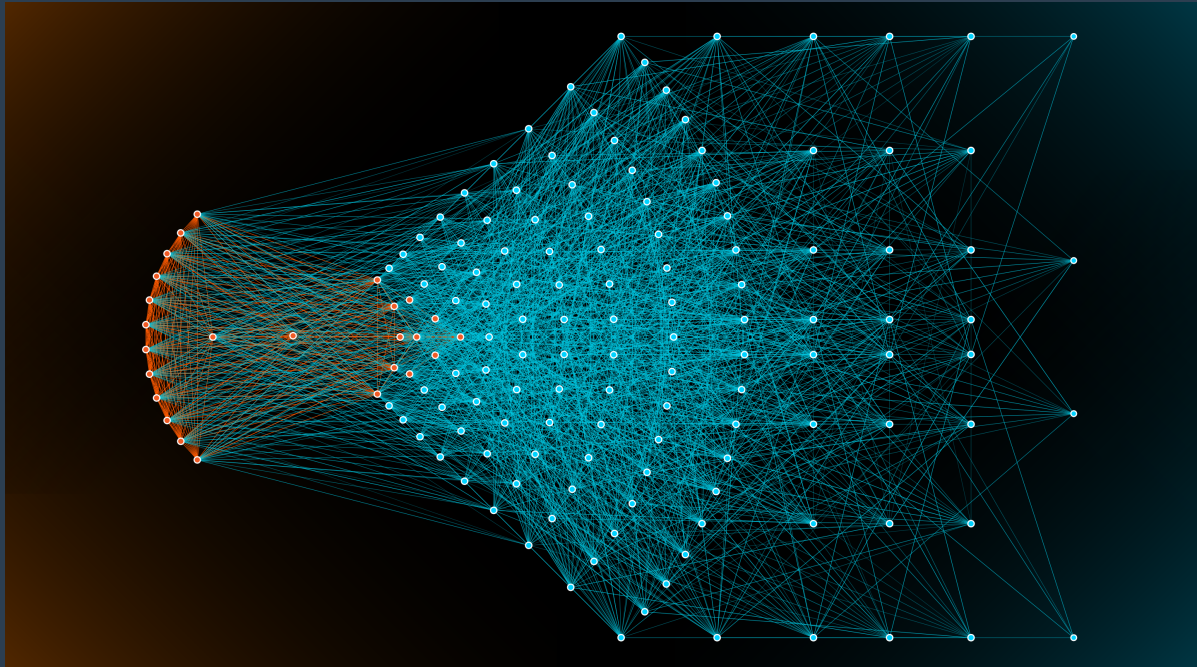


# Description of Work

Sound Detection and Classification  
using Spiking Neural Networks



COURREGE Téo  
GANDEEL Lo'aï

Date: November 22, 2023

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Chosen theme</b>	<b>2</b>
2.1	The SNN	2
2.1.1	Definition	2
2.1.2	Advantages / Disadvantages	3
2.1.3	snnTorch	3
2.2	Neural networks applied to audio classification	4
2.2.1	Definition	4
2.2.2	How it works	4
2.3	Relevance of this choice	4
2.3.1	Defining the project's pros and cons	4
2.3.2	Cases of use to define the relevance of the project	5
<b>3</b>	<b>Data used</b>	<b>5</b>
3.1	Choosing the type of training data	5
3.2	Feasibility	5
3.2.1	Copyright	5
3.2.2	Data collection/extraction	5
3.2.3	Uncertainties about data quality	6
3.3	Open to other databases	6
<b>4</b>	<b>State of the art</b>	<b>6</b>
4.1	ANNs Used for Audio Classification	6
4.2	SNNs Used for Audio Classification	6
<b>5</b>	<b>Technical objectives of the project</b>	<b>7</b>
5.1	Technical objectives achievable with pre-existing ANNs	7
5.2	Technical objectives with data	7
5.3	Technical objectives with SNNs	7
5.4	Feasibility	7
5.4.1	Describing uncertainties	7
5.4.2	Looking for a working model	8
5.4.3	Estimate computation time	8
<b>6</b>	<b>Theoretical study</b>	<b>8</b>
6.1	Theoretical study on SNN	8
<b>7</b>	<b>Matching with available resources</b>	<b>8</b>
7.1	Work time	9
7.2	Computer resources	9
7.3	Human resources	9
<b>8</b>	<b>Appendix</b>	<b>10</b>
8.1	Data	10
8.2	Existing models	10
8.3	Code	10
8.4	Links	10
	<b>References</b>	<b>11</b>

