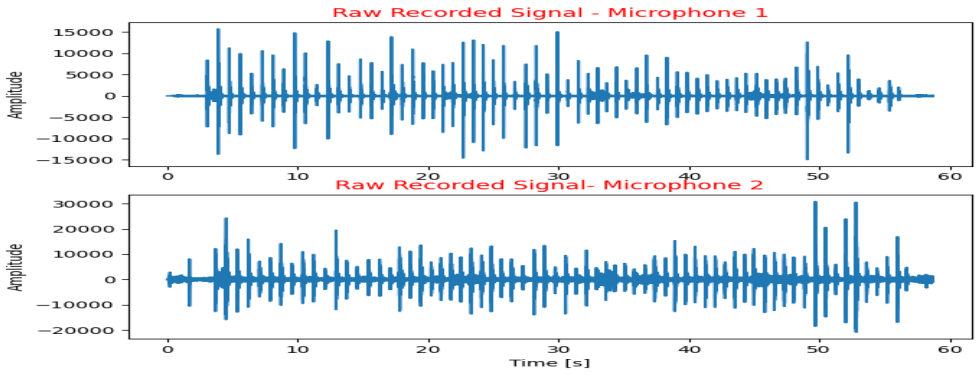
The Sound Source, Microphones Arrangements, and Movement Path of the Sound Source in 2D x- y axes



The signals, recorded in a real environment, are depicted in Figure 15. In this figure, the upper figure is the recorded signal by microphone 1, which is located at (2,3), and the lower figure is the recorded signal by microphone 2, which is located at (0,0).

Figure 16

Recorded Raw Signals by Microphones 1 and 2



As shown in Figure 16, the signal-to-noise ratio (SNR) is low and there is background noise in both recorded signals; so, filtering is necessary to clean the data. The first 10 seconds of the recorded signals include only a single clap hand audio signal in position (0,1).

We use a band-pass filter to remove the background noise of the recorded audio. The filter we use is Butterworth, 5 degrees with band pass 400 Hz – 1000 Hz. Figure 17 shows 1 second filtered signals, recorded by two sensors 1 and 2.

Figure 17

Filtered Signals

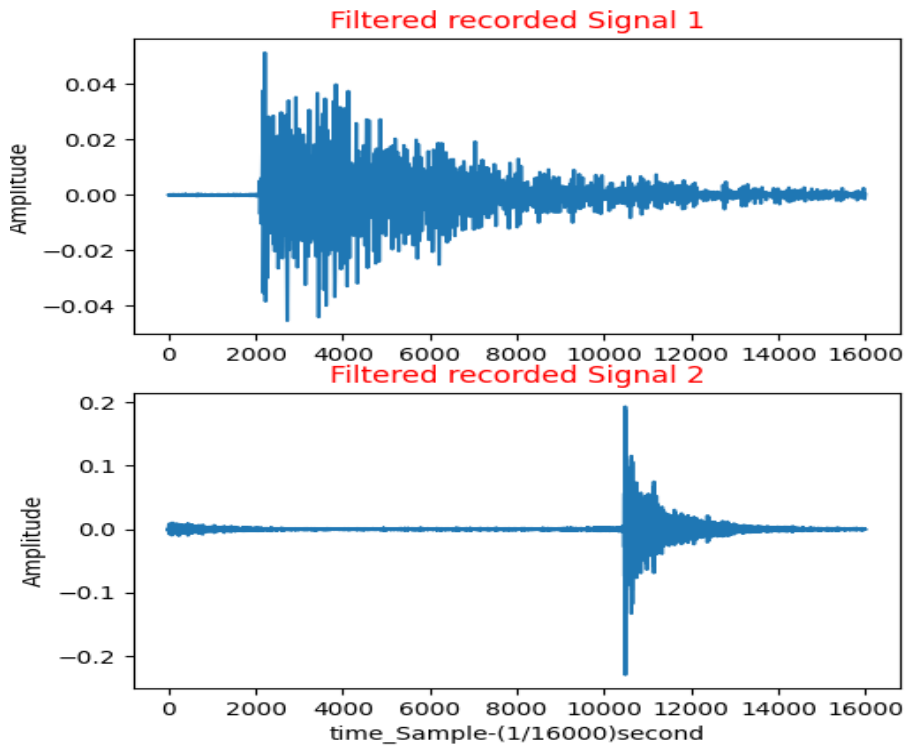
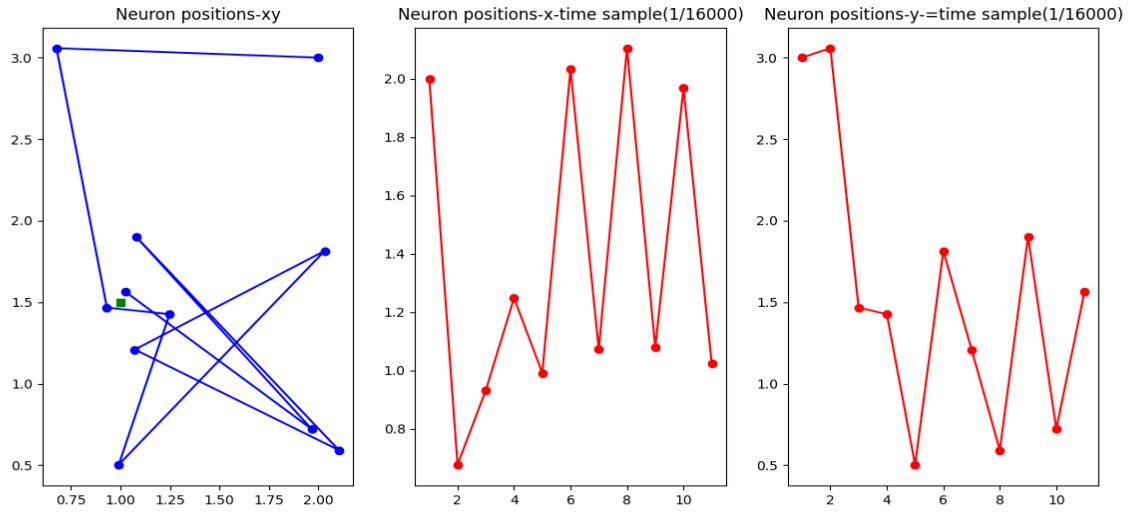


Figure 17 reveals that microphone 1, which receives the sound signal at time sample 2000 and with a high signal intensity, is closer to the sound source than microphone 2, which receives the sound signal with a delay and lower intensity. In the upper block of Figure 17, signal 1 is recorded at position (2,3) and in the lower block of Figure 17 signal 2 is recorded at (0,0). The signals are filtered by a Butterworth band pass filter. One second of the recorded signal are depicted.

In Figure 17, signals are shown before the normalization stage. We applied tanh normalization to the presentation signal, as it did not appear to be properly normalized.

In the next section, the performance of the proposed recurrent spiking neural network (RSNN) is investigated in a real-world environment. To evaluate and find the performance of this dynamic structure, including spatial data, we compare both fixed and dynamic architectures of a RSNN.



6.3 Evaluation of the Role of a Dynamic Structure in RSNNs

In this section, the performance of a RSNN is evaluated with two different real sample data in an environment. Sound sources are recorded at (0,1.5) and (1,1.5), considering the location of microphone 2 at (0,0) as the reference node. Figure 18 (a, b, c) depicts how the proposed architecture localizes a sound source located at (1,1.5).

Figure 18

Neuron Position Tracking for a sound source located at (1,1.5).

(a) indicates the hidden neurons' arrangements after growing network in each iteration in x-y coordination, (b) and (c) respectively indicate the x- direction neuron arrangements and the y-direction of the new neuron arrangement.

a) x-y neuron positions b) x-axis of neuron position c) y-axis of neuron

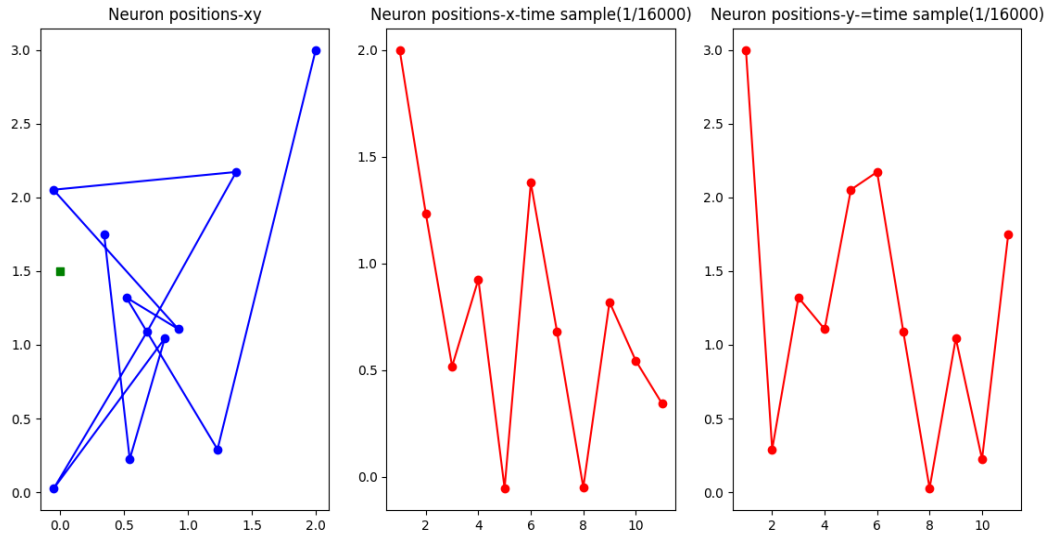
a) x-y neuron positions b) x-axis of neuron position c) y-axis of neuron

Figure 19 reveals that how the high potential neuron position, created by the proposed dynamic architecture, is updated. The sound source (green square) is located at (1,1.5) after 10 sample time, new neuron created at location (1.03,1.57), which the norm 2 of sound source tracking error is almost 0.1. To work out the performance of the proposed architecture, Figure 18 depicts another example of sound source localization by the proposed method, when the sound source is located at (0,1.5).

Figure 19

Neuron position tracking for a sound source located at (0,1.5)

(a) indicates the hidden neurons' arrangements after growing network in each iteration in x-y coordination, (b) and (c) respectively indicate the x- direction neuron arrangements and the y-direction of the new neuron arrangement



a) x-y neuron positions b) x-axis of neuron position c) y-axis of neuron

Figure 19 (a, b, and c) shows that the final location of the new generated neuron is at (0.31, 1.7), which the relevant norm2 of the error of tracking is almost 0.4. The sound source (green square) is located at (0,1.5).

In the fixed structure, hidden neurons are randomly arranged and the location of neuron with the higher membrane potential is considered as the best estimation of sound

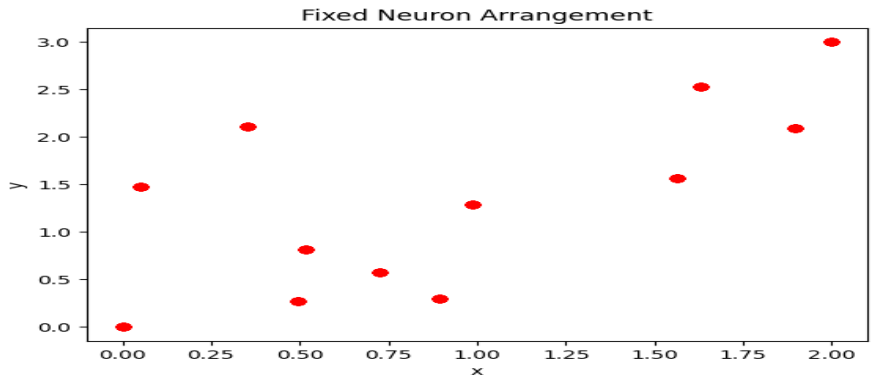
sources. With the aim of integrating spatial data, instead of a mere time difference cue, the STDP updating law is modified according to the following equation.

$$w_{ij}(t) = Ae^{-(\frac{d_{ij}}{c_0})}, \quad c_0 > 0, \quad A > 0 \tag{6-1}$$

where c_0 is the tuning law, chosen based on the acoustic velocity in the environment. A sample of the arrangement of the fixed structure RSNN is indicated in Figure 20.

Figure 20

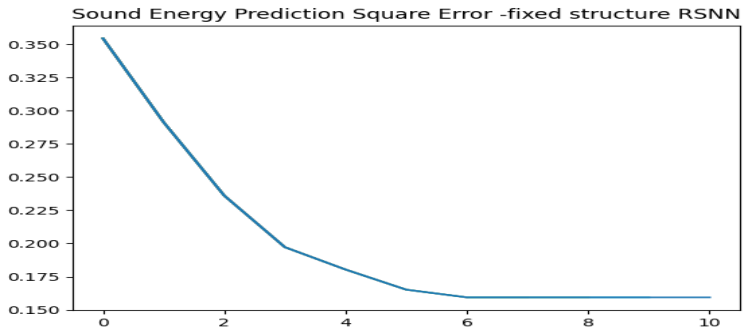
Fixed Neuron Arrangement with 10 Hidden and 2 I/O Neurons (Located at (0,0) and (2,3))



As shown in Figure 20, 10 hidden neurons are considered as well as two I/O neurons. Received energy is predicted by the observed neurons, which the relevant mean square Error (MSE) is illustrated in Figure 21.

Figure 21

Sound Energy Prediction MSE, Calculated by the Fixed Structure RSNN



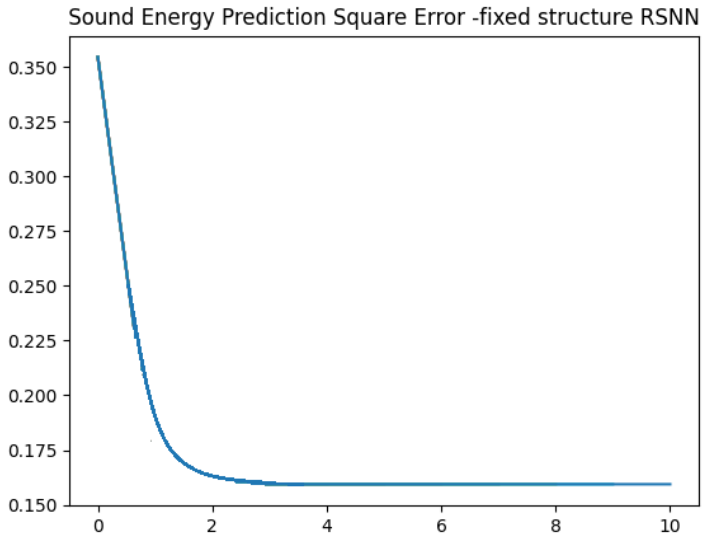
In Figure 21, the Y axis indicates the MSE amplitude and the X axis represents the iteration number. In addition, Figure 21 indicates that the calculated MSE converges to

a specific value which denotes that the proposed estimator is biased; therefore, we should have normalized the input properly to have an unbiased estimation. MSE of energy converges to 0.15 which indicates we have biased estimation of energy in our model in this regard we can consider unbiased approaches or adjust specific threshold for the future work.

Although MSE quickly converges to its steady state value, the question still arises of what happens if the number of hidden neurons increased? To respond this question, we raise the number of hidden neurons to 100. Figure 22 indicates the MSE of sound energy prediction by the 100 hidden neurons.

Figure 22

MSE of Sound Energy Prediction by Fixed Structure RSNN Method with 100 Hidden Neurons



In Figure 22, the Y axis indicates the MSE amplitude and the X axis represents the iteration number. Higher convergence speed is clearly detectable in figure 22; however, utilizing the time process function of the time library in Python 3.10 indicates a logarithmical increase of computational time cost from approximately 0.024 to 0.079 seconds at each iteration in the same processor. Calculating the processing time of the first fixed structure and the second larger structure indicates that although the iteration numbers of the smaller network are higher, the incremental time process of the smaller network is not significantly much more than the larger one, while their accuracy is almost the same. Therefore, knowing how much we can increase the network size can reduce computational costs. So, comparing the computational time cost and

convergence speed of fixed and dynamic structures, Figures 21 and 22 indicate that although the computational cost of the proposed strategy is not much lower than the fixed one with 10 hidden neurons, possibly due to integrating the ART section computation costs to the ART-SNN method, the precision of sound localization has increased. The simulated network parameters are given in Table 3.

Table 3

Network Parameters of ART-SNN

Parameter	Value
τ	0.01
A	100
TRAIN	70%
TEST	30%
C_0	342

Table 3 presents the parameter optimization results for the model, specifically focusing on the process of fine-tuning the hyperparameters in order to achieve optimal performance. The optimization was carried out using a trial-and-error approach, where we iterated through different combinations of parameters to determine the most effective configuration. After performing 10 separate runs, the parameters were adjusted based on the performance metrics observed during each iteration, ensuring that the final set of parameters provided the best results for the given dataset.

It is important to highlight that the optimization process is highly sensitive to the conditions of the environment in which the model is applied. Given that different environments can introduce variations such as changes in background noise, acoustic conditions, and the types of sound events being processed, the parameters are often optimized for each specific environment to account for these factors. However, when working with external datasets in this study, we utilized a fixed set of parameters that

had been optimized for the primary dataset, as these parameters performed sufficiently well across the various environments tested.

We compare some well-known and conventional SSL algorithms – namely known energy-based, MUSIC, GCC-PHAT, LS-SVM, and conventional SNN – to better understand their performance in at least five sample examples. Table 4 compares the proposed method with two conventional sound localizing methods for sound source steady-state error averages and the standard deviations for the four mentioned recorded data. Table 4 indicates the superiority of the proposed method in localizing the sound source with only two sensors in comparison to both SNN-based and non-SNN-based approaches.

Table 4

Comparison of Sound Source Localization Methods

Items	Energy - Based	MUSIC	GCC-PHAT	LS-SVM	Conventional SNN	Fixed-RSNN-10 hidden Neurons	ART_rSNN-10 Hidden Neurons
Average error of direction angle (deg)	-	33	27	15	15	17	14
Average error of distance (m)	Depends on the distribution method and random search resolution 1.03 (m) for uniform distribution search and (0.7 to 4m) in other distributions	-	-	-	-	0.45	0.31
Standard deviation of the error	1.2 (m)	15 (deg)	12.02 (deg)	10 (deg)	15 (deg)	0.5 (m)	0.3 (m)
Real-time applicability	Yes in small search area	No	No	No	Yes	No	Yes
Execution Time for each iteration (ms)	32	15	18.8	17.3	49	40.9	32

Furthermore, it seems that sensitivity to the sensors' arrangements in both MUSIC and GCC-PHAT algorithms should have triggered the higher error in sound localizing.

The network proposed in this study utilizes an adaptive approach for sound source localization. It starts with the smallest possible size and gradually grows based on the estimation error. The system employs an event trigger-based approach, utilizing the encoding procedure's threshold. These features enhance the ability of the new architecture to be utilized in event-triggered sound source localization, activating neurons in different positions based on the target path trajectory.

To evaluate the proposed method, a fixed structure network and four other conventional algorithms – energy-based by normal and random search distribution strategy, GCC-PHAT, MUSIC algorithms, and a conventional STDP-based SNN and LSSVM – were investigated. The results demonstrate that the proposed ART-rSNN method achieves higher estimation accuracy and converges to the target location in fewer iterations compared to the fixed structure SNNs and other classic methods.

Table 4 compares various Sound Source Localization (SSL) methods across several performance metrics, including directional accuracy, distance estimation, standard deviation of error, real-time applicability, and execution time. Among the methods tested, the ART_rSNN-10 Hidden Neurons model showed the best performance in terms of average error for direction angle (14°) and distance estimation (0.31 meters). It also exhibited the lowest standard deviation in distance error (0.3 meters), highlighting its consistency. While methods like MUSIC and GCC-PHAT showed competitive directional accuracy, they were not suitable for real-time applications. In contrast, ART_rSNN-10, along with the energy-based method, demonstrated real-time applicability, making it more practical for dynamic environments. Execution times for ART_rSNN-10 were also favourable, comparable to faster methods like MUSIC. Overall, the ART_rSNN-10 Hidden Neurons model outperforms the other approaches, particularly in real-time performance and localization accuracy, while maintaining competitive computational efficiency.

6.4 ART_rSNN Performance Analysis

To assess the proposed ART_rSNN method on the L3DAS22 dataset, we employed two key metrics: Accuracy and Mean Error at 20 degrees (ER_{20}). The results are visually presented in Figure 23. Accuracy, calculated as the percentage of correct predictions among the total, serves as a comprehensive indicator of the system’s overall

performance. A higher accuracy percentage signifies better alignment between predicted and true sound source locations. The accompanying chart also illustrates the Mean Error at 20 degrees, providing insights into the average angular deviation between predicted and true angles. This metric offers a nuanced evaluation, emphasizing the system's accuracy specifically at the critical angle of 20 degrees. The bar chart collectively provides a comprehensive view of our sound localization system's effectiveness, facilitating interpretation and comparison under various conditions.