# 1 Introduction

In the constantly evolving landscape of neural network studies, the project outlined in this Description of Work (DoW) proposes to explore the field of spiking neural networks (SNNs). This effort will at term combine both a study and a prospective implementation, with a focus on the emerging field of neuromorphic computing.

The overall goal of this project is to perform a thorough investigation and subsequent implementation of spiking neural networks. SNNs, which are inspired by the neural signaling patterns of the human brain, show promising potential in various applications, especially in real-time processing and pattern recognition.

Before getting into the specifics of our project, let us first give a brief overview of what spiking neural networks are and how they work:

A Spiking Neural Network (or SNN) is a type of artificial neural network that mimics the functionality of biological neural networks. Unlike traditional neural networks (ANNs - Artificial Neural Networks), SNNs incorporate the concept of time into their operating model. The neurons in SNNs generate spikes of activity and communicate through these spikes (like brain neurons stimulating each other with electrical impulses), allowing them to process information in a more complex and potentially more efficient manner.

This type of neural network is theoretically less energy consuming than some ANNs and potentially more efficient in terms of processing power (in practice, it will be necessary to create an adapted hardware architecture to take advantage of this potential). Therefore, it seems that the use of this type of neural network could be a viable solution to a classification processing problem (especially in real time ones).

Our project focuses on a specific problem within the broader domain of audio classification.

Adopting an exploratory approach, we will begin this DoW with a brief introduction of our topic, followed by a detailed presentation of the dataset used and its construction. We will review the state of the art in the field of SNNs and audio classification. We will then present the technical goals of our project and a simple theoretical study of SNNs. Finally, we will present the resources we have at our disposal to carry out this project, as well as a description of the work to be done. Finally, we will propose a schedule for the project.

## 2 Chosen theme

### 2.1 The SNN

#### 2.1.1 Definition

Spiking Neural Network (SNN) is a variant of artificial neural network that try to mimic more closely the biological neural networks. More precisely, it tries to copy the functioning of neurons and synapses. As a result, rather than working with continually changing time values as ANNs do, SNNs work with discrete events that happen at defined times. SNNs take a set of spikes as input and produces a set of spikes as output

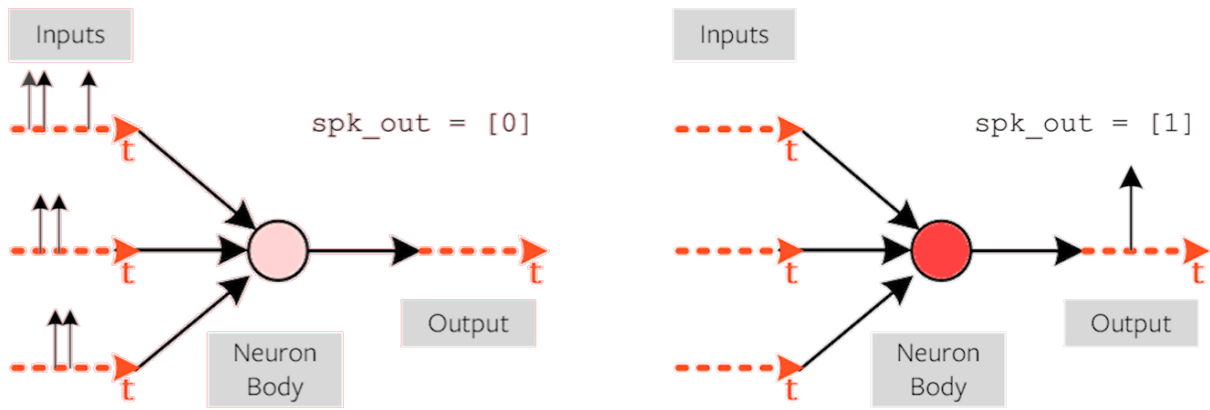
The spiking behavior of a neuron in an SNN is modeled by the membrane potential equation:

$$V(t) = \sum_i W_i \cdot \delta(t - t_i)$$

Here,  $V(t)$  is the membrane potential,  $W_i$  is the synaptic weight, and  $\delta(t - t_i)$  is the Dirac delta function representing the occurrence of a spike at time  $t_i$ .

The output spikes  $Y$  can be obtained based on a threshold  $\theta$ :

$$Y(t) = \begin{cases} 1 & \text{if } V(t) \geq \theta \\ 0 & \text{otherwise} \end{cases}$$



### 2.1.2 Advantages / Disadvantages

SNNs, when well constructed, should offer the following benefits:

- Energy efficiency: SNNs are highly energy efficient because they only consume energy when a spike is generated, unlike "classical" neural networks which consume energy continuously. This makes them ideal for low-power embedded devices.
- Event-driven processing: While ANNs typically operate on a fixed time interval, SNNs only process information when there is an input spike. This makes them highly efficient at processing information in an event-driven environment, such as processing visual or audio data.
- Temporal coding: SNNs are able to process information based on the timing of spikes, allowing them to encode temporal information and be temporally accurate. This is particularly useful for tasks that require the processing of sequential data, such as speech or gesture recognition.
- Robustness to noise: SNNs are inherently robust to noise and can effectively filter out irrelevant information. This makes them useful in environments where noise is a common problem, in our case it may be an instance of audio not well recorded.
- Neuroplasticity: SNNs are able to adapt to new inputs and change their behaviour over time, much like the brain. This allows them to learn and adapt to new tasks and environments, making them highly versatile.

Disadvantages:

- SNNs are difficult to train, mainly due to backpropagation issues, which are caused by the discrete, temporal nature of spikes. Backpropagation in time (BPTT) is the typical method used, but it suffers from issues such as the temporal credit assignment problem (the determination of which past events or actions contributed most significantly to the current outcome or state)
- There is currently no learning algorithm designed specifically for this task, SNNs lack the "maturity" of more classic neural networks.
- Building a small SNN is impractical. Indeed, we need a sufficient number of neurons to retrieve useful informations.

We will try to highlight these specificities during our project

### 2.1.3 snnTorch

We will primarily work in Python and use the snnTorch package to work on SNNs. snnTorch is a Python package for performing gradient-based learning with spiking neural networks. snnTorch is built on top of Pytorch and takes advantage of its GPU-accelerated tensor computation. Pre-defined spiking neuron models are integrated into the PyTorch framework and can be treated as recurrent activation units.

## 2.2 Neural networks applied to audio classification

### 2.2.1 Definition

Audio classification is a machine learning and signal processing task that involves categorizing or labeling audio data into different predefined classes or categories. The goal is to automatically assign a label to an audio segment based on its content or characteristics. This can be useful in various applications :

- Speech Recognition: Identifying spoken words or phrases in audio recordings.
- Music Genre Classification: Categorizing music into different genres, such as rock, pop, jazz, etc.
- Environmental Sound Analysis: Detecting and classifying sounds in the environment, such as sirens, footsteps, or car engines.
- Anomaly Detection: Identifying unusual or unexpected sounds in a given context.

### 2.2.2 How it works

The process of audio classification is relatively similar to image classification and typically involves the following steps

Data collection: Collecting a dataset of audio samples, each labeled with its corresponding class or category.