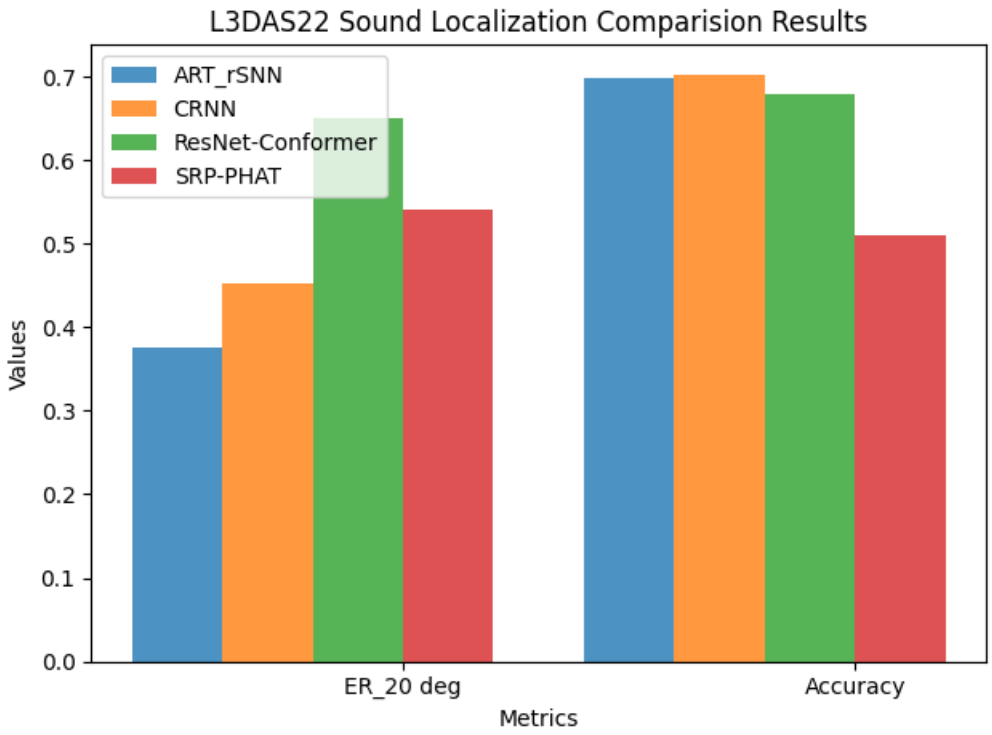
In Fig. 23, we present a comparative analysis of sound localization results achieved by the CRNN, ART-rSNN, ResNet-Conformer, and SRP-PHAT methods. The plot showcases the mean error at 20 degrees for each method and their corresponding accuracy values. Notably, our proposed method, ART-rSNN, exhibits a lower mean error at 20 degrees compared to the other methods, indicating superior performance in terms of localization precision. The accuracy of our method stands out, showing results nearly identical to CRNN, with only a marginal 0.01 decrease in accuracy compared to CRNN.

Furthermore, our method outperforms ResNet-Conformer and SRP-PHAT in accuracy. These outcomes affirm the efficacy of incorporating Mag features in the L3DASS dataset for sound localization. Specifically, our method achieves an accuracy of 69.8%, slightly below the 70.3% accuracy achieved by CRNN, while maintaining a notable advantage in computational efficiency.

The proposed ART-rSNN method demonstrates a calculation time, of approximately one-tenth that of deep learning methods like CRNN and ResNet-Conformer. This significant reduction in computation time not only attests to the computational efficiency of our approach but also positions it as a promising solution for real-time applications where speed is crucial. In summary, the results presented in Figure 23 underscore the favourable trade-off between accuracy and computational efficiency offered by our ART-rSNN method when compared to existing state-of-the-art techniques in sound localization.

Figure 23

Mean Error and Accuracy Comparison on L3DAS22 in 2D Sound Source Localization Task



The ART_rSNN method, a novel architecture designed and evaluated in this thesis, has been published in the IEEE ACCESS journal by Roozbehi et al., 2024. This method has been a significant contribution to the field of sound localization and source separation, as demonstrated in the results presented in this thesis.

6.5 Classification Module

To evaluate the classification capacity of the new designed structure, first we compare the generated output of the observed neurons and the measured input, as indicated in Figure 24. The left block in Figure 24 depicts firing output of the observed neuron and the right block in Figure 24 indicates the input measured power of a sound source.

Figure 24

Input and output of an observed neuron potential

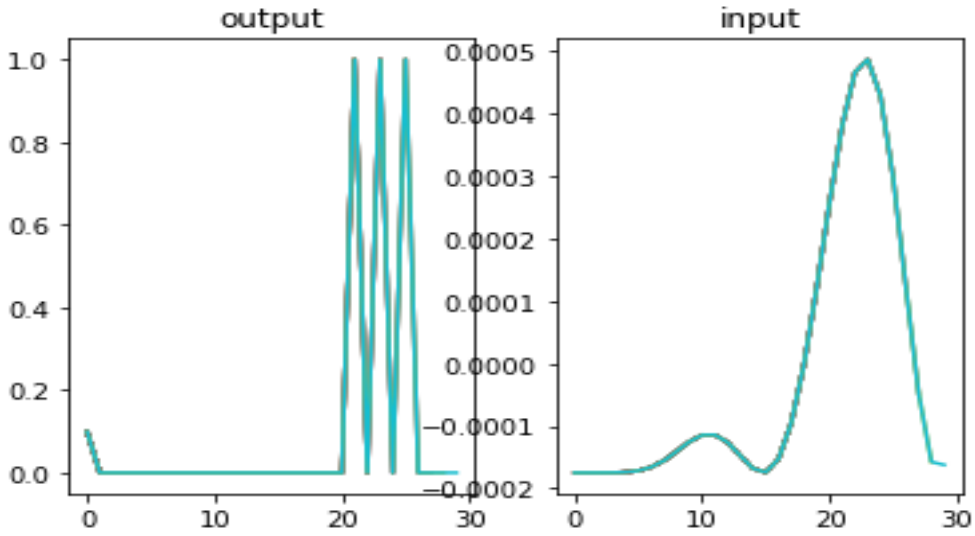


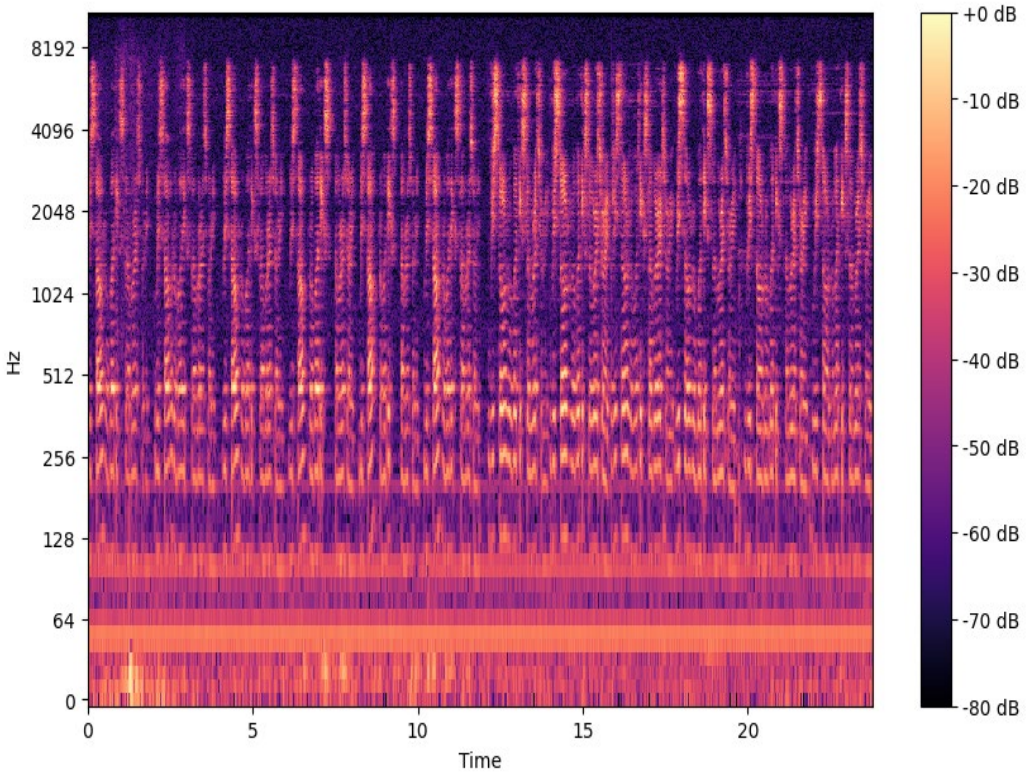
Figure 24 indicates that firing rate of the output neuron corresponds to the measured input signal. The left figure depicts firing output of the observed neuron and the right figure indicates the input measured power of a sound source.

6.6 Multiple Sound Source Time-Frequency Spectrogram Features

This section conducts a thorough examination of the spectrogram characteristics of three mixed sound source signals. The examination is predicated on the magnitudes of the spectrograms corresponding to the signals, which are derived by implementing a logarithmic transformation to the magnitude spectrogram of the signal within a time-frequency domain framework. The principal objective of this examination is to discern distinct patterns inherent within the signals that may facilitate the disaggregation of individual sound sources. Figures 25-27 illustrate the spectrogram analysis of the three samples, emphasizing the identified patterns and furnishing a visual representation of the manifestation of mixed signals within the frequency spectrum. These figures play an integral role in elucidating the intricacies of sound signal interactions, thereby augmenting our comprehension of the fundamental acoustic phenomena. By presenting these spectrograms, we not only validate our analysis but also offer a visual abstract that assists in elucidating the pertinence of funding and antecedent discussions prior to the implementation of AI methodologies for subsequent analysis. This visual context is indispensable for readers to apprehend the subtleties of signal mixing and its ramifications for source separation.

Figure 25

Spectrogram of Three Speakers Sound Sources



As depicted in Figure 25, the configurations of the three auditory sources, intermingled with three instances of human vocalization, exhibit a congruent conceptual framework while occupying distinct spatial positions. A majority of frequency bands reveal a similar configuration; however, within the frequency range of 512-4069 Hz, there exists a discrepancy in power levels as analysed over the temporal domain. More specifically, fluctuations in power are observed during the initial (0-12) sample interval in contrast to the intervals extending beyond 12 samples. In this context, the objective of the presentation is to underscore that, although three instances of human speech conveying identical content from disparate locations display analogous patterns, the differentiation between these patterns and classifications proves to be a formidable challenge. Despite the patterns being nearly identifiable, the quantity of sound sources remains visually indeterminate; it is only through the analysis of power levels that one can infer a shift in the location of the sound source. As a result, our investigation indicates that the detection of multiple sound events may not achieve visual efficacy in this instance, due to the overlapping characteristics of the signals which obstruct the

unambiguous identification of individual sources. Therefore, our proposed method can demonstrate its potential efficiency even in challenging conditions, similar to this case..

Figure 26

Spectrogram of Four Sound Events in Succession

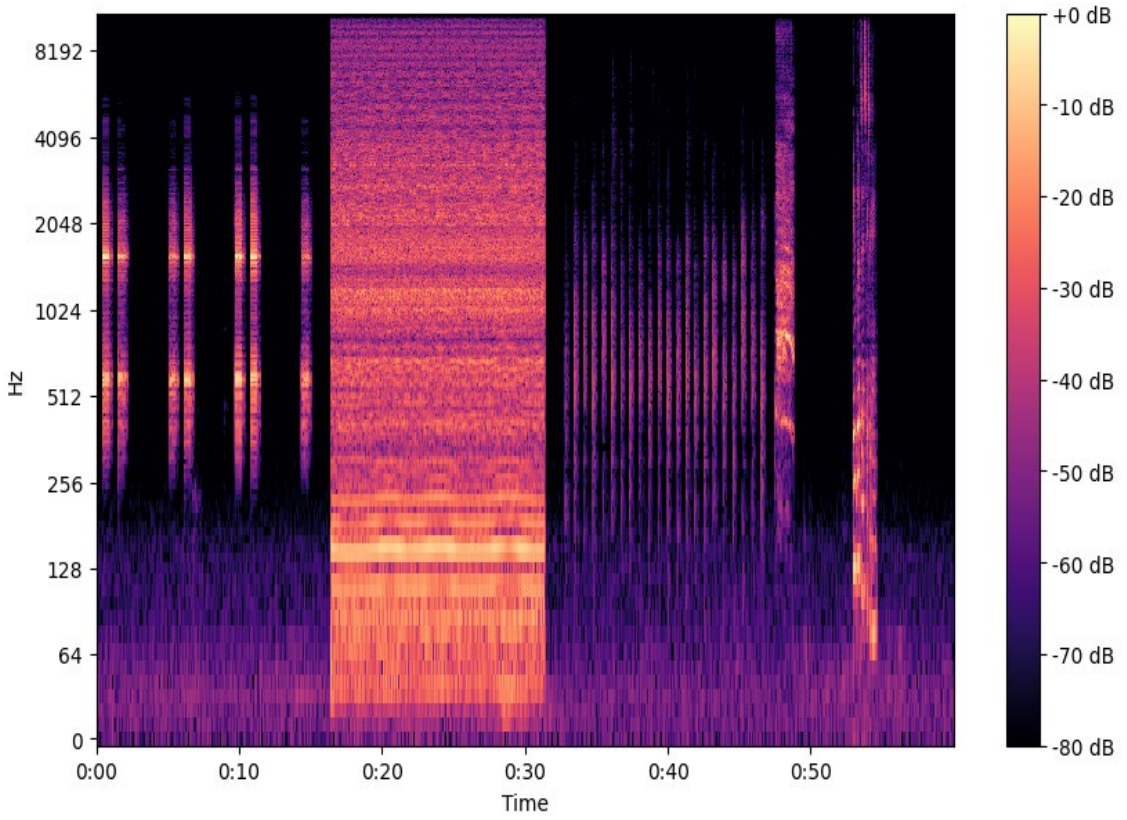


Figure 26 demonstrates that corresponding to four sound events in succession—a ringing bell, a machine, a man walking, and shouting—four distinct frequency-time-power patterns occur within every 15-time sample period. In contrast to Figure 25, which illustrates overlapping sound sources with similar patterns that complicate source separation, Figure 26 depicts clearly separable sound events. This clarity allows for easier identification of the classes and the number of sound sources present. The distinct frequency signatures and power levels associated with each event enhance our ability to differentiate between them, making it evident that these sounds can be analyzed independently. Thus, the representation in Figure 26 underscores the potential for effective sound event detection when the underlying patterns are sufficiently distinct.

Figure 27

Spectrogram of Multiple Sound Events

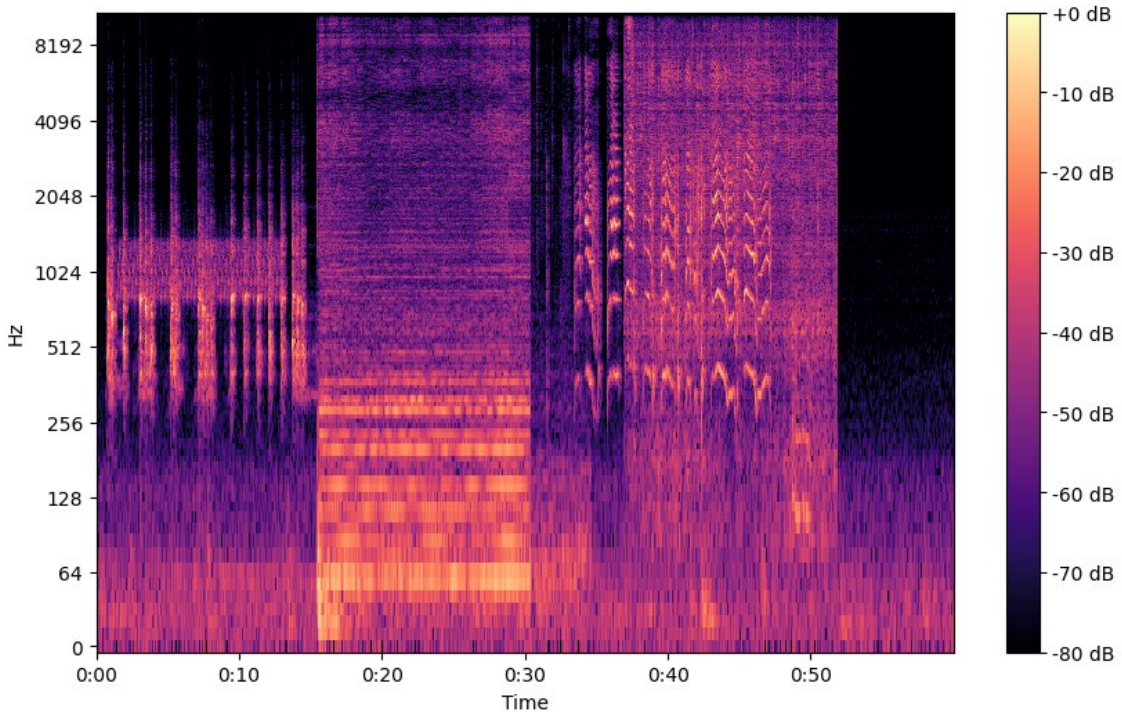


Figure 27 illustrates multiple simultaneous sound events, including two car alarms, a dog barking, machine noise, a baby crying, and washing sounds, alongside periods of silence. Notably, during the initial 0 to 15 seconds, two distinct sound patterns are observed in the time domain. Following this period, the machine sound pattern becomes more pronounced, while the baby crying and washing sounds emerge as mixed events after 40 seconds. The figure also reveals that noise power is predominantly concentrated in the frequency band of 0 to 256 Hz, with its strength remaining below -50 dB. To enhance sound event detection, it is advisable to implement a threshold of -50 dB. However, relying solely on time domain features may raise concerns regarding their effectiveness in accurately distinguishing between overlapping sounds. To address this issue, Figures 28-30 provide a detailed examination of the intensity of each sound signal in the time domain, allowing for a more comprehensive analysis of their characteristics when considered individually.

Figure 28

Ringing Bell Sample in time domain

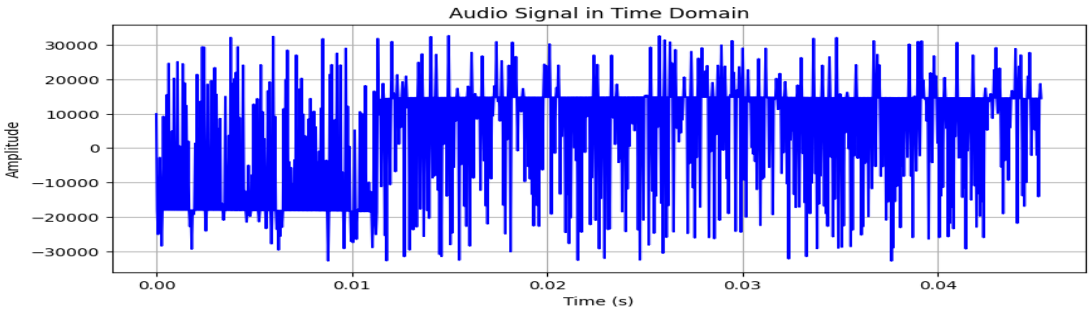


Figure 29

Machine Sound Sample in time domain

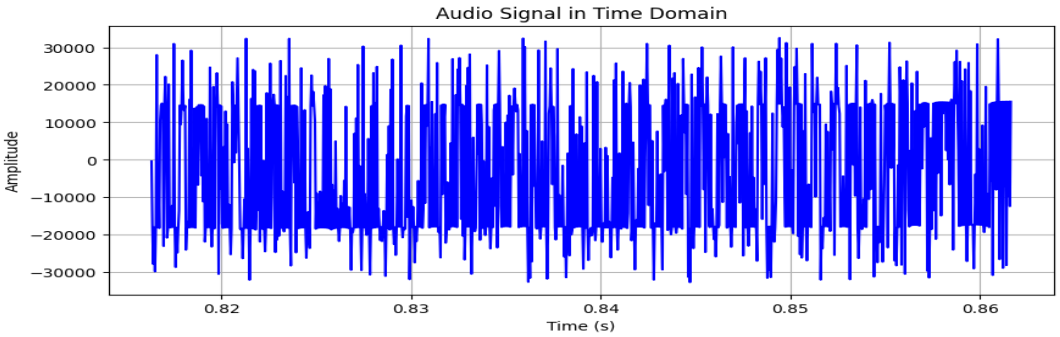
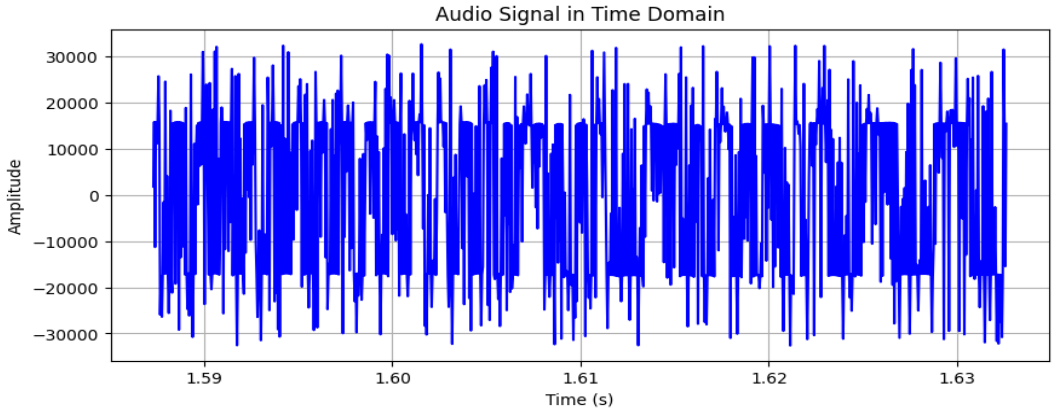


Figure 30

Baby Crying Sound Intensity in time domain



Figures 28-30 depict three instances of auditory events represented as time-domain waveforms, facilitating a visual examination of disparate acoustic patterns. Each waveform demonstrates distinct attributes, encompassing differences in drift and oscillatory frequency.