Feature extraction: Transforming the raw audio data into a set of relevant features that can be used to represent the content of the audio. Common features include spectrogram representations, Mel Frequency Cepstral Coefficient, and other time or frequency domain features.

Model training: Using machine learning algorithms or deep learning architectures to train a model on the extracted features and their corresponding labels. Apart from neural networks, others algorithms can be used for audio classification such as support vector machines, decision trees, or random forests for example.

Validation and testing: Evaluate the performance of the trained model on a separate set of data that it has not seen before. This helps evaluate the model's ability to generalize to new, unseen examples.

Deployment: Integrate the trained model into applications or systems that require real-time or batch audio classification.

## 2.3 Relevance of this choice

### 2.3.1 Defining the project's pros and cons

Pros :

- SNN's can be effective to detect and recognize different types of audio (especially on embedded devices).
- The comparison between SNN and other types of neural networks will be particularly important to us, and we will try to show what can be the strengths and drawbacks of SNNs compared to other networks.

Cons:

- SNNs are new to us, we need to understand the package `snnTorch`.
- Training SNNs is harder, and we will need to find the most efficient encodings/decodings, training method, and parameters.

### 2.3.2 Cases of use to define the relevance of the project

Our final goal is primarily to compare the efficiency and energy use of SNNs compared to others networks (CNNs, RNNs, LSTMs...) in audio classification, and to evaluate if SNNs are a good solution in our case.

## 3 Data used

### 3.1 Choosing the type of training data

When we first read the survey about *SNNs*<sup>[1]</sup>, it was understood that training this type of model would be more difficult than some ANNs due to the temporal dynamics involved in SNNs. This made us want to work on a type of data that would be less computationally or resource intensive than videos (so that we could store and train our models on many of them). This type of data also has the advantage of being more adaptable to pre-recorded and real-time processing.

### 3.2 Feasibility

In our case, we decided to apply our analysis to the [Google AudioSet](#) database, which contains a large number of audio samples classified into different categories. The structure of this dataset is as follows: each audio sample is associated with a label, and each label is associated with a set of audio samples. In practice, this dataset is composed of '.csv' files that contain one or more labels, a timecode (10 seconds of duration), and a part of an url that would help us find an associated Youtube video. This gave us 2 problems: a copyright problem and the fact of extracting the data and formatting it in a way that would be usable for our project.

#### 3.2.1 Copyright

According to [the AudioSet website](#) (download page): "The dataset is made available by Google Inc. under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, while the ontology is available under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license".

#### 3.2.2 Data collection/extraction

Fortunately for us, we found [some GitHub repositories](#) whose code aims to download, format, and crop the various sound files so that it "unifies the formats of the datasets so that it is possible to analyze them in the same pipelines". The given script also included a parallelization of the calculation.

After some slight modifications we were able to check the impact on memory resources, for example for [this youtube video](#) we get

- Resource usage (RAM, GPU, CPU, time)
  - Size of 144p: 2-4 MB (depending on format)
  - Size of the same video as an audio file (wav in best quality): 1-3 MB (depending on the format)

Without parallelizing the code we used, we were able to download, format, and trim 6 audio files in about 60 seconds (and our first attempts at parallelizing seemed promising, though there are still bugs to work out). At this pessimistic rate, we think we should be able to download, format and trim about 2000 10-second audio files in about 5-6 hours. By limiting the number of labels to study to about ten, we think we might get some results (to be tested with a pre-trained Pytorch model).

### 3.2.3 Uncertainties about data quality

While looking up to the data, we found that a lot of these sounds are taken out of context and other sounds are mixed with them (for example, [this](#) sample is labeled with as a music sound but a person is talking at the end of the sample). Also, each sample can have one or multiple labels, which can be related to multi-clustering problems. Another problem we face is the fact that the data is always balanced (there exists a balanced version of the dataset) and that there are some Weak and Strong Label annotations.