To emphasize the salient features of each auditory event within the temporal waveforms, Table 5 provides a comprehensive summary of essential characteristics for each acoustic pattern.

This summary encompasses metrics such as mean, variance, root mean square (RMS), and the duration of the signals. These features collectively provide insights into the distinctiveness of each sound pattern, aiding in their classification and analysis. They help us capture both the central tendency and variability in the data, as well as assess the accuracy of our models. To provide clarity, we have summarized these statistical time domain features in Table 5, which can be valuable for sound classification tasks.

Table 5

Time Domain Feature Analysis of Three Sound Signal Patterns

Sound signal	**Time duration (sec)**	**Mean**	**Variance**	**RMS**
Ringing bell	1.14	-4.46815e-05	5.11933e-08	0.000231
Machine	1.14	6.037052e-06	5.32732e-08	0.000231
Baby crying	1.14	-0.000121	6.732694e-08	0.000286

In Table 5, it is evident that the mean values of the signals analysed with the same window time differ significantly. This variation in means suggests distinct central tendencies in the three signals under consideration. Additionally, the variance values exhibit slight differences, indicating variations in the spread or dispersion of data points. Interestingly, the RMS values show a remarkable similarity, with precision up to six decimal places, between the three signals. This implies that despite differences in mean

and variance, the accuracy of predictions or models based on these signals is closely aligned.

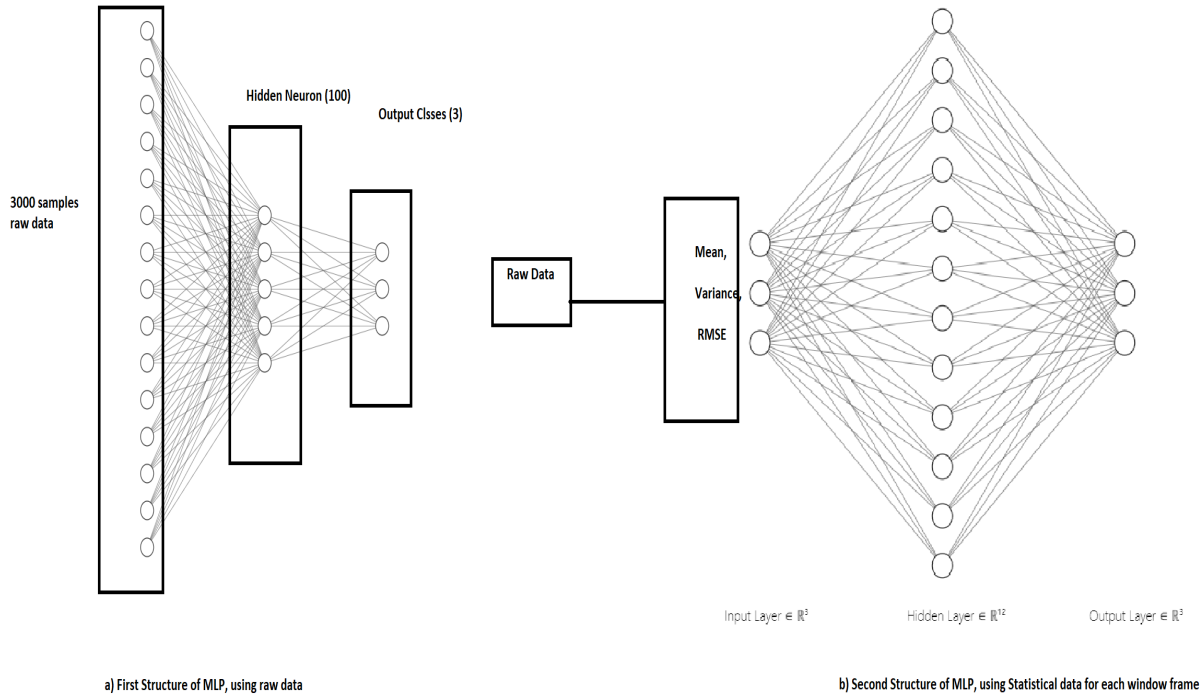
Moving forward, our analysis will encompass single sound source classification using machine learning techniques such as multilayer perceptron (MLP), convolutional neural networks (CNN), and recurrent neural networks (RNN). Subsequently, we will delve into the analysis of mixed data, further exploring the intricacies of our dataset.

6.7 MLP in a Single Sound Source Classification Approach

The utilized MLP classification analysis has been conducted in two structures. The first structure has been utilized for raw data analysis with a time window of 3 seconds. The second structure receives feature inputs of signal. The classification accuracy is shown in Table 5. The overall structure of the MLP utilized in this study is depicted in Figure 31. The two structures referred to in Figure 30 are (a) using raw data with 3,000 sample data (left side) and (b) using statistical features of 1,000 samples (right side).

Figure 31

Two structures of MLP Classifier



Both structures of MLP constructed of one hidden layer with 100 Relu neurons. The initial structure employs a dataset comprising 7,000 samples, each representing a distinct sound class signal. This dataset is divided equally, with 50% allocated for training purposes and the remaining 50% serving as a test dataset. Figures 33 and 34 illustrate the weights of the MLP classifier.

While we posit that the optimization of MLPs could potentially augment their efficacy for specific temporal tasks, our primary objective herein was to position MLPs as a foundational benchmark for ANNs within the domain of Sound source classification, rather than as an endeavour to maximize the capabilities of MLP architectures for SSL applications. This section primarily focuses on comparing our SNN with state-of-the-art (SOTA) models in SSLs, including CRNN, which have been carefully crafted to perform exceptionally well in these targeted areas. CRNNs epitomize an example of the current SOTA in Sound Source detection and classification, able to be superseded MLP variants, even in their most refined iterations.

Moreover, our research accentuates the aspect of explainability inherent to SNNs. We undertook experiments utilizing MLPs and furnished weight spectrograms to demonstrate the adequacy of training and to assess interpretability, which constitutes a pivotal criterion in our investigation. While CRNNs and other SOTA of deep learning methods models predominantly function as black boxes, thereby constraining interpretability, SNNs provide a notable degree of transparency. For example, the spike patterns or denser neuronal populations in SNNs can be directly linked to predictions concerning target locations, thereby rendering them particularly advantageous in an era that increasingly emphasizes model explainability.

In the discourse surrounding the SOTA of MLPs in SSL applications, it is imperative to recognize that although MLPs have been employed for a myriad of classification tasks, they frequently encounter challenges in addressing the temporal dependencies that are intrinsic to audio signals. Contemporary research indicates that deep learning architectures such as CRNNs significantly surpass traditional MLPs by proficiently capturing both spatial and temporal characteristics from audio data (Khan. et al,2022).

Figure 32

Weights of Input to Hidden Neurons in MLP (Structure a)

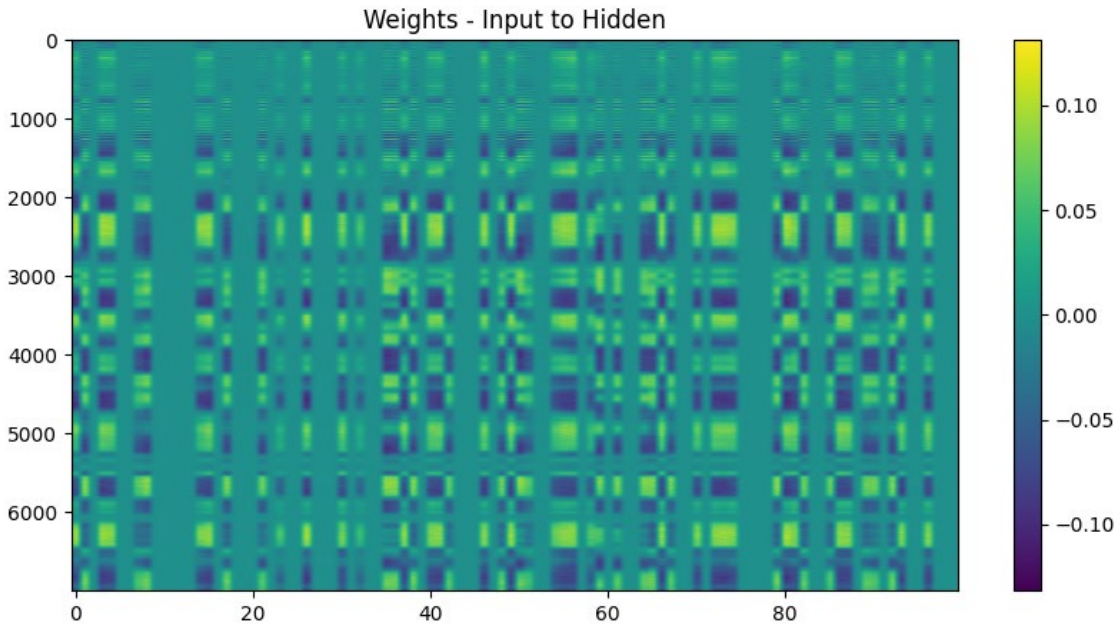
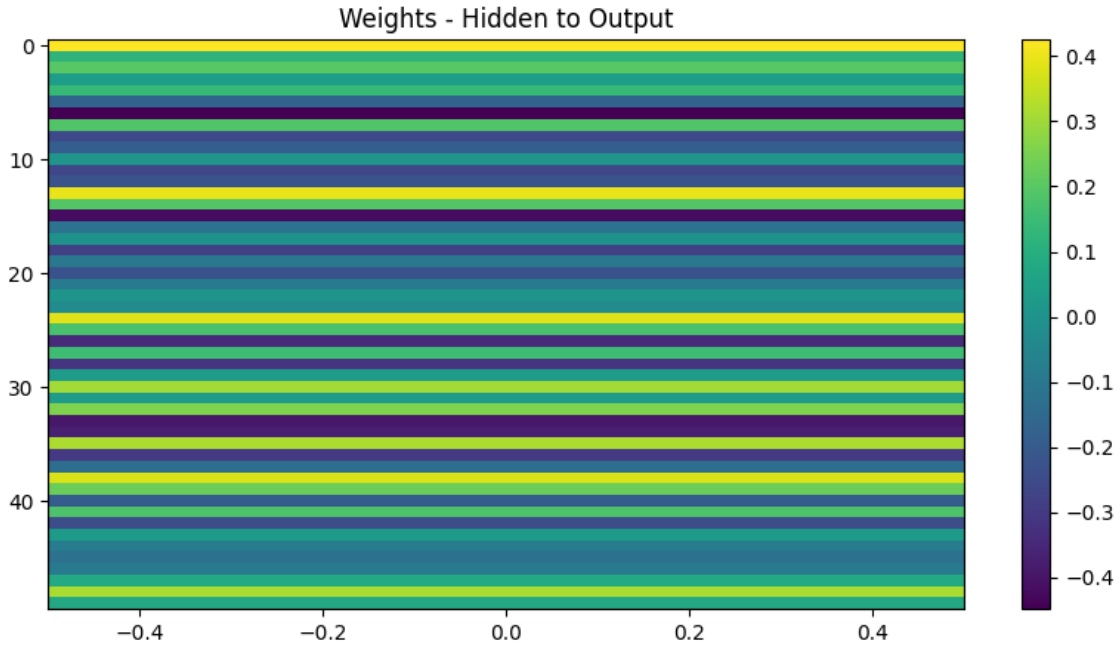


Figure 33

Weights of Hidden to Output Neurons in MLP (Structure a)



In structure (a), the classification accuracy is 50%. As for structure (a), Figures 34 and 35 display its respective weights. The results demonstrate that when using 12 sample data points for training and 6 sample data points for output classes, the accuracy achieved is also 50%.

Figure 34

Weights of Input to Hidden Neurons in Structure a

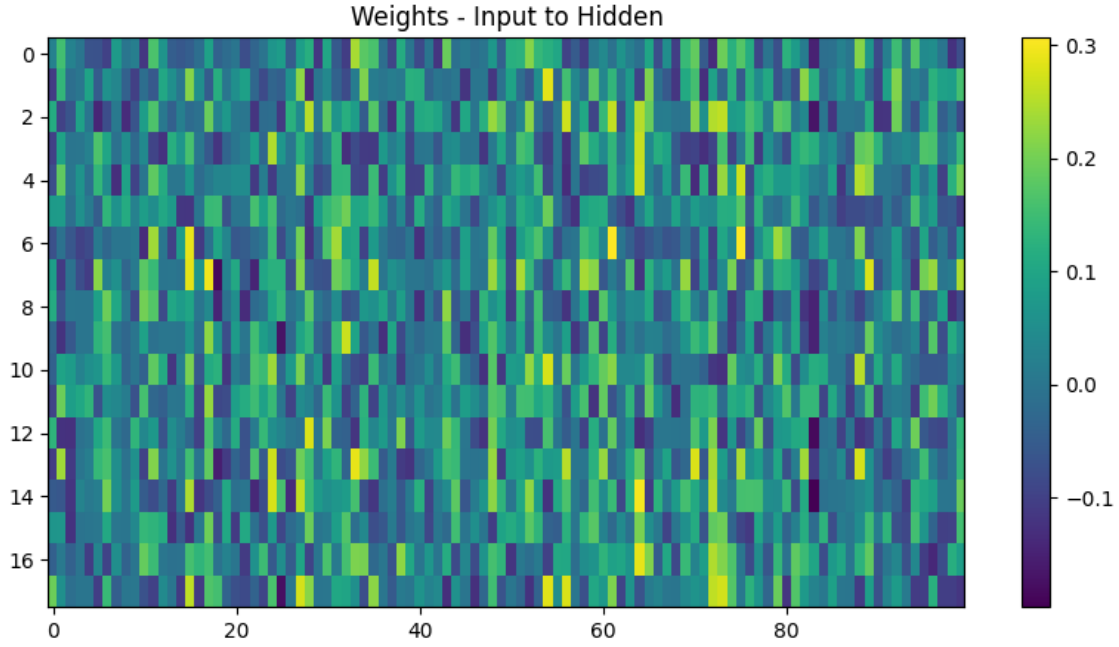
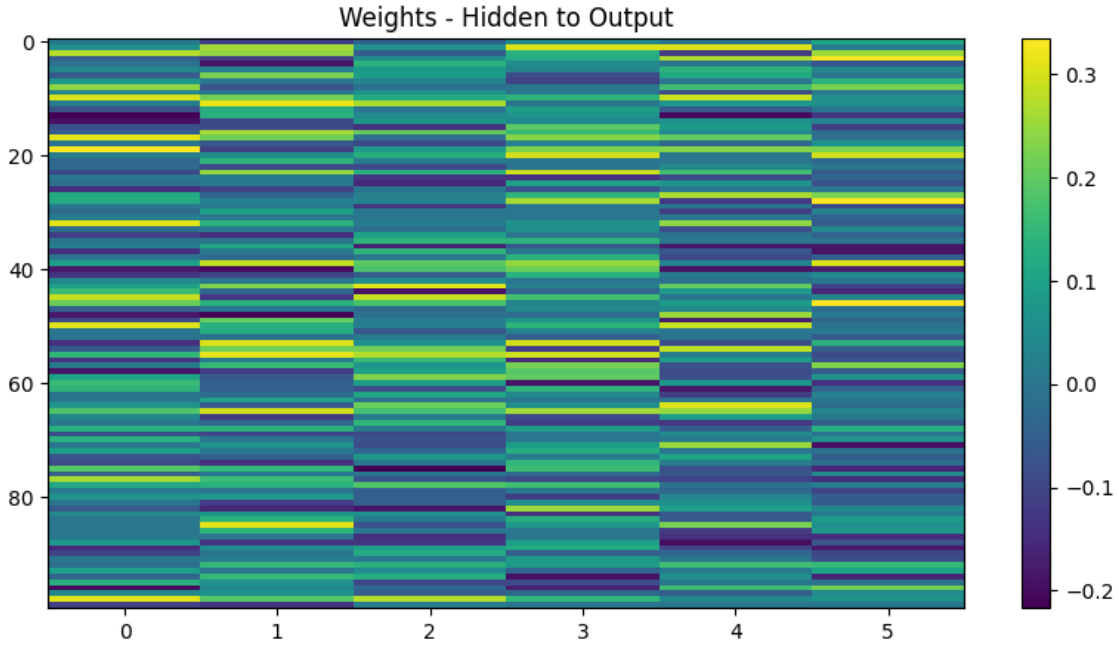


Figure 35

Weights of Hidden to Output Neurons Classification Model with 6 Output Tests



The results in this case indicate that achieving high accuracy with MLP classification for time-domain features and raw data in our dataset is challenging.

6.8 Recurrent Neural Networks for the Single Sound Classification

Recurrent neural networks (RNNs) are a type of deep learning that can model sequences and are capable of modelling the sequential dynamics of acoustic scene signals. They have been used for various audio classification tasks such as audio scene classification, multimedia event detection, and sound classification (Scarpiniti et al., 2021; Phan et al., 2017).

RNNs have been shown to be effective in capturing the temporal dynamics of audio signals and can be used to classify simple sound signals into different classes (Gimeno et al., 2020). The foundation of our classification task lies in the input signals, which are represented as raw audio waveforms. These waveforms are sourced from the dataset detailed in the Zenodo repository (Trowitzsch et al., 2019).

Each input signal comprises 500 samples, and we have a total of 20 such signals from categories of baby and dog sound. These 500-sample-long raw audio waveforms offer a high-resolution depiction of the sound, encompassing its nuances and variations over time. Initially, we attempted to employ traditional time domain features to classify these sound signals. However, despite their widespread use in audio analysis, time domain features fell short in delivering the desired classification accuracy. In our experiments, the accuracy of the model on the test data hovered around 33%. This outcome underscored the limitations of relying solely on time domain features for single sound source classification.

To better comprehend our dataset and the challenges we faced, we visualized the input signals. In Figures 36 and 37 we present a selection of these input signals. Each subplot represents a different sound signal, showcasing its amplitude variations across the 500 samples. The complex patterns and fluctuations within the waveforms emphasize the need for a model that can effectively extract and learn from this temporal information.

Figure 36

Train Audio Signals

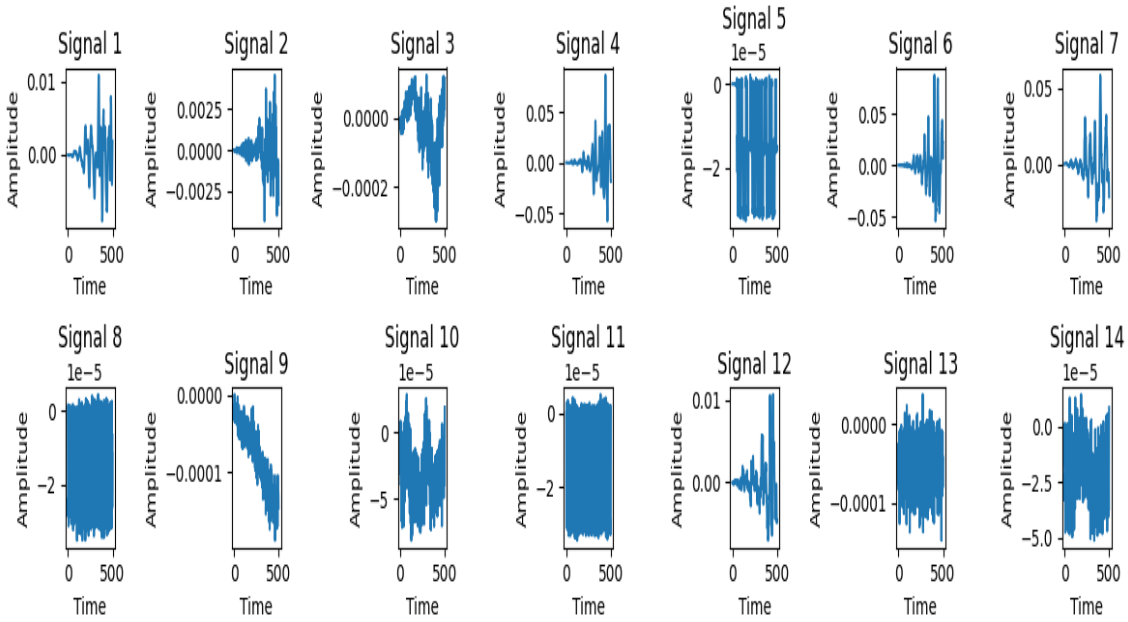
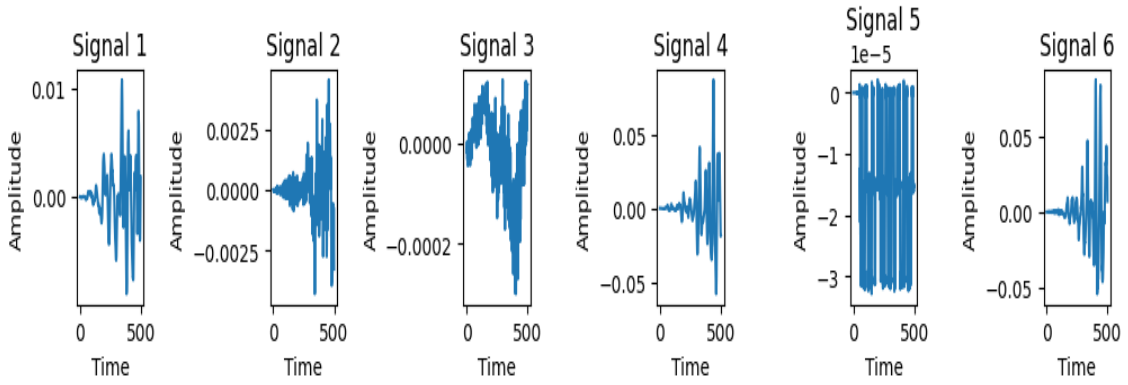


Figure 37

Test Audio Signals



6.8.1 RNN Model Architecture

The utilized RNN model architecture is designed to process these 500-sample input sequences and make binary classifications (baby or dog). The model consists of an RNN layer as its core component, followed by a series of dense layers for classification.

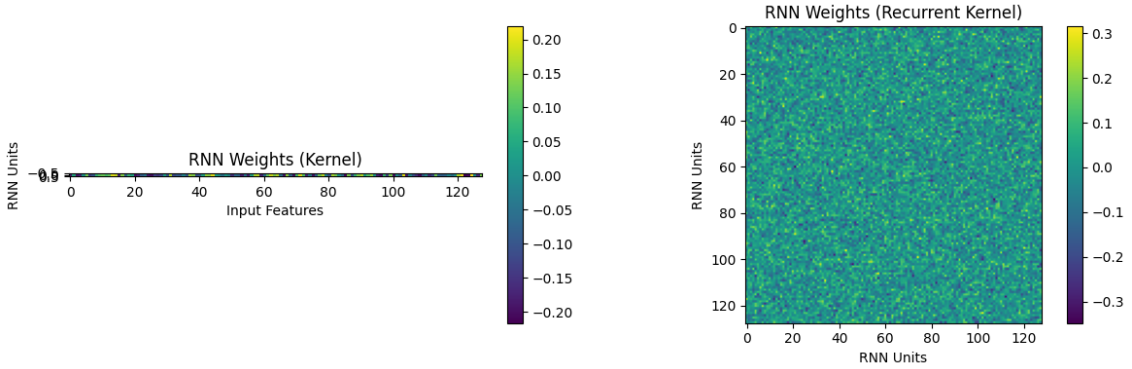
6.8.2 Visualization of RNN Weights

To gain insight into how the RNN model processes these input signals, we visualized the RNN layer's weights. Figure 38 presents two weight matrices: the kernel weights (representing input connections) and the recurrent kernel weights (representing

recurrent connections). These weights capture the learned patterns and relationships within the input signals, crucial for making accurate classifications.

Figure 38

RNN Weights



The overall structure of utilized RNN classifier is indicated in the Figure 38.

Figure 39

The Overall Structure of the RNN Classifier

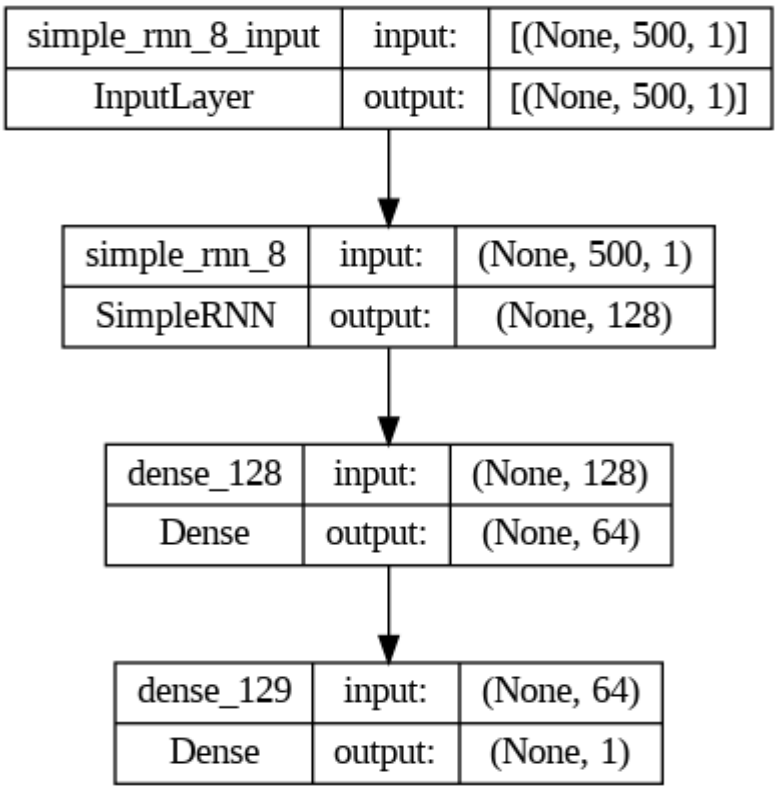


Figure 39 provides a detailed description of the RNN layers utilized in this section. The RNN model consists of three layers: the input layer, Simple RNN layers, and the output layer. The input layer receives the raw audio data, and the input shape is defined as

(batch_size, sequence_length, num_features), where batch_size is the number of samples processed in each training batch, sequence_length denotes the length of the input sequence, and num_features represents the number of features extracted from the audio data at each time step. The Simple RNN layers are fundamental building blocks of RNNs that introduce recurrent connections to capture the temporal dependencies in the audio data. The output layer generates predictions, and in sound classification tasks it typically consists of one or more neurons with a softmax activation function to produce probability distributions over the target classes.

The RNN layers, through their recurrent connections, enable the network to consider the entire sequence of audio data when making predictions, which is crucial for sound classification as it captures the temporal characteristics of different sound events. However, training RNNs can be challenging due to issues like vanishing gradients. In practice, other variants of RNNs, such as long short-term memory (LSTM) or gated recurrent unit (GRU), and integration of convolutional neural network and recurrent neural network (CRNN) are often preferred for sound classification tasks as they have improved mechanisms for handling long-range dependencies in the data.

It is also crucial to stress that traditional RNNs have certain limitations, especially in the context of the vanishing gradient problem. The combination of RNNs and CNNs has led to the emergence of CRNNs. These CRNN architectures have shown significant improvements toward sequential data processing as well as the capture of their temporal characteristics, and thus their high accuracy in the tasks of auditory classification.

In the following section, we perform a comparative study of our proposed method compared with a CRNN, which can be seen as an example of a state-of-the-art model. The CRNN architecture used as a baseline includes a two-stage LSTM network architecture. This design of the architecture is crucial since it can help the model to better absorb complex temporal dependencies inherent in acoustic data for effective sound classification. This comparison analysis is described in Table 7, where we report the performance measures of both approaches. The SNN method is carefully designed to make use of the time evolution that is present in auditory data and yet to alleviate the difficulties raised by conventional RNN paradigms. Focusing on the strengths of

SNNs in temporal information processing, we aim to demonstrate the place of our method with respect to existing CRNN architectures.

6.9 CRNN Classification Using Time Domain Features

The model begins with a series of convolutional layers that learn spatial features from the input data. These layers use small filters to convolve over the input audio signals, capturing local patterns in the data. After each convolutional layer, max-pooling layers are employed to downsample the feature maps, reducing the spatial dimensions while retaining the most important information. Batch normalization layers are inserted to stabilize and accelerate training by normalizing the activations of the previous layers. Following the convolutional and pooling layers, recurrent layers are added.

In this section, two LSTM (long short-term memory) layers are utilized. LSTMs are a type of RNN designed to capture long-range dependencies in sequential data. The model concludes with dense layers that perform the final classification. The number of neurons in the output layer corresponds to the number of classes (sound sources) in the dataset, and the softmax activation function is applied for multi-class classification. The dataset used in this experiment comprises audio recordings of two sound sources: baby and dog. A total of 20 audio signals were collected, with 10 samples from each category. The audio signals were pre-processed by extracting mean, variance, and RMS (root mean square) features, which were used as input features for the CRNN model. The model was trained on a portion of the dataset, with a split of 80% for training and 20% for validation. Training was conducted for 10 epochs using a batch size of 32. The model was compiled with the Adam optimizer and the sparse categorical cross-entropy loss function. After training, the model's performance was evaluated on the remaining 20% of the dataset. The accuracy metric was used to assess the model's ability to correctly classify sound sources. The CRNN model achieved an accuracy of 33% on the test dataset and 81% accuracy in the training dataset, indicating the time domain features, such as mean, variance, and RMS in sound source classification by CRNN are inefficient. Time delta is considered a constant value 0.05. The structure of the CRNN model is indicated in Figure 40.

This hybrid architecture is a modified version of the method, established by (Muqing Deng. et al,2020). This model is particularly well-suited for audio classification tasks due to

its ability to capture both temporal and spectral features of audio signals. Relying solely on time domain features may not be a feasible approach for this structure. Our evaluations have confirmed this notion, especially when dealing with small datasets. To address this limitation, we have incorporated mel-frequency cepstral coefficients (MFCCs) as a feature representation to assess the effectiveness of this structure in sound classification applications. It is worth noting that, while this example focuses on batch processing, the proposed method in this thesis leverages online data sampling as a core component of its processing pipeline.

Using time-domain features alone may not fully harness the potential of sound classification tasks. In our initial experiments with a dataset of 500 samples, we observed that the accuracy remained at a modest 33%. This result indicated that relying solely on temporal characteristics of audio signals did not yield satisfactory outcomes. To improve the performance of our sound classification model, we decided to increase the dataset size by doubling it to 1,000 samples. This adjustment had a significant impact, as we observed a notable increase in accuracy, reaching up to 66%. This substantial improvement reaffirmed the importance of having a sufficiently large and diverse dataset for training our hybrid architecture. Figures 41 and 42 indicate MFCCS features of the utilized dataset.

While the CRNN architecture utilizing MFCC inputs may exhibit improved performance relative to our proposed methodology within this dataset, it is crucial to emphasize that our approach adeptly manages raw time-domain signals—an intricate challenge inherent to sound classification. The empirical findings reveal that, even when constrained to time-domain features, our methodology attains outcomes that are only marginally inferior to those of the CRNN employing MFCCs. This observation signifies a considerable potential for our approach, especially in the context of temporal signal classification tasks, implying that it can contend effectively with leading methodologies in this domain. The findings accentuate the versatility of our SNN-based approach for time-domain data, outstripping the CRNN when restricted to sequential temporal inputs. In forthcoming research endeavors, enhancing our methodology with designs that integrate frequency-time domain features could significantly bolster its efficacy and competitiveness against established methods, such as CRNNs utilizing MFCCs.

Figure 40

Layers of Utilized CRNN in the Single Sound Classification Application

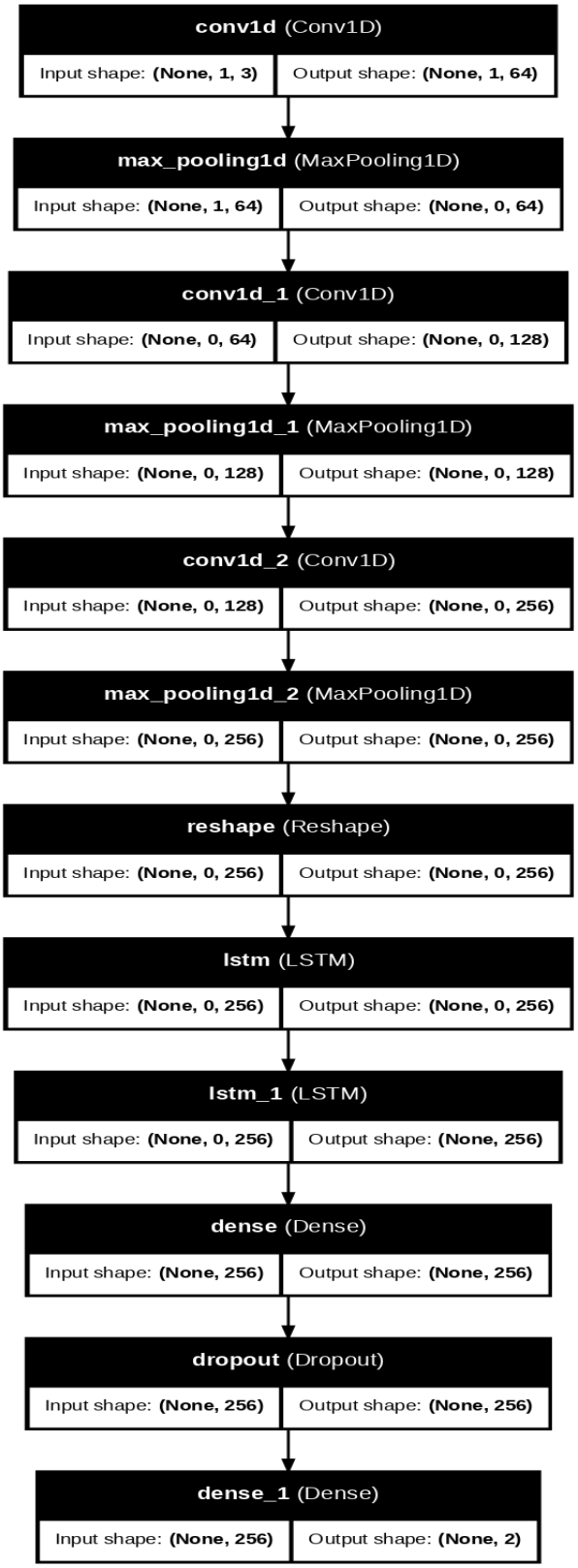


Figure 41

MFCCs Features of the Training Dataset

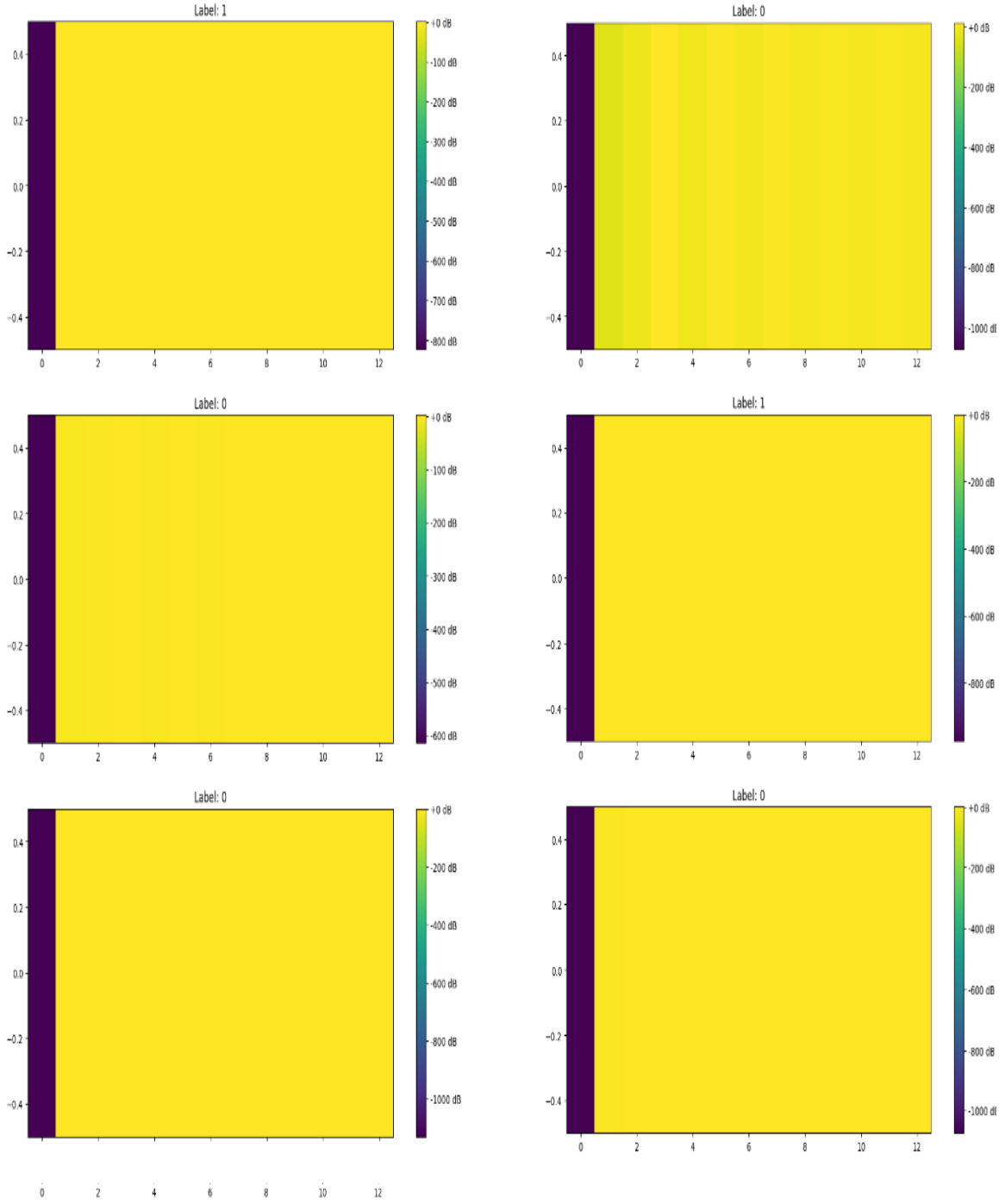
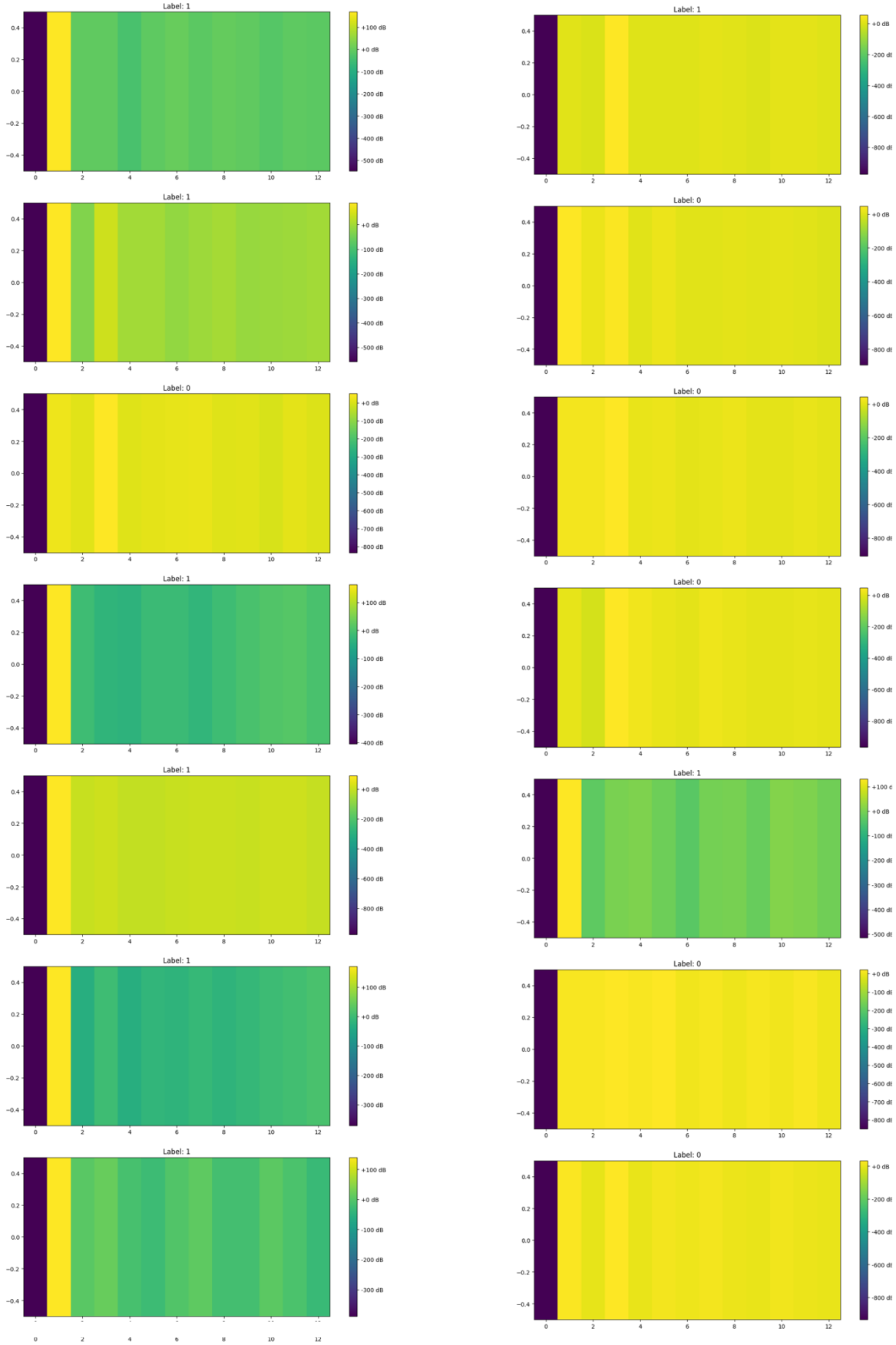


Figure 42

MFCCs Features of the Training Dataset



By incorporating MFCCs to consider both frequency and time domain features and increasing the dataset size, we were able to leverage the strengths of a CRNN architecture effectively. This approach showcases the significance of feature selection and dataset size in achieving better results in sound classification tasks.

6.10 Tempotron SNN Classifier

In this section, we showcase the performance of the tempotron SNN structure to illustrate visual perception related to tempotron efficacy and to compare two SNN models used for sound classification: a traditional tempotron SNN and our enhanced tempotron SNN structure. This comparison aims to underscore the advantages of our modified SNN model (which integrates the tempotron mechanism with STDP); as a result, it can more effectively capture temporal patterns within the raw time-domain signal. The tempotron model, functioning as a fundamental spiking classifier, offers a valuable benchmark because it exemplifies a conventional method of spike-based classification.

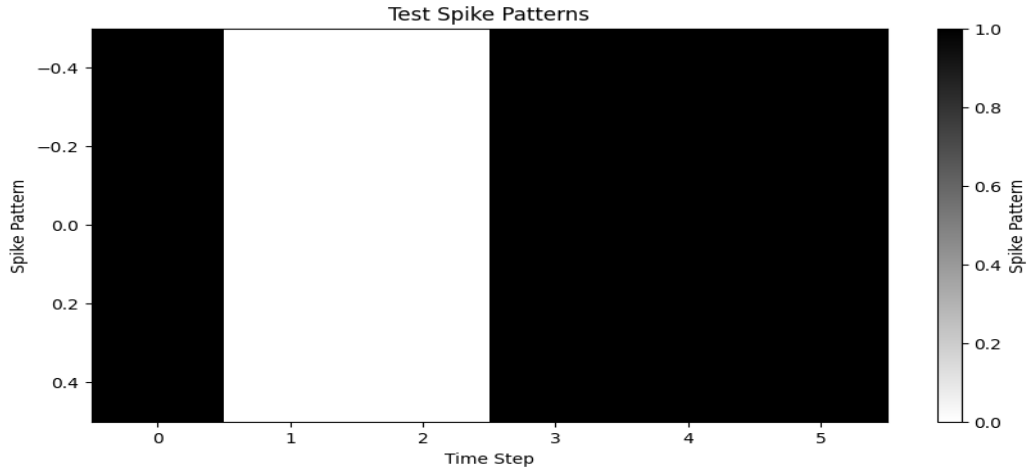
The tempotron SNN is a simplified neural network model inspired by the biological behaviour of neurons. It is used for binary classification tasks, where it learns to discriminate between two classes of data. The tempotron SNN model is based on a simplified type of neuron, often referred to as a "tempotron" or "threshold neuron".

It is inspired by the integrate-and-fire neuron model, where the neuron accumulates incoming signals (integration) and fires (produces an output) when the accumulated signal reaches a certain threshold. As described in the previous chapter, the tempotron SNN uses a simple learning law that adjusts the weights of its input connections during training. The tempotron SNN is primarily used for binary classification tasks, where it learns to distinguish between two classes (Class 0 and Class 1). It does this by finding a decision boundary in the input space defined by its weights and threshold.

The simple tempotron SNN classifier, consisting of a single tempotron layer, achieves an accuracy of 66.67% for the best random test example of classification in the test dataset. However, the average accuracy across different runs indicates an average accuracy of 52% for the test dataset and 48% for the training dataset, both of which consist of 1,000 samples. To provide further clarity, the single-layer output is depicted in Figure 43.

Figure 43

Output Spike Patterns of a Single Layer Tempotron SNN Classifier



By introducing a hidden layer consisting of 50 neurons, the reliability of the classifier model was significantly improved. This means that the accuracy and performance of the model were enhanced. Specifically, when evaluating the model using the test dataset, the hidden layer consistently achieved a remarkable accuracy of 66.67%. It is worth noting that this accuracy remained consistent regardless of the initial weights or the random arrangement of the dataset. This is a significant improvement that demonstrates the robustness and stability of the classifier when the hidden layer is included.

The structure of the classifier model follows a specific principle known as the tempotron-STDP law. This law governs the behaviour and interactions of neurons within the network. In this case, the aim of the classifier model is to accurately classify single sound sources. The implementation involves using LIF (leaky integrate-and-fire) neuron models in the hidden layer, which help capture important features of the input data. The output layer utilizes a tempotron neuron, which is responsible for making the final classification decision.

It's important to mention that this network structure is fixed, meaning that the number of layers and the connections between them are predetermined. The data flows in a unidirectional manner through the network, starting from the input layer, passing through the hidden layer, and culminating in the output layer. This architecture is commonly referred to as an integration of recurrent network for hidden layer connected to the tempotron layer as the output layer. The learning law, set at 0.1, plays a vital role

in adjusting the weights and biases of the neurons during the training phase. It governs how the network learns from the input data and improves its classification performance over time. A learning law of 0.1 indicates a specific learning rate, which determines the magnitude of weight adjustments based on the feedback received during training.

To further visualize and understand the behaviour of the network, the spike patterns of the train samples and test samples are depicted in Figures 43 and 44, respectively. These figures help illustrate the activation patterns and temporal dynamics of the neurons in response to the input samples.

Overall, the addition of the hidden layer significantly enhances the reliability of the classifier model, as indicated by its consistent accuracy and adherence to the tempotron-STDP law. The fixed network structure, forward architecture, learning law of 0.1, and spike pattern visualizations all contribute to a comprehensive understanding of the classifier model's performance.

Figure 44

Hidden Neurons Mean of Spike Patterns for Each Neurons and Input Sample Data of Training Dataset

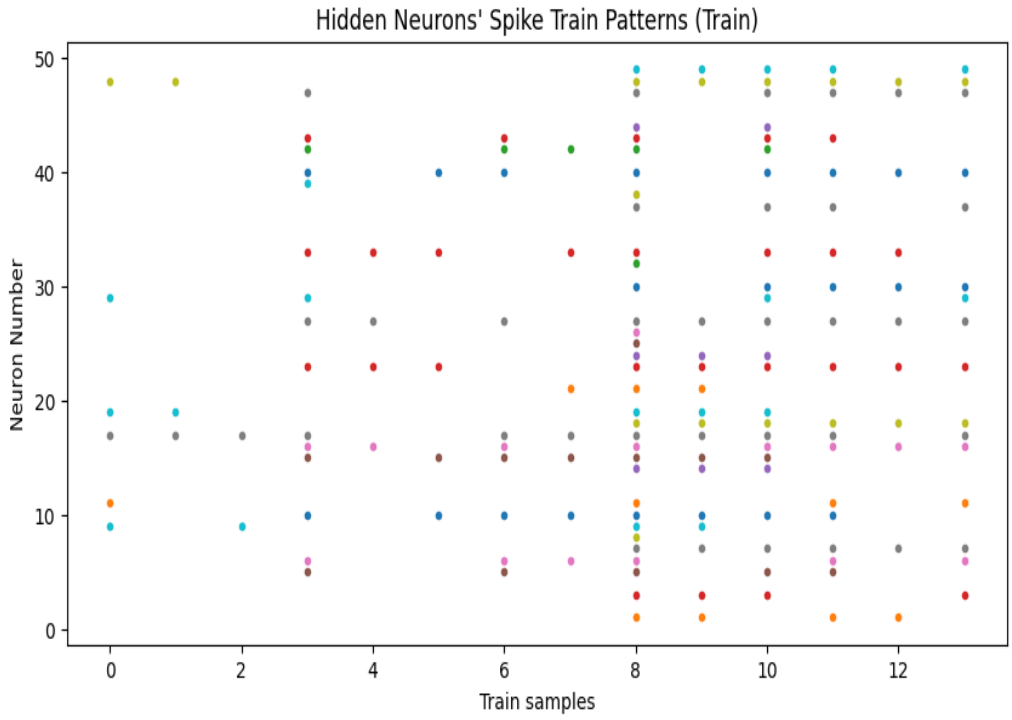
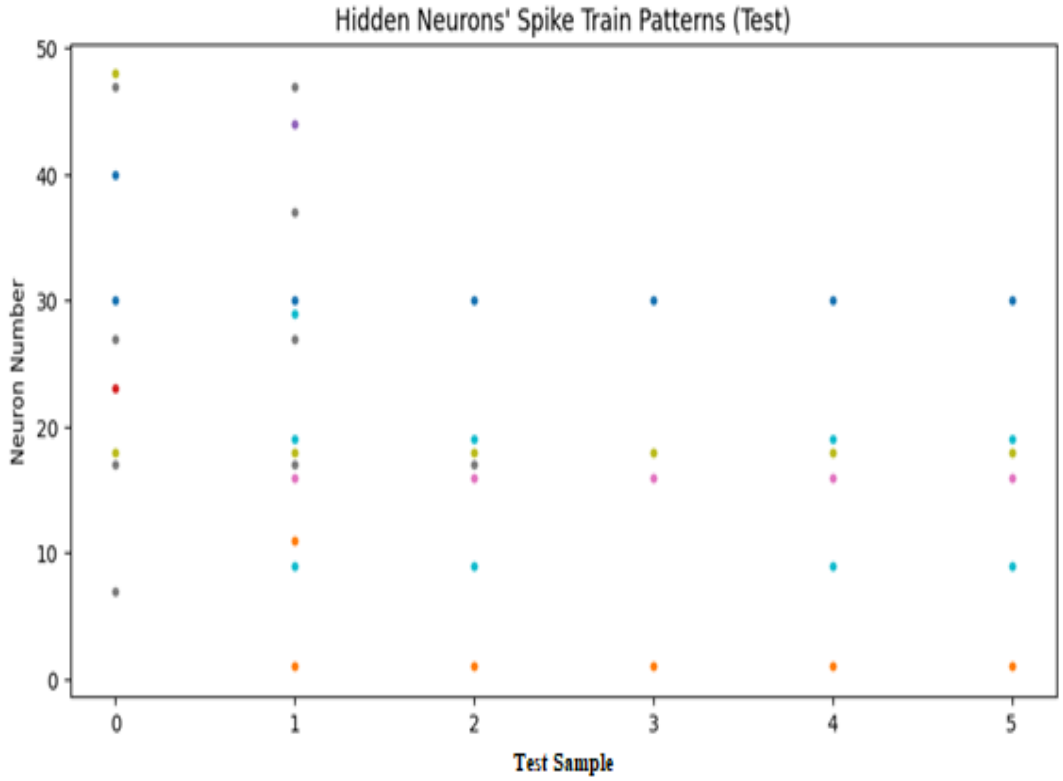


Figure 45

Hidden Neurons Mean of Spike Patterns for Each Neurons and Input Sample Data of Test Dataset



Figures 44 and 45 depict spike patterns. In these figures, the Y axis represents the numbers of neurons in the hidden layer, while the X axis represents the sample numbers in the dataset. The spike patterns of the neurons are represented by dots or scatter points in the plots. These dots or scatter points indicate the average spike rates of the neurons for each corresponding sample.

In the following section, we will analyse the performance of the proposed method in both single and multiple sound source localization and classification tasks.

6.11 Evaluation of Adaptive Physic-Informed ART-rSNN Performance

To evaluate the newly designed structure, we first investigate firing patterns of several examples to recognize the features and detect threshold to distinguish the criteria of the event classification. Figures 46-48 illustrate some of the patterns.

Figure 46

Power Pattern of the Network - A Ringing Bell Sound Event

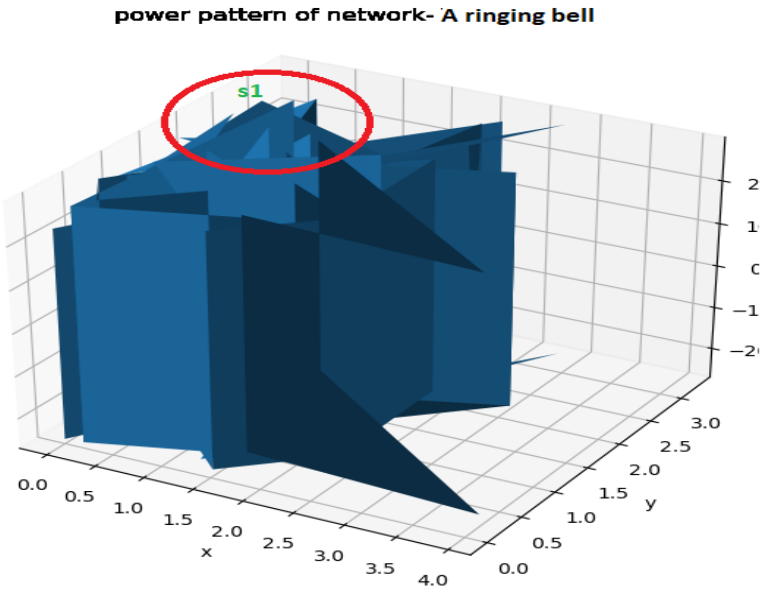


Figure 47

Power Pattern Example of Three Sound Sources in a Physic-Informed ART-rSNN

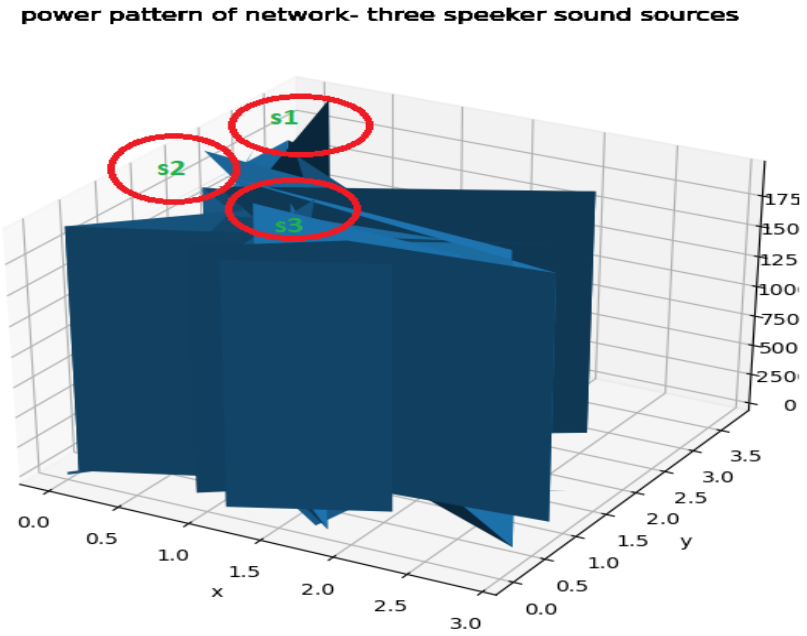


Figure 48

Power Pattern of Two Mixed Sound Events – Dog Barking and Ambulance Car

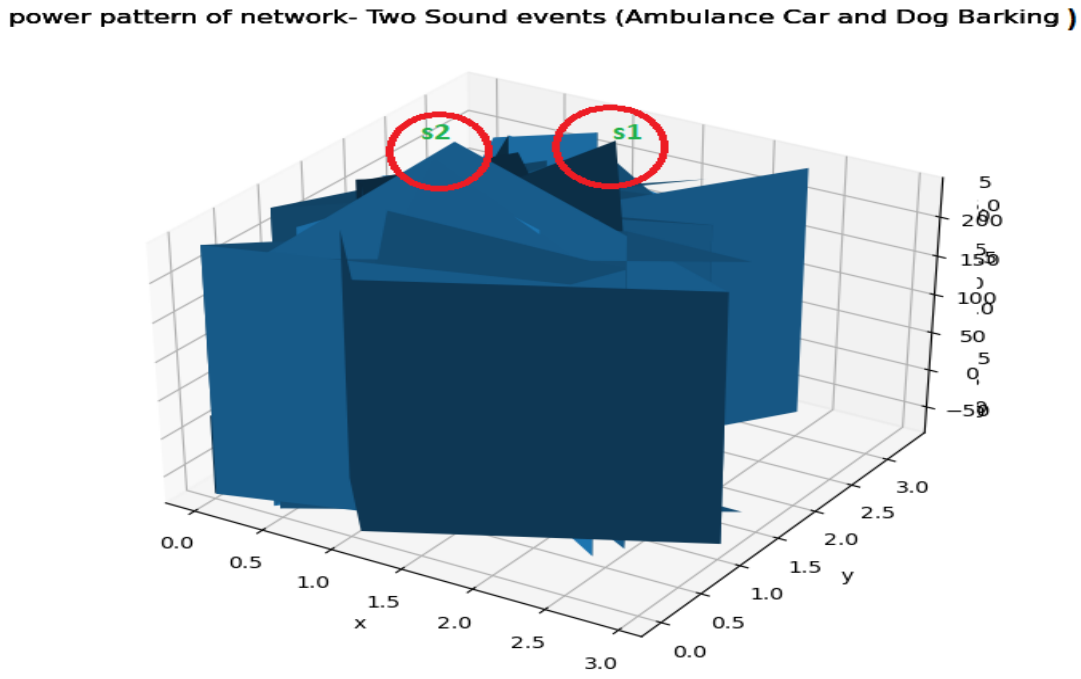
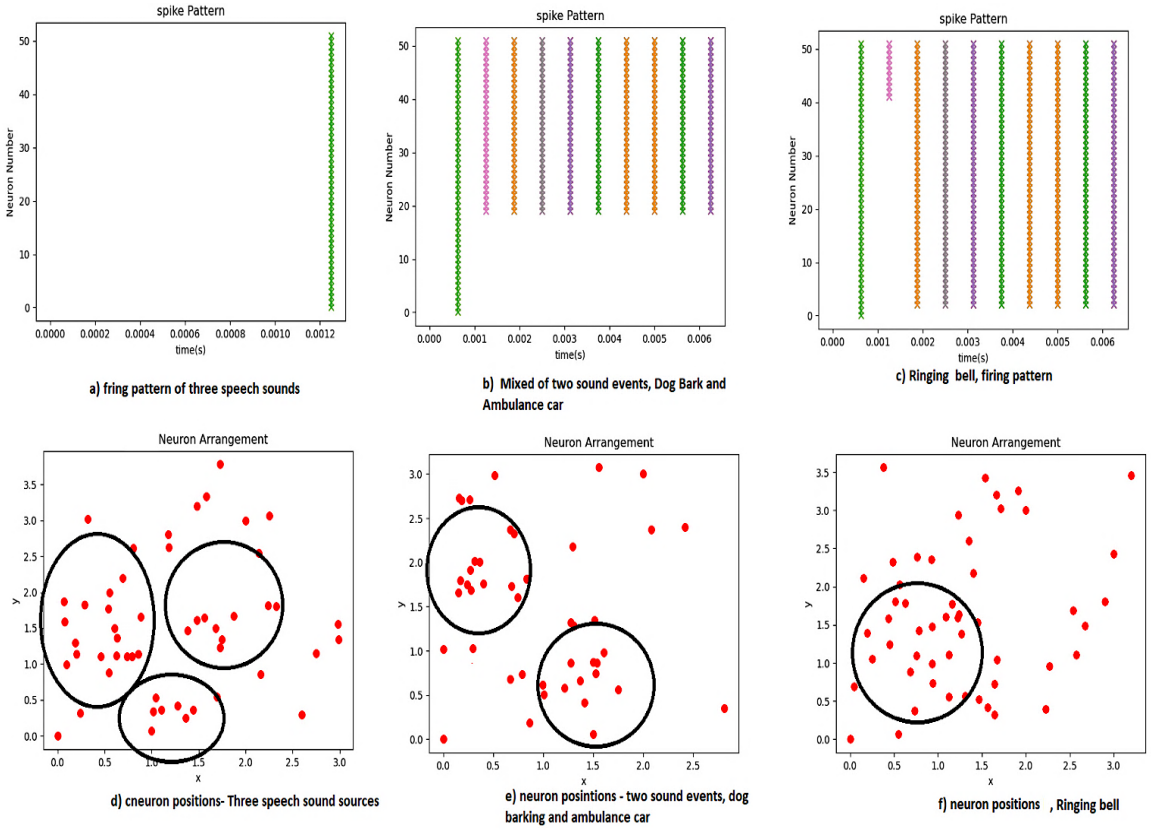


Figure 46-48 indicate that the new strategy is able to detect the number of sound sources by illustration of power peaks in the hidden network structure. Figure 49 depicts firing patterns of the proposed network as well as neuron position arrangements at epoch time 20.

In Figure 48, (a) represents the firing pattern of three speakers, (b) depicts the firing pattern of a dog barking and an ambulance siren, and (c) illustrates the firing pattern of a ringing bell sound along with the network arrangement at epoch time=20 for (d) three speakers, (e) dog barking and ambulance, and (f) ringing bells. Examining Figure 48 reveals that this configuration effectively determines the number of available sound sources. Additionally, it is evident that the neuron arrangement density allocates all sources within different zones, even in the presence of multiple sound sources.

Figure 49

Sound Events Firing Patterns and Neuron Arrangements in Three Different Sound Events and Environments



To assess the performance of this structure in comparison to well-known machine learning (ML) algorithms, the classification performance of the proposed strategy is evaluated and compared with multilayer perceptron (MLP) (Wu, 2018), convolutional neural network (CNN) (Wu, 2018), and recurrent neural network (RNN) (Wu, 2018) classification algorithms in the environmental examples of the FOA dataset. Parameters of the simulated network are given in the Table 6 and Table 7 compares MLP, CNN, and Physic informed classification accuracy.

Table 6

Network Parameters of Physic-informed ART-rSNN

PARAMETER	VALUE
τ_m	1
τ_s	0.9
λ_h	0.3
β	0.5
C_0	342
HIDDEN NEURONS NUMBERS	50
f_d	30

Table 7

Comparison of Classification Accuracy of the Proposed Structure Against Other Neural Network Systems

Model	Multiple SED	Single sound classification accuracy	Multiple sound classification
MLP	No	20%	20%
CRNN	Yes	66.67%	55%
RNN	No	33.3%	30%
Single tempotron SNN	No	50%	15%
Tempotron SNN with one hidden layer	No	66.67	33%
Physic-informed ART-rSNN	Yes	73.44%	33%

Table 7 shows that this structure, as a kind of modified recurrent neural network, exhibited higher accuracy of the pattern recognition rather than RNN and MLP. However, multiple patterns may trigger further possible errors. This structure provides the information of sound sources quantities. It can determine the class of a sound event as well as localize the sources. According to the comparison results, it can be concluded that the modified structure can enhance efficiency of a recurrent neural network.

6.12 Further Classification Evaluation

To ensure a fair comparison for validating our proposed classifier, we performed a detailed analysis comparing the performance of our Physic-Informed ART-rSNN with SOTA models, including MLP, RNN, and CRNN. The efficacy of our newly designed SNN classifier was evaluated by detecting and classifying power patterns and firing rates of various sound events using the TAU-NIGENS Spatial Sound Events 2020 database (Politis & S., 2020). The audio scenes were constructed from individual sounds sourced from the NIGENS public audio event database. All simulations were conducted using Python 3.10, with the performance of our Adaptive PI RSNN method compared to MLP (Pahuja & Kumar, 2021), CRNN (Sang, Park, & Lee, 2018), and RNN (Phan et al., 2017). Table 8 provides the detailed parameters and specifications of these networks.

A key distinction between our proposed method and traditional methods, such as MLP, CRNN, and RNN, lies in our approach to training. Our method uses the first 350 samples for online training and the remaining samples for testing, enabling real-time processing for various applications. In contrast, MLP, CRNN, and RNN typically follow an offline approach, where 90% of the data is allocated for training, and the remaining 10% is used for testing through cross-validation. This difference in training methodology highlights the adaptability and real-time capabilities of our proposed method when compared to the batch-processing approach of traditional models.

Additionally, this evaluation was conducted on a 500-sample batch process, using 500 observed neurons arranged randomly without any specific mapping to the location of sound sources. In this validation stage, the structure did not take into account the spatial data of sound sources, focusing solely on the classification process.

Table 8

Networks' specifications

MODEL	PARAMETERS AND LAYERS	VALUE
ADAPTIVE PHYSIC INFORMED ART-rSNN	τ_m	0.5
	τ_s	2.5
	λ_h	20
	K_0	100
	λ_0	10
	Hidden neurons numbers	50
	f_d	30
MLP **(Phan et al., 2017)**	Hidden Layer	2
	Hidden neurons numbers	24
	Output layer neurons	3
	Input layer Neurons (Input samples)	500
	Learning rate	0.95
CRNN2(0.25M) **SANG, PARK, & LEE, 2018)**	Input samples	500×1
	Convolutional layer	C (64, 80/4)
	Max Pooling	4×1
	Convolutional layer	C (64,3)
	Max Pooling	4×1
	RNN (128) and Output layer	Fully Connected (activation: softmax)
RNN **(PHAN ET AL., 2017)**	Number of Layers	3
	Size of Hidden state vector	256
	The learning rate for Adam optimizer	0.0001
	Dropout rate	0.1
	Regularization parameter	0.001

In the classification algorithms in a second scenario, including a batch process comparison with 500 samples, ensuring a fairer comparison between online and offline processes. Figures 50 to 52 illustrate the firing patterns generated by the proposed methods in the second scenario (batch process) with a 10-cross-fold validation of three classes of sounds in the network with 500 observed neurons and 500 hidden neurons,

highlighting the network's ability to effectively process and classify complex data. Table 9 compares the performance of MLP, CRNN, RNN, and Physic Informed ART- rSNN classification.

Figure 50

Firing pattern of encoded input and generated output spike trains by the proposed method for class 0

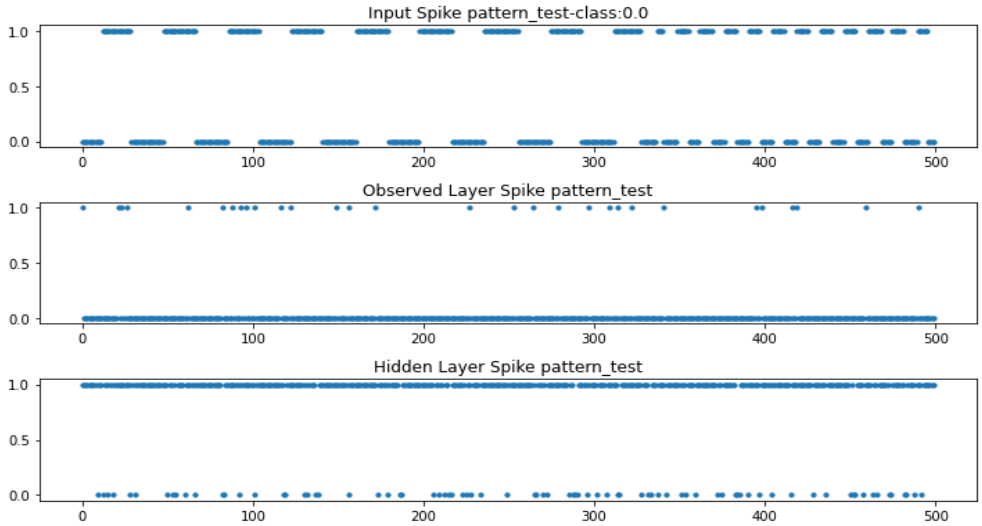


Figure 51

Firing pattern of encoded input and generated output spike trains by the proposed method for class 0.5

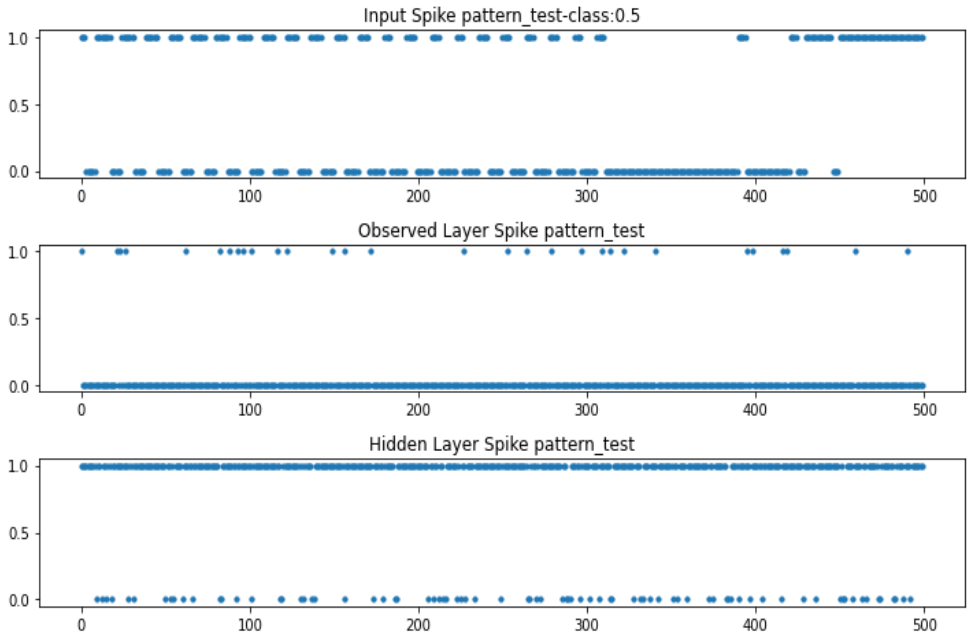


Figure 52

Firing pattern of encoded input and generated output spike trains by the proposed method for class 1

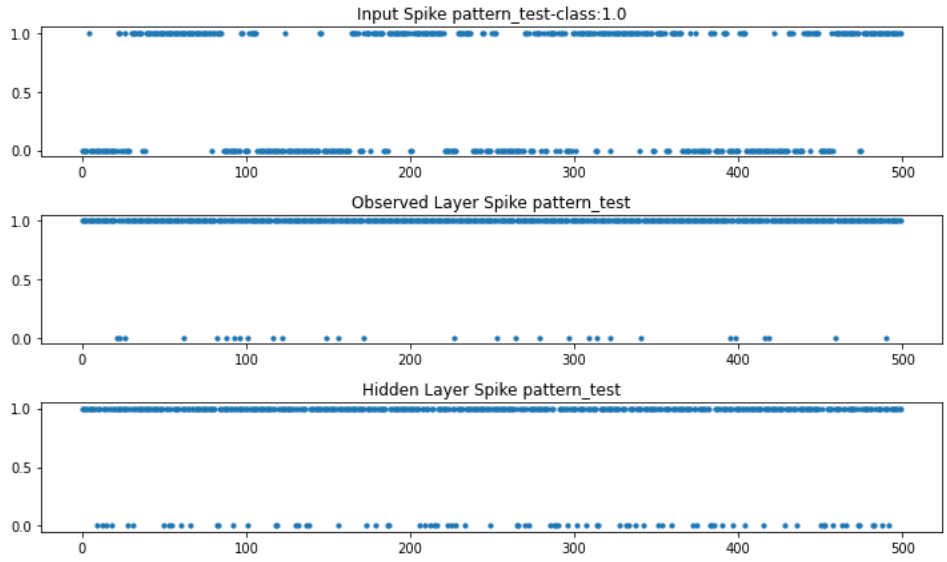


Table 9

Comparison of averaged classification accuracy of the proposed structure against other neural network systems for 10 random partitions of a dataset containing 500 samples, with 90% of the data used for training and 10% for testing. Statistically significant differences between Physic-Informed ART- rSNN metrics and other networks using ranked T tests for the 10 runs are indicated by asterisks (=p≤0.05, **=p≤0.01)*

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
MLP (Pahuja & Kumar, 2021))	0.571	0.321*	0.528	0.439
CRNN (Sang, Park, & Lee, 2018)	0.651**	0.728	0.636**	0.611**
RNN (Phan et al., 2017)	0.335**	0.108**	0.333**	0.163**
Physic-Informed ART- rSNN	0.622	0.738	0.604	0.557

Table 9 indicates the Physic-Informed ART- rSNN, model achieves the highest precision of 73%, underscoring its effectiveness in classification tasks. Each model, namely MLP,

CRNN, RNN, and PI RSNN exhibits distinct classification capabilities. Notably, CRNN demonstrates a precision of 72%, accompanied by balanced precision, recall, and F1 score values, indicating its robustness in accurately identifying and classifying instances. Conversely, RNN lags with a precision of 11% and the lowest precision, recall, and F1 score values, suggesting limitations in its classification effectiveness. MLP metrics, while not significantly different from Physic-Informed ART- rSNN, display a trend to perform at a lower level. The comparison underscores the importance of selecting appropriate neural network architectures tailored to specific classification tasks, as different models exhibit varying performance levels. Further analysis and experimentation may be necessary to understand the underlying factors contributing to the observed performance differences among the models.

In addition to the Tempotron classifier, we also explored several other classifiers to evaluate their performance in our sound event classification and localization tasks. Specifically, we experimented with the mean firing rate threshold method, which was tested alongside the Tempotron classifier. In some cases, the mean firing rate method even outperformed the Tempotron classifier, though overall, the results were quite similar across all methods when evaluated using a 10-fold cross-validation analysis. This indicates that while different classifiers may exhibit slight variations in performance, they generally produced comparable outcomes for the task at hand.

Furthermore, we also utilized K-means clustering and Support Vector Machine (SVM) classifiers to further assess the effectiveness of different approaches. However, both K-means and SVM classifiers performed less well than the Tempotron classifier, with accuracy scores approximately 10% to 15% lower compared to Tempotron in this evaluation. These results highlight the advantages of the Tempotron classifier, particularly in the context of sound event detection, as it demonstrated superior accuracy and robustness in handling the classification tasks compared to the other methods tested.

6.13 Handling Reverberation, Absorption, and Reflection

In this work, we handled the issues of reverberation, absorption, and reflection mainly by the nature of the dataset and the inherent design of our methods for classification and localization. The TAU-NIGENS dataset used in our experiments involves recordings from real-world rooms of different shapes, sizes, and with a variation in sound absorption properties. Reverberation and reflection common in environmental acoustics are naturally caught in those recordings. We refrained from using all possible artificial filters, which would have isolated or minimized such effects. Rather, we tried to retain the realistic acoustic conditions in their most natural states, where reverberation and reflection were allowed to persist.

Our model does not aim at the explicit removal of reverberation or reflection but instead takes those into account indirectly. We drew on ITD- and ILD-fusion techniques shown in the literature to enhance the robustness to reverberation and reflections for various spatial sound processing tasks. These acoustic cues are able to provide spatial and temporal information about the sound sources; thus, inherently, the model can fuse the environmental characteristics. We further incorporated an energy-based cue into our model, which enabled the latter to identify and capture the environmental context of the sound signals without explicitly filtering out or pre-processing.

Our architecture also considered the temporal consistency of environmental effects to help the model adapt to persistent reverberation and absorption characteristics in small steps. This is the temporal consistency that would introduce the classifier to an improved way of handling the long-term behaviour of the environmental factors, which may affect the classification and localization. While indirect methods are the basis for the current approach, further studies will be carried out, which also involve adaptive filtering techniques, with the view to extract and measure the exact influence of reverberation, absorption, and reflection. This will provide further insight into the factors that play a role in sound classification tasks and probably result in future improvements in dealing with real-world environmental acoustic challenges.

6.14 Discussion

In this chapter, a novel idea for multisource sound localization has been presented by a new dynamic structure of a SNN. The method is inspired by the way humans approximate the location of sound sources by using binaural features such as interaural time difference (ITD) and interaural level difference (ILD). As well, an energy-based sound localization cue is integrated within a new structure of rSNNs to jointly detect and trigger sound events. Since a SNN presents a more realistic representation of human hearing by mimicking the binaural time delays in its simulations, it has been selected as the main structure of the proposed system. The significant feature of this newly designed architecture is its ability to modify and change the network size and arrangements. This special feature fosters the efficiency of considering spatial information as well as temporal data. The proposed structure works based on the time domain classification approaches, which can be developed by modification of the time encoding process.

A sound event-trigger and classification process includes three overarching steps. The first step is to detect the sound events and determine the number of sources in the incoming signal by analysing the SNN power pattern as well as considering the arrangements of the neurons. Second, the centroid of the high density neuron arrangements' neighbourhood should be intended as the approximation of sound source location. Third, firing patterns are compared to several predefined thresholds with the aim of classification.

Two types of datasets are utilized to evaluate the proposed method versus other well-known localization and classification strategies. Different sound events have been tested to be detected and localized by only two microphones. First, only a single clapping hand sound source is considered to estimate the sound source. Then, in the classification part, two and three sound events are determined and localized by the new structure.

The process of designing a new dynamic structure of a SNN consists of two main evolutionary stages.

First, ART-rSNN, a specialized SNN architecture, was meticulously crafted with a primary focus on the challenging task of sound localization. This initial design showed promising

results, particularly in the context of localizing a single sound source. Evaluation involved a meticulous comparison of mean squared errors (MSEs) with deep and non-deep learning approaches. The results revealed a remarkable achievement: ART-rSNN exhibited significantly higher accuracy in this specialized task, outperforming other methodologies.

Second, recognizing the inherent limitations of a convolutional recurrent neural network (CRNN) in the context of audio classification, a novel approach was developed. Leveraging insights from the physical concept of power within network strategies, the innovative physics-informed ART-rSNN was introduced. This novel structure was engineered not only to detect and classify sound sources, but also to localize them.

Simulation results yielded fascinating insights. The physics-informed ART-rSNN demonstrated its versatility, excelling in scenarios involving multiple sound sources. More impressively, it displayed an extraordinary ability to precisely determine the locations of sound events, achieving an impressive accuracy rate of 73.44%.

It's worth noting that the performance of CRNN models often exhibits limitations that are highly dependent on the available data samples. In this context, mel-frequency cepstral coefficients (MFCCs) emerged as a powerful tool to enhance audio classification by capturing both temporal and spectral features. Interestingly, the time-frequency data alone did not yield significant improvements in classification results. Time-domain data provided an accuracy of 33%, which aligns with the baseline. However, in the realm of SNNs, temporal raw data can be harnessed effectively, leading to identical classification results as when using MFCCs. The proposed physics-informed ART-rSNN' structure further augments this capability, offering a compelling solution to address these limitations.

Throughout the thesis, both the recorded dataset and the external datasets utilized a two-microphone configuration. This choice was driven by considerations of efficiency, minimizing hardware requirements while maintaining effective performance. However, the algorithms developed are designed to scale and work with configurations involving two or more microphones. Increasing the number of sensors would undoubtedly enhance accuracy, especially for energy-based methods that benefit from richer spatial data. While the potential for integrating moving sensors—similar to the constant

adjustments seen in animal auditory systems—was not addressed in this study, it represents an exciting avenue for future exploration. The next chapter will conclude the study and offer insights into avenues for future research

7. Conclusion and Future Works

7.1 Introduction

The main objective of this thesis was to design a new rSNN structure, able to integrate spatial data as well as temporal information within incoming signals. A spiking neural network (SNN) was utilized as a feature extractor, in which output firing rates of the network were processed and adjusted so that it was able to trigger events by eliminating noise effects on the network. Various machine learning methods – including GCC-PHAT, MUSIC, and SNN algorithms – were trained to predict the locations of a single sound source. This novel dynamic-structure network was tested with different datasets and compared with different machine learning algorithms – namely, CRNN, RNN, tempotron SNN, and MLP – and the network exhibited excellent performance for single and multisource localization and classification.

This work has developed a new feature extraction model to solve the multisource localization and classification problem, a model which is robust and has applicable real-time applications. Different machine learning approaches have been compared and their effectiveness in sound source localization in the presence of environmental noise examined. The utilized database (TU-NIGEN) contains audio speech as well as indoor and outdoor events samples recorded in different environments. Real data receded by mobile microphones are also investigated in this work. The following section provides a summary of the thesis.

7.2 Thesis Summary

Chapter 1 of this study begins by highlighting the limitations and shortcomings of existing methods in the context of sound source localization (SSL) and sound classification (SC). The first chapter introduces the use of spiking neural networks (SNNs) as a potential solution to address these issues. It outlines the study's goals and poses research questions that guide the investigation. One of the key features of SNNs discussed in this chapter is their inherent ability to perform time coding and handle time delays. These characteristics make them suitable for applications involving interaural time difference (ITD)-based methods, which are important for sound localization tasks.

The chapter also emphasizes the significance of neural network structure and connectivity in determining the performance of SNNs in SSL and SC.

Chapter 2 provides a comprehensive review of the literature related to SSL and SC methods. It categorizes different approaches in these domains and discusses the challenges associated with each. Notably, the chapter covers both deep learning and non-deep learning methods in SSL and SC. It highlights important features and trends in sound classification.

Chapter 3 delves into the theoretical foundations relevant to the research study. It explores the principles of adaptive resonance theory (ART), spiking neural networks (SNNs), learning rules, and encoding laws. These concepts are crucial for understanding the underlying theory and mechanisms of the proposed methods.

Chapter 4 focuses on recurrent neural network (RNN) structures and their various components, with a specific emphasis on their applications in sound processing. This chapter provides insights into how RNNs can be utilized in sound-related tasks.

Chapter 5 is divided into two main sections. The first section discusses sound event tracking and localization. It introduces a novel dynamic-structure rSNN that integrates the energy-based method of sound event localization (SEL) with ITD-based methods. The second section modifies this ART-rSNN by incorporating the concept of signal power and electrical powers to enable event-triggered sound event localization and classification. This chapter highlights the evolution of the proposed methods.

Chapter 6 presents the results obtained from the simulation of both strategies discussed in Chapter 5. It compares these results against well-established machine learning (ML) methods. The chapter showcases the high performance of the proposed methods, particularly in the detection and localization of multiple sound sources. While the clear classification of multiple sound sources may not surpass convolutional recurrent neural networks (CRNNs), the proposed methods offer more efficient results compared to MLP, RNN, and other fixed-structure SNN methods. The chapter emphasizes the significant improvement in the accuracy of classifications achieved by this method compared to traditional RNN-based approaches.

Overall, these chapters provide a comprehensive overview of the research study, from its motivations and theoretical foundations to the development and evaluation of novel SNN-based approaches for sound localization and classification.

7.3 Response to the Research Questions

1. Sound Localization

- How can the accuracy of sound localization be improved by synergizing energy-based cues with traditional interaural time difference (ITD) and interaural intensity difference (IID) methods in spiking neural network (SNN) structures?

  This research study adopted an energy- based decay model as the main part of the cost function of the learning law. The objective function estimates the measured power of the incoming signals within a group of observed neurons, and a gradient descent approach is utilized to reduce squared estimation errors. Then, an ITD cue is utilized to modify the spatial arrangement and connectivity of the SNNs.

- What factors influence the determination of the azimuth sign when locating a sound?

  The estimated azimuth angles are validated based on a given function to detect the sign of the azimuth angle.

- How should neurons be arranged and connected in an rSNN structure to effectively process spatio-temporal data in incoming signals?

  By integration of ITD and IID cues, the hidden neuron set in new arrangements, and connectivity of the neurons are amplified or attenuated, based on the Euclidean distance from each other.

2. Sound Event Triggering

- What methods can be used to encode audio data into electrical impulses suitable for rSNNs in noisy environments?

  In fact, two approaches of encoding strategies – membrane potential and temporal encoding methods – are fused to increase accuracy of encoding. This integration can be seen in observed neuron and hidden neuron connectivity.

- What threshold should be considered to initiate neuronal activity in rSNNs?

  The threshold voltage is updated based on the firing rates of the network in several epochs, so that it can be stable during processing.

3. Sound Event Classification

   - What is the most effective approach for classifying sound event patterns in the developed network structure?
     Pattern of power, spatial arrangement, and firing rate can be utilized individually, or integrated to each other, to find the criteria of classification. We utilized firing rate criteria methodology in simulations to classify numerically.

4. System Evaluation

- How does the dynamic spatio-temporal rSNN architecture perform in sound localization compared to conventional approaches?
  Results indicate that the RNN classification method manifests lower accuracy, and the proposed method has higher accuracy in localization as compared to conventional methods. In fact, due to ability to modify the arrangement of neurons, the connectivity is modified so that it increases the number of effective neurons in synaptic activity within a neighborhood and eliminates the neurons that are far away from the activated zone by the sound sources.

- How do the adaptive size increase of the neural network and dynamically assigned neuron positions impact the learning process and overall performance of the model?

  These approaches are often robust in noisy and reverberant environments, especially under uncertain conditions. The main advantages of these methods are in their capabilities for accurate detection and classifications. The major challenges of the methods are their high computational costs and their need to be trained. Most of these methods are sensitive to parameter tuning and their structures.

- How does the dynamic spatio-temporal rSNN performs on sound localization compared to conventional approaches?

  Results indicate that by utilizing a lower hidden number quantity, here 50 neurons (maximum), rather than other AI methods, this structure reduces the time computation costs of processing. However, some other processing stages for updating arrangements influence an increase in computational costs. Meanwhile, results indicate that for complex applications, the proposed method surpasses other neural network- based strategies because a lower population of networks can provide accurate results.

7.4 Conclusion and Limitation

As mentioned earlier, the main objective of this research study was to design a new dynamic architecture that can modify the arrangement of neurons, based on the spatial information which is concealed within the incoming signals. The strategy includes a combination of data and feature fusion approaches, which can handle complex sound event localization and detection problems. By adopting tempotron-STDP, as well as ART, the proposed method is able to be utilized for accurate classification and localization of sound sources. The main limitation of the proposed method is in tuning the classification parameters and the learning rate of the network. As well, training is still an open issue in neural networks.

7.5 Future Works

The main contribution of this research study is combining cues and encoding methodologies, as well as considering spatial information, to present a self-modifying architecture to reduce the burden of utilizing a high network population in complex problems. To foster efficiency of the proposed network architecture and address the issues of the proposed method, this study suggests several future approaches as follows:

- Consider frequency domain and time domain methods to increase classification accuracy. In fact, time-frequency approaches can potentially increase the ability to distinguish different categories in complex incoming data.

- Utilize a physic-informed structure. Dealing with duplex theory and concerning acoustic wave physic laws can reduce computation costs, instead of utilizing nonlinear neurons.
- Utilize probabilistic neural networks or stochastic approaches to eliminate or at least reduce training process time.
- Utilize other cues, or monaural approaches within the network to reduce dependency on microphone arrangements. However, the proposed method is also able to efficiently work in different microphone arrangements.

References

A. J. Watt and N. S. Desai. (2010). Homeostatic plasticity and STDP: keeping a neuron's cool in a fluctuating world. Frontiers in synaptic neuroscience, 5.

Abayomi-Alli, O. O., Damaševičius, R., Qazi, A., Adedoyin-Olowe, M., & Misra, S. (2022). Data augmentation and deep learning methods in sound classification: A systematic review. Electronics, 11(22), 3795.

Adavanne, S. (2018). Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. 2018 26th European Signal Processing Conference (EUSIPCO), 1462-1466.

Adavanne, S. (2018). Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. IEEE Journal of Selected Topics in Signal Processing, 12(3), 34-48.

Adavanne, S. (2019). Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network. arXiv preprint arXiv:1904.12769.

Adel, H., Souad, M., Alaqeeli, A., & Hamid, A. (2012). Beamforming techniques for multichannel audio signal separation. arXiv preprint arXiv:1212.6080.

Aggarwal, M., Rai, A., & Yadav, M. P. (2020). Urban sound classification using neural networks. International Journal for Research in Applied Science and Engineering Technology, 8, 1670-1673.

Al-Abboodi, H. M. A. (2019). Binaural sound source localization using machine learning with spiking neural networks features extraction. University of Salford (United Kingdom).

Algobail, A., Soudani, A., & Alahmadi, S. (2019). Energy-efficient scheme for target recognition and localization in wireless acoustic sensor networks. International Journal of Distributed Sensor Networks, 15(11), 1550147719891406.

Amidi, A., & Amidi, S. (2018). VIP cheatsheet: Recurrent neural networks.

Arbel, N. (2018). How LSTM networks solve the problem of vanishing gradients. Medium, December.

Arnault, A., Hanssens, B., & Riche, N. (2020). Urban sound classification: Striving towards a fair comparison. arXiv preprint arXiv:2010.11805.

Atito, S., Awais, M., Wang, W., Plumbley, M. D., & Kittler, J. (2024). ASiT: Local-Global Audio Spectrogram Vision Transformer for Event Classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Averbeck, B. B. (2006). Neural correlations, population coding and computation. Nature reviews neuroscience, 358-366.

Bansal, A., & Garg, N. K. (2022). Environmental Sound Classification: A descriptive review of the literature. Intelligent Systems with Applications, 200115.

Barroso, V. R., Xavier, F. C., & Ferreira, C. E. L. (2023). Applications of machine learning to identify and characterize the sounds produced by fish. ICES Journal of Marine Science, 80(7), 1854-1867.

Belloch, J. A. (2015). On the performance of multi-GPU-based expert systems for acoustic localization involving massive microphone arrays. Expert Systems with Applications, 5607-5620.

Bianco, M. J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M. A., Gannot, S., & Deledalle, C. A. (2019). Machine learning in acoustics: Theory and applications. The Journal of the Acoustical Society of America, 146(5), 3590-3628.

Bing, Z., Baumann, I., Jiang, Z., Huang, K., Cai, C., & Knoll, A. (2019). Supervised learning in SNN via reward-modulated spike-timing-dependent plasticity for a target reaching vehicle. Frontiers in neurorobotics, 13, 18.

Bruce, M. M. (1969). *Estimation of variance by a recursive equation* (Vol. 5465). National Aeronautics and Space Administration.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing*, *37*(1), 54-115.

Carpenter, G. A., & Grossberg, S. (2012). Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world. Neural Networks, 31, 1-24. https://doi.org/10.1016/j.neunet.2012.09.017

Carpenter, G. A., & Grossberg, S. (2016). Adaptive Resonance Theory. In Encyclopedia of Machine Learning and Data Mining (pp. 1-6). Springer US. https://doi.org/10.1007/978-1-4899-7687-1_1

Chan, V. Y. S., Jin, C. T., & Schaik, A. V. (2010). Adaptive sound localization with a silicon cochlea pair. Frontiers in neuroscience, 4, 196.

Chen, J., Takashima, R., Guo, X., Zhang, Z., Xu, X., Takiguchi, T., & Hancock, E. R. (2021). Multimodal fusion for indoor sound source localization. Pattern recognition, 115, 107906.

Chen, J., Takashima, R., Guo, X., Zhang, Z., Xu, X., Takiguchi, T., & Hancock, E. R. (2021). Multimodal fusion for indoor sound source localization. *Pattern recognition*, *115*, 107906.

Chiariotti, P., Martarelli, M., & Castellini, P. (2019). Acoustic beamforming for noise source localization – Reviews, methodology, and applications. Mechanical Systems and Signal Processing, 120, 422-448.

Cho, K. (2014). On the Properties of Neural Machine Translation: Encoder-decoder Approaches. *arXiv preprint arXiv:1409.1259*.

Chun C, Jeon KM, Choi W. Configuration-Invariant Sound Localization Technique Using Azimuth-Frequency Representation and Convolutional Neural Networks. Sensors (Basel). 2020 Jul 5;20(13):3768. doi: 10.3390/s20133768.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Chung, M. A., Chou, H. C., & Lin, C. W. (2022). Sound localization based on acoustic source using a multiple microphone array in an indoor environment. Electronics, 11(6), 890.

Cobos, M. (2017). A survey of sound source localization methods in wireless acoustic sensor networks. Wireless Communications and Mobile Computing.

Cook, P. U. (2015). Unsupervised learning of digit recognition using spike-timing-dependent plasticity. Frontiers in Computational Neuroscience, 99.

Cooper, S. J. (2005). Hebb's synapse and learning rule: a history and commentary. Neuroscience & Biobehavioral Reviews, 29(6), 851-874.

Cooper. S. J. (2005). Hebb's synapse and learning rule: a history and commentary. Neuroscience & Biobehavioral Reviews, 29(6), 851-874.

Cucchi, M., Abreu, S., Ciccone, G., Brunner, D., & Kleemann, H. (2022). Hands-on reservoir computing: a tutorial for practical implementation. Neuromorphic Computing and Engineering.

Cui, Y. (2016). The time delay neural networks (TDNNs) are another way for sequence information processing, which organize sequential memory information in a multilayer feedforward structure. The current machine learning methods obtain impressive perf. Neural computation, 2474-2504.

Dang, X. (2019). Indoor multiple sound source localization via multi-dimensional assignment data association. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 1944-1956.

Das, J. K., Ghosh, A., Pal, A. K., Dutta, S., & Chakrabarty, A. (2020). Urban sound classification using convolutional neural network and long short term memory based on multiple features. 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS).

De Silva, L. C., Morikawa, C., & Petra, I. M. (2012). State of the art of smart homes. Engineering Applications of Artificial Intelligence, 25(7), 1313-1321.

Deng, M., Meng, T., Cao, J., Wang, S., Zhang, J., & Fan, H. (2020). Heart sound classification based on improved MFCC features and convolutional recurrent neural networks. Neural Networks, 130, 22-32.

Denk, F. (2020). Characterizing and conserving the transmission properties of the external ear with hearing devices.

Desai, D., & Mehendale, N. (2022). A review on sound source localization systems. Archives of Computational Methods in Engineering, 29(7), 4631-4642.

Desai, U., & Mehendale, N. (2021). Sound source localization: A review. Journal of the Acoustical Society of America, 149(3), 1806-1824. doi: 10.1121/10.0003657

Diehl, P. U. (2015). Unsupervised learning of digit recognition using spike-timing-dependent plasticity. Frontiers in computational neuroscience, 99.

DiPietro, R., & Hager, G. D. (2020). Deep learning: RNNs and LSTM. In Handbook of medical image computing and computer-assisted intervention (pp. 503-519). Academic Press.

Dynamic Sound Localization during Rapid Eye-Head Gaze Shifts. (n.d.). National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6730098/

Egger, V. (1999). Coincidence detection and changes of synaptic efficacy in spiny stellate neurons in rat barrel cortex. Nature neuroscience, 1098-1105.

Ekpezu, A. O., Katsriku, F., Yaokumah, W., & Wiafe, I. (2022). The Use of Machine Learning Algorithms in the Classification of Sound: A Systematic Review. International Journal of Service Science, Management, Engineering, and Technology (IJSSMET), 13(1), 1-28.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179-211.

Elsaraiti, M., & Merabet, A. (2021). Application of long-short-term-memory recurrent neural networks to forecast wind speed. Applied Sciences, 11(5), 2387.

Escobar, F. A., Chang, X., Ibala, C., & Valderrama, C. (2013). Accuracy study of a real-time hybrid sound source localization algorithm. In Intelligent Technologies for Interactive Entertainment: 5th International ICST Conference, INTETAIN 2013, Mons, Belgium, July 3-5, 2013, Revised Selected Papers 5 (pp. 146-155). Springer International Publishing.

Escudero, E. (2018). Real-time neuro-inspired sound source localization and tracking. Neurocomputing, 129-139.

Fei-Fei, L. (2006). One-shot learning of object categories. IEEE transactions on pattern analysis and machine intelligence, 594-611.

Ferreira, A. A., Ludermir, T. B., & De Aquino, R. R. (2013). An approach to reservoir computing design and training. Expert systems with applications, 40(10), 4172-4182.

Florian, R. V. (2005). A reinforcement learning algorithm for spiking neural networks. Seventh International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, (p. 8). SYNASC 2005.: IEEE.

Francart T, Lenssen A, Wouters J. Enhancement of interaural level differences improves sound localization in bimodal hearing. J Acoust Soc Am. 2011 Nov;130(5):2817-26. doi: 10.1121/1.3641414.

Francl, A., & McDermott, J. H. (2022). Deep neural network models of sound localization reveal how perception is adapted to real-world environments. Nature human behaviour, 6(1), 111-133.

Gao, S. (2018). A Modified Frequency Weighted MUSIC Algorithm for Multiple Sound Sources Localization. In 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP).

Gaver, W. W. (2013). Sounding bodies. MIT Press.

Gehrig, T., Nickel, K., Ekenel, H. K., Klee, U., & McDonough, J. (2005, October). Kalman filters for audio-video source localization. In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005 (pp. 118-121). IEEE.

Gerstner, W. (2000). Population of Dynamics of Spiking Neurons: Fast Transient, Asynchronous State And Locking. Neural Comput, 43-89.

Gerstner, W. (2002). Spiking neuron models: Single neurons, populations, plasticity. Cambridge University press.

Glackin, Brendan and Wall, Julie A and McGinnity, Thomas M and Maguire, Liam P and McDaid, Liam J. (2010). A spiking neural network model of the medial superior olive using spike timing dependent plasticity for sound localization. Frontiers in computational neuroscience, 18.

Goodman, D. (2009). Sound localization with spiking neural networks. BMC Neuroscience, 10(1), 1--1.

Goodman, D. F. (2010). Spike-timing-based computation in sound localization. PLoS computational biology, e1000993.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, *28*(10), 2222-2232.

Grondin, F., & Michaud, F. (2019). Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations. *Robotics and Autonomous Systems*, *113*, 63-80.

Gronh, A., & Buchholz, J. M. (n.d.). Static and dynamic sound source localization in a virtual room. ResearchGate.

Grossberg, S. (2013). Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world. Neural Networks, 37, 1-47. https://doi.org/10.1016/j.neunet.2012.09.017

Grossberg, S. (2013). Recurrent neural networks. Scholarpedia, 8(2), 1888.

Grumiaux, P. A., Kitić, S., Girin, L., & Guérin, A. (2022). A survey of sound source localization with deep learning methods. The Journal of the Acoustical Society of America, 152(1), 107-151.

Guo, L., Gao, Z., Qu, J., Zheng, S., Jiang, R., Lu, Y., & Qiao, H. (2023). Transformer-based Spiking Neural Networks for Multimodal Audio-Visual Classification. IEEE Transactions on Cognitive and Developmental Systems.

Gütig, R., & Sompolinsky, H. (2006). The tempotron: a neuron that learns spike timing–based decisions. Nature neuroscience, 9(3), 420-428.

Hamedani, K. (2020). Energy-efficient deep spiking recurrent neural networks: A reservoir computing-based approach (Doctoral dissertation, Virginia Tech).

Hammer, H. (2021). Dynamically localizing multiple speakers based on the time-frequency domain. EURASIP Journal on Audio, Speech, and Music Processing, 1-10.

Han, B. (2020). Rmp-snn: Residual membrane potential neuron for enabling deeper high-accuracy and low-latency spiking neural network. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, (pp. 13558-13567).

Hanson, J.-N., Gauvrit, F., Renard, C., Lemesre, P.-E., Bonne, N.-X., & Vincent, C. (2018). Sound source localization. European annals of otorhinolaryngology, head and neck diseases, 135(4), 259-264.

He, W., Motlicek, P., & Odobez, J. M. (2019, May). Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 770-774). IEEE.

Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen Netzen. *Diploma, Technische Universität München*, *91*(1), 31.

Hochreiter, S. (1997). Long Short-term Memory. *Neural Computation MIT-Press*.

https://www.researchgate.net/publication/236169360_Static_and_dynamic_sound_source_localization_in_a_virtual_room

Huang, Z., Xu, Y., Shi, J., Zhou, X., Bao, H., & Zhang, G. (2019). Prior guided dropout for robust visual localization in dynamic environments. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2791-2800).

Ishi, C. (2009). Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments. 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2027--2032.

J. Wu, Y. (2018). A spiking neural network framework for robust sound classification. Frontiers in Neuroscience, 836.

Jaeger, H. (2007). Special issue on echo state networks and liquid state machines. Neural Networks, 287-289.

Jin, W., Wang, X., & Zhan, Y. (2022). Environmental Sound Classification Algorithm Based on Region Joint Signal Analysis Feature and Boosting Ensemble Learning. Electronics, 11(22), 3743.

Karanasiou, I. S. and J. C. Brown (2021). "Deep learning for acoustic source classification: A review." Applied Acoustics 183: 108256.

Kasabov, N. (2014). NeuCube: A spiking neural network architecture for mapping, learning, and understanding of spatio-temporal brain data. Neural Networks, 62-76.

Khan, M. S., Shah, M., Khan, A., Aldweesh, A., Ali, M., Tag Eldin, E., ... & Hussain, L. (2022). Improved Multi-Model Classification Technique for Sound Event Detection in Urban Environments. *Applied Sciences*, *12*(19), 9907.

Kheradpisheh, S. R. (2018). STDP-based spiking deep convolutional neural networks for object recognition. Neural Networks, 56-67.

Kholkin, V., Druzhina, O., Vatnik, V., Kulagin, M., Karimov, T., & Butusov, D. (2023). Comparing reservoir artificial and spiking neural networks in machine fault detection tasks. Big Data and Cognitive Computing, 7(2), 110.

Kita, S., & Kajikawa, Y. (2021). Fundamental study on sound source localization inside a structure using a deep neural network and computer-aided engineering. Journal of sound and vibration, 513, 116400.

Kwak, K. (2008). Sound localization based on excitation source information for intelligent home service robots. In International Conference on Image and Signal Processing. 115. Berlin, Heidelberg: Springer.

Kwon, A. M., & Kang, K. (2022). A temporal dependency feature in lower dimension for lung sound signal classification. Scientific reports, 12(1), 7889.

Lee, J. H., Delbruck, T., & Pfeiffer, M. (2016). Training deep spiking neural networks using backpropagation. *Frontiers in neuroscience*, *10*, 508.

Lee, R. (2020). Sound source localization based on GCC-PHAT with diffuseness mask in noisy and reverberant environments. IEEE Access, 7373--7382.

Lee, S. Y., Chang, J., & Lee, S. (2021). Deep learning-based method for multiple sound source localization with high resolution and accuracy. Mechanical Systems and Signal Processing, 161, 107959.

Legenstein, R. (2005). What Can a Neuron Learn with Spike - Time - Dependent – Plasticity? Neural Comp, 2337-2382.

Li, D., & Hu, Y. H. (2003). Energy-based collaborative source localization using acoustic microsensor array. EURASIP Journal on Advances in Signal Processing, 2003, 1-17.

Liaquat, M. U. (2021). Localization of Sound Sources: A Systematic Review. Energies, 14(13), 3910.

Liaquat, M. U., Munawar, H. S., Rahman, A., Qadir, Z., Kouzani, A. Z., & Mahmud, M. P. (2021). Localization of sound sources: A systematic review. Energies, 14(13), 3910.

Lillicrap, T. P., & Santoro, A. (2019). Backpropagation through time and the brain. Current Opinion in Neurobiology, 55, 82-89.

Lim, S. (2021). Hebbian learning revisited and its inference underlying cognitive function. Current Opinion in Behavioral Science, 96-102.

Liu, J., Perez-Gonzalez, D., Rees, A., Erwin, H., & Wermter, S. (2009, June). A biomimetic spiking neural network of the auditory midbrain for mobile robot sound localisation in reverberant environments. In *2009 International Joint Conference on Neural Networks* (pp. 1855-1862). IEEE.

Lopatka, K., Kotus, J., & Czyzewski, A. (2016). Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations. Multimedia Tools and Applications, 75, 10407-10439.

Lukoševičius, M., & Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. Computer Science Review, 3(3), 127-149.

Ma, L., Hu, C., & Cheng, F. (2021). State of charge and state of energy estimation for lithium-ion batteries based on a long short-term memory neural network. Journal of Energy Storage, 37, 102440.

Ma, W., Bao, H., Zhang, C., & Liu, X. (2020). Beamforming of phased microphone array for rotating sound source localization. Journal of sound and vibration, 467, 115064.

Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural networks*, *10*(9), 1659-1671.

Masquelier, T. (2009). Competitive STDP-based spike pattern learning. Neural computation, 1259-1276.

Masuyama, Y. (2020). Self-supervised neural audio-visual sound source localization via probabilistic spatial modeling. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 1-8). IEEE.

McAnally, K. I., & Martin, R. L. (2014). Sound localization with head movement: implications for 3-d audio displays. *Frontiers in neuroscience*, *8*, 210.

McLoughlin, I., Zhang, H., Xie, Z., Song, Y., Xiao, W., & Phan, H. (2017). Continuous robust sound event classification using time-frequency features and deep learning. Plos one, 12(9), e0182309.

Medsker, L. R., & Jain, L. C. (2001). Recurrent neural networks. Design and Applications, 5(64-67), 2.

Meedeniya, D., Ariyarathne, I., Bandara, M., Jayasundara, R., & Perera, C. (2023). A Survey on Deep Learning-based Forest Environment Sound Classification at the Edge. ACM Computing Surveys.

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2023). A review of deep learning techniques for speech processing. Information Fusion, 101869.

Melandri, L. (2014). Introduction to reservoir computing methods.

Meng, W., & Xiao, W. (2017). Energy-based acoustic source localization methods: a survey. Sensors, 17(2), 376.

Merolla, P. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. Science, 345(6197), 668-673.

Mesaros, A., Heittola, T., Virtanen, T., & Plumbley, M. D. (2021). Sound event detection: A tutorial. IEEE Signal Processing Magazine, 38(5), 67-83.

Middlebrooks, J. C. (2015). Sound localization. Handbook of clinical neurology, 129, 99-116.

Mostafa, H. (2017). Supervised learning based on temporal coding in spiking neural networks. IEEE transactions on neural networks and learning systems, 28(12), 3227-3235.

Munappy, A. R., Bosch, J., Olsson, H. H., Arpteg, A., & Brinne, B. (2022). Data management for production quality deep learning models: Challenges and solutions. Journal of Systems and Software, 191, 111359.

Musicant, A. D., & Butler, R. A. (1984). The influence of pinnae-based spectral cues on sound localization. *The Journal of the Acoustical Society of America*, *75*(4), 1195-1200.

Nakamura, K., Nakadai, K., Asano, F., Hasegawa, Y., & Tsujino, H. (2009). Intelligent sound source localization for dynamic environments. 2009 IEEE/RSJ international conference on Intelligent Robots and Systems.

Nasiri, A., & Hu, J. (2021). SoundCLR: Contrastive learning of representations for improved environmental sound classification. arXiv preprint arXiv:2103.01929.

Nawrot, M. P. (2009). Precisely timed signal transmission in neocortical networks with reliable intermediate-range projection. Frontiers in Neural Circuits, 3, 3.

Ngo, L., Cha, J., & Han, J.-H. (2019). Deep neural network regression for automated retinal layer segmentation in optical coherence tomography images. IEEE transactions on image processing, 29, 303-312.

Nguyen, T. N. T., Jones, D. L., Ranjan, R., Jayabalan, S., & Gan, W. S. (2019). A two-step system for sound event localization and detection. arXiv preprint arXiv:1911.11373.

Nogueira, A. F. R., Oliveira, H. S., Machado, J. J., & Tavares, J. M. R. (2022). Sound Classification and Processing of Urban Environments: A Systematic Literature Review. Sensors, 22(22), 8608.

Nunes, L. O., et al. (2021). "An overview of sound source localization methods." Applied Acoustics 184: 108194.

Nuntalid, N. (2011). EEG classification with BSA spike encoding algorithm and evolving probabilistic spiking neural network. Neural Information Processing, 6803, 451-460.

Orchard, G. (2015). HFirst: A temporal approach to object recognition. IEEE transactions on pattern analysis and machine intelligence, 37(10), 2028-2040.

Orta Alemán, D., & Horne, R. (2020, October). Improved Robustness In Long-term Pressure Data Analysis Using Wavelets and Deep Learning. In SPE Annual Technical Conference and Exhibition? (p. D021S009R002). Society of Petroleum Engineers.

Pahuja, R., & Kumar, A. (2021). Sound-spectrogram based automatic bird species recognition using MLP classifier. Applied Acoustics, 180, 108077.

Palatucci, M. (2009). Zero-shot learning with semantic output codes. Carnegie Mellon University.

Pan Z, Chua Y, Wu J, Zhang M, Li H, Ambikairajah E. An Efficient and Perceptually Motivated Auditory Neural Encoding and Decoding Algorithm for Spiking Neural Networks. Front Neurosci. 2020 Jan 22;13:1420. doi: 10.3389/fnins.2019.01420.

Pan, Z., Zhang, M., Wu, J., Wang, J., & Li, H. (2021). Multi-tone phase coding of interaural time difference for sound source localization with spiking neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 2656-2670.

Pang, C., Liu, H., & Li, X. (2019). Multitask learning of time-frequency CNN for sound source localization. IEEE Access, 7, 40725-40737.

Pérez-López, A., Fonseca, E., & Serra, X. (2019). A hybrid parametric-deep learning approach for sound event localization and detection. arXiv preprint arXiv:1908.10133.

Pfeiffer, M., & Pfeil, T. (2018). Deep learning with spiking neurons: opportunities and challenges. *Frontiers in neuroscience*, *12*, 409662.

Phan, H., Koch, P., Katzberg, F., Maass, M., Mazur, R., & Mertins, A. (2017). Audio scene classification with deep recurrent neural networks. arXiv preprint arXiv:1703.04770.

Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. In 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1-6). IEEE.

Planinec, V., Reijniers, J., Horvat, M., Peremans, H., & Jambrošić, K. (2023). The Accuracy of Dynamic Sound Source Localization and Recognition Ability of Individual Head-Related Transfer Functions in Binaural Audio Systems with Head Tracking. Applied Sciences, 13(9), 5254. https://doi.org/10.3390/app13095254

Politis, A. (2020). A Dataset of Reverberant Spatial Sound Scenes with Moving Sources for Sound Event Localization and Detection. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020). Tokyo, Japan.

Ponulak, F. (2010). Supervised learning in spiking neural networks with ReSuMe: sequence learning, classification, and spike shifting. Neural computation, 22(2), 467-510.

Portello, A., Danes, P., & Argentieri, S. (2011, September). Acoustic models and Kalman filtering strategies for active binaural sound localization. In 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 137-142). IEEE.

Praveen, R. G., de Melo, W. C., Ullah, N., Aslam, H., Zeeshan, O., Denorme, T., ... & Granger, E. (2022). A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2486-2495).

Presannakumar, K., & Mohamed, A. (2023). Deep learning based source identification of environmental audio signals using optimized convolutional neural networks. Applied Soft Computing, 143, 110423.

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S. Y., & Sainath, T. (2019). Deep learning for audio signal processing. IEEE Journal of Selected Topics in Signal Processing, 13(2), 206-219.

Rascon, C., & Meza, I. (2017). Localization of sound sources in robotics: A review. Robotics and Autonomous Systems, 96, 184-210.

Reid, A. (2017). Directional hearing at the micro-scale: bio-inspired sound localization.

Reynolds, J. J. (2019). Reservoir Computing in an Evolutionary Neuromorphic Framework.

Rhinehart, T. A., Chronister, L. M., Devlin, T., & Kitzes, J. (2020). Acoustic localization of terrestrial wildlife: Current practices and future opportunities. Ecology and Evolution, 10(13), 6794-6818.

Risoud, M., (2018). Sound source localization. European annals of otorhinolaryngology, head and neck diseases, 135(4), 259-264.

Roozbehi, Z., Narayanan, A., Mohaghegh, M., & Saeedinia, S.-A. (2024). Dynamic-Structured Reservoir Spiking Neural Network in Sound Localization. IEEE Access.

Salehinejad, H., Sankar, S., Barfett, J., Colak, E., & Valaee, S. (2017). Recent advances in recurrent neural networks. arXiv preprint arXiv:1801.01078.

Sang, J., Park, S., & Lee, J. (2018). Convolutional recurrent neural networks for urban sound classification using raw waveforms. In 2018 26th European Signal Processing Conference (EUSIPCO) (pp. 2444-2448). IEEE.

Sarker, I. H. (2021). Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Computer Science, 2(6), 420.

Schliebs, S., Hamed, H. N. A., & Kasabov, N. K. (2011, November). Reservoir-Based Evolving Spiking Neural Network for Spatio-temporal Pattern Recognition. In ICONIP (2) (pp. 160-168).

Schrauwen, B., Verstraeten, D., & Van Campenhout, J. (2007). An overview of reservoir computing: Theory, applications, and implementations. In Proceedings of the 15th European Symposium on Artificial Neural Networks (pp. 471-482).

Seung, H. S. (2003). Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. Neuron, 40(6), 1063-1073.

Shaukat, M. A. (2021). Cluster Analysis and Model Comparison Using Smart Meter Data. Sensors, 21(9), 3157.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of big data, 6(1), 1-48.

Simon HJ, Levitt H. Effect of dual sensory loss on auditory localization: implications for intervention. Trends Amplif. 2007 Dec;11(4):259-72. doi: 10.1177/1084713807308209.

Soures, N., & Kudithipudi, D. (2019). Spiking reservoir networks: Brain-inspired recurrent algorithms that use random, fixed synaptic strengths. IEEE Signal Processing Magazine, 36(6), 78-87.

Substance Abuse and Mental Health Services Administration. (2019). Communicating in a Crisis. Rockville, MD.

Sun, A. Y., & Scanlon, B. R. (2019). How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environmental Research Letters, 14(7), 073001.

Sun, Y. (2017). Indoor sound source localization with probabilistic neural network. IEEE Transactions on Industrial Electronics, 64(8), 6403-6413.

Takahashi, T. (2021). Moving Sound Source Tracking in Wide Space by Multiple Microphone Arrays. In 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE.

Takeda, R., & Komatani, K. (2016, March). Sound source localization based on deep neural networks with directional activate function exploiting phase information. In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 405-409). IEEE.

Tan, T. H., Lin, Y. T., Chang, Y. L., & Alkhaleefah, M. (2021). Sound source localization using a convolutional neural network and regression model. Sensors, 21(23), 8031.

Tanaka, G. (2019). Recent advances in physical reservoir computing: A review. Neural Networks, 115, 100-123.

Tanaka, G., Yamane, T., Héroux, J. B., Nakane, R., Kanazawa, N., Takeda, S., ... & Hirose, A. (2019). Recent advances in physical reservoir computing: A review. Neural Networks, 115, 100-123.

Tang, R. (2022). Efficient energy-based orthogonal matching pursuit algorithm for multiple sound source localization with unknown source count. Measurement Science and Technology, 33(4), 045018.

Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., & Maida, A. (2019). Deep learning in spiking neural networks. *Neural networks*, *111*, 47-63.

Thorpe, S. &. (1998). Rank order coding. The Proceedings of the sixth annual conference on Computational neuroscience, Big Sky, Montana, United States.

Thorpe, S. D. (2001). Spike-based strategies for rapid processing. Neural Networks, 14(6-7), 715-725.

Valenti, M., Tonelli, D., Vesperini, F., Principi, E., & Squartini, S. (2017, August). A neural network approach for sound event detection in real life audio. In 2017 25th European Signal Processing Conference (EUSIPCO) (pp. 2754-2758). IEEE.

Vesperini, F. (2016). A neural network based algorithm for speaker localization in a multi-room environment. In 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1-6). IEEE.

Villa, A. E. (1999). Spatiotemporal activity patterns of rat cortical neurons predict responses in a conditioned task. Proceedings of the National Academy of Sciences, 96(3), 1106-1111.

Vliegen, J., Van Grootel, T. J., & Van Opstal, A. J. (2004). Dynamic sound localization during rapid eye-head gaze shifts. Journal of Neuroscience, 24(42), 9291-9302.

Von Der Malsburg, C. (1994). The correlation theory of brain function. New York, NY: Springer.

Wall, J. A., McDaid, L. J., Maguire, L. P., & McGinnity, T. M. (2012). Spiking neural network model of sound localization using the interaural intensity difference. IEEE transactions on neural networks and learning systems, 23(4), 574-586.

Wang, D. (2006). Computational auditory scene analysis: Principles, algorithms, and applications. Wiley-IEEE Press.

Wang, D. and K. C. Soh (2017). "Sound source localization using microphone arrays: A review." IEEE Signal Processing Magazine 34(4): 113-126.

Wang, L. (2021). Robotic odor source localization via adaptive bio-inspired navigation using fuzzy inference methods. Robotics and Autonomous Systems.

Wang, L., & Cavallaro, A. (2022). Deep-learning-assisted sound source localization from a flying drone. IEEE Sensors Journal, 22(21), 20828-20838.

Wang, Q., Du, J., Wu, H.-X., Pan, J., Ma, F., & Lee, C.-H. (2023). A four-stage data augmentation approach to ResNet-Conformer based acoustic modeling for sound event localization and detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 1251-1264.

Wikipedia contributors. (2021, August 28). 3D sound localization. Wikipedia. https://en.wikipedia.org/wiki/3D_sound_localization

Willert, V., Eggert, J., Adamy, J., Stahl, R., & Korner, E. (2006). A probabilistic model for binaural sound localization. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 36(5), 982-994.

Wu J, Yılmaz E, Zhang M, Li H, Tan KC. Deep Spiking Neural Networks for Large Vocabulary Automatic Speech Recognition. Front Neurosci. 2020 Mar 17;14:199. doi: 10.3389/fnins.2020.00199. PMID: 32256308; PMCID: PMC7090229.

Wu, J. M. (2019). Competitive STDP-based feature representation learning for sound event classification. In 2019 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

Wu, J., Chua, Y., Zhang, M., Li, H., & Tan, K. C. (2018). A spiking neural network framework for robust sound classification. Frontiers in neuroscience, 12, 836.

Wu, Y., Ayyalasomayajula, R., Bianco, M. J., Bharadia, D., & Gerstoft, P. (2021, June). Sslide: Sound source localization for indoors based on deep learning. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4680-4684). IEEE.

Xie, B. (2013). *Head-related transfer function and virtual auditory display*. J. Ross Publishing.

Xiong, C., Lu, W., Zhao, X., & You, Z. (2022). Miniaturized multi-topology acoustic source localization network based on intelligent microsystem. Sensors and Actuators A: Physical, 345, 113746.

Yalta, N., Nakadai, K., & Ogata, T. (2017). Sound source localization using deep learning models. Journal of Robotics and Mechatronics, 29(1), 37-48.

Yamazaki K, Vo-Ho VK, Bulsara D, Le N. Spiking Neural Networks and Their Applications: A Review. Brain Sci. 2022 Jun 30;12(7):863. doi: 10.3390/brainsci12070863. PMID: 35884670; PMCID: PMC9313413.

Yan, Y. (2018). On the Semidefinite Programming Algorithm for energy-based acoustic source localization in sensor networks. IEEE Sensors Journal, 18(21), 8835-8846.

Yiwere, M., & Rhee, E. J. (2019). Sound source distance estimation using deep learning: An image classification approach. Sensors, 20(1), 172.

Yu M, Xiang T, P S, Chu KTN, Amornpaisannon B, Tavva Y, Miriyala VPK, Carlson TE. A TTFS-based energy and utilization efficient neuromorphic CNN accelerator. Front Neurosci. 2023 May 5;17:1121592. doi: 10.3389/fnins.2023.1121592. PMID: 37214405; PMCID: PMC10198466.

Yu, Q. (2015). A spiking neural network system for robust sequence recognition. IEEE Transactions on Neural Networks and Learning Systems, 26(3), 621-635.

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. Neural Computation, 31(7), 1235-1270.

Yudanov, D. M. (2010). GPU-based simulation of spiking neural networks with real-time performance & high accuracy. In The 2010 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

Zhang, L., et al. (2021). "Sound source localization and tracking based on machine learning: A survey." Signal Processing 186: 108187.

Zhang, Y., Xiong, R., He, H., & Pecht, M. G. (2018). Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries. IEEE Transactions on Vehicular Technology, 67(7), 5695-5705.

Zhang, Y., Zeng, Z., & Tang, D. (2021, August). Application of Neural Network and Kalman Filtering in Sound Source Localization. In 2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA) (pp. 5-8). IEEE.

Zhao, B. (2014). Feedforward categorization on AER motion events using cortex-like features in a spiking neural network. IEEE Transactions on Neural Networks and Learning Systems, 25(11), 1963-1978.

Zhao, S. (2012). A real-time 3D sound localization system with a miniature microphone array for virtual reality. 7th IEEE Conference on Industrial Electronics and Applications (ICIEA).

Zhuo, D. B., & Cao, H. (2021). Fast sound source localization based on SRP-PHAT using density peaks clustering. Applied Sciences, 11(1), 445.

Appendices

Appendix A : Maximum Error Boundary Calculation

Our methodology relies on an energy-based framework, wherein the manipulation of Leaky Integrate-and-Fire (LIF) neuron voltages assumes a pivotal role. To offer a familiar analogy, we conceptualize observed neurons as akin to input-output (I/O) entities, drawing parallels with loudspeakers. This analogy finds its grounding in the inherent resemblance between the LIF neuron model and the dual loudspeaker lumped model.

In the realm of our proposed algorithm, the behavior of the LIF neuron aligns with the dual circuit of a loudspeaker's lumped element model. Specifically, the equation Lumped model of load speaker captures the dynamics of the electrical circuit, mirroring how the LIF neuron adjusts its voltage in response to incoming signals. This logical connection between the LIF model and the dual loudspeaker lumped model underpins our energy-based framework.

Expanding further, the adjustment of LIF neuron voltage stands out as a pivotal element of our approach. This voltage modulation, reminiscent of tuning parameters in a loudspeaker, facilitates dynamic adaptation to received signals. The presented energy-based law provides a systematic method to enhance energy simulation by manipulating LIF neuron parameters. Consequently, the error boundary of our proposed algorithm is intricately tied to these tunable parameters, allowing for fine-tuning through controlled experimentation on a limited dataset.

For calculating the maximum distance error, we consider the lumped model of the load speaker, as denoted above:

$$L_c \frac{di}{dt} + Ri + Bl.\frac{dx}{dt} = E(t) \qquad (A-1)$$

$$V + \frac{R}{C}\frac{dV}{dt} + Bl.\frac{dx}{dt} = E(t) \qquad (A-2)$$

$$\tau \frac{dV}{dt} = -V \underbrace{-Bl.\frac{dx}{dt} + E(t)}_{I} = -V + I \rightarrow LIF\ Model \quad and \quad I \propto \text{ Sound Energy} \qquad (A-3)$$

$$I = \tau \frac{dV}{dt} + V \rightarrow V = k.e^{\frac{\Delta t}{\tau}}.I \rightarrow I = W.V.e^{-\frac{\Delta t}{\tau}} \qquad (A-4)$$

$$y_{sound\_Energy} = \frac{S}{\left|d - d_s\right|^2} + \varepsilon \approx \frac{S}{\left|d\right|^2} = W.V.e^{-\frac{\Delta t}{\tau}} \qquad (A-5)$$

$$Taylor\ \text{Expansion}\{e^{-\frac{\Delta t}{\tau}} = 1/\exp(\frac{\Delta t}{\tau})\} \approx \frac{1}{1 + \frac{\Delta t}{\tau} + \frac{1}{2}\left(\frac{\Delta t}{\tau}\right)^2} \approx \frac{1}{d^2} \qquad (A-6)$$

if $\frac{1}{\tau}$ = sound wave spead=c $\rightarrow$ we expect that : cΔt ≈ d in the best Δt calculation, (TDOA)

Under this assumption, Error boundary is calculated as follows:

$$\left|Er\right| : \left| d^2 - \frac{1}{2}\left(d^2 + 2d + 2\right) \right| = \left| \frac{1}{2}\left(d^2 - 2d - 2\right) \right| = \frac{1}{2}\left|(d-1)^2 - 3\right| \qquad (A-7)$$

According to our dataset, maximum of d is 3, So:

$\left|Er\right| \leq 0.5$