Strong labels are precise and indicate the exact start and end times of the event in the audio file whereas the Weak labels, only indicate the presence of a sound event somewhere in the audio file, without specifying the exact time.

## 3.3 Open to other databases

In the case that we use for this classification purpose is not sufficient (for example when collecting all the data related to a unique label), We might try to complete it with other open-source label databases such as [Freesound](#) or [Kaggle](#).

## 4 State of the art

While searching for existing work on the topic, we found that there are already some studies on the subject of audio classification using SNNs. We will present a couple of them here.

### 4.1 ANNs Used for Audio Classification

First of all, we made some researches to find some existing ANNs made for audio classification for comparison with the model we plan to build. We found the following models:

- A VGG model already trained of the Google AudioSet dataset ([GitHub](#)).
- Rearranged Resnet, inception, densenet pretrained models ([GitHub](#) and [research paper](#))
- An audio classifier that uses LSTM (made in class of Deep Learning): ([Google Colab](#))
- others: [1](#) [2](#) [3](#) [4](#)

### 4.2 SNNs Used for Audio Classification

Secondly, we made some researches to find some existing SNNs made for audio classification, again, for comparison with the model we plan to build. We found the following models:

- A spiking convolutional neural network (SCNN) for audio classification ([GitHub](#)).
- A multi-layer SNN for audio classification using SpiNNaker ([GitHub](#)).
- Shadow training ("conversion" of an ANN to an SNN) ([GitHub](#) and [research paper](#)).
- Spiking used for image classification ([GitHub](#) and [doc](#)).
- SNN simulators : [BindsNET](#) and [NEST](#)

The results we may get from the aforementioned networks may serve as a basis of comparison for some of them.



## 5 Technical objectives of the project

### 5.1 Technical objectives achievable with pre-existing ANNs

We should find pre-trained ANN's that can attain a good accuracy with AudioSet.

### 5.2 Technical objectives with data

In its raw form, a sound is an analog signal, needing to be converted to digital. Analog signals are continuous waves. An analog wave has a discrete value at any given point in time. It's impossible to convert every value to a digital representation because there are infinite points in a continuous wave. You must convert the audio signal to digital through sampling and quantization. In our case, since we extract the audio from Youtube, the conversion step is already done.

After conversion, the next step is to extract useful features from the audio. You can extract many features, so you must focus on the specific problem you want to solve. Here are some critical features used for audio classification with ML:

Time domain features are extracted directly from the raw audio (waveform). Examples are the amplitude envelope, and the root mean square energy. Time domain features are not enough to represent the sound because they do not include frequency information.

Frequency domain features are also called the spectrum. Examples include band energy ratio, spectral flux, and spectral bandwidth. These features lack time representation. They are extracted from a time domain representation through Fourier transform. Fourier transform converts a waveform from a function of time to a function of frequency.

Time and frequency domain features extract a spectrogram from the waveform using a short-time Fourier transform, like Spectrogram or Mel Spectrogram. The spectrogram shows both the frequency and time domain. Spectrograms represent frequency linearly, but humans perceive sounds logarithmically. This means that humans can tell the difference between lower frequencies like 500-1000 Hertz (Hz) but find it harder to differentiate between sounds at a higher frequency, from 10,000 Hz to 15,000 Hz. This difficulty is why the Mel-scale was introduced. A Mel Spectrogram is a spectrogram on the Mel-scale, which measures how different pitches sound to the listener. It maps pitches that, to listeners, sound equidistant.

### 5.3 Technical objectives with SNNs

We need to either find a pre-trained model or create it ourselves. We found a framework for the creation of a SNN for robust sound classification.[?]

## 5.4 Feasibility

### 5.4.1 Describing uncertainties

Uncertainties are mostly due to the dynamic nature of audio signals, varying acoustic environments, and the complexity of distinguishing between diverse sound patterns. Additionally, challenges in precisely modeling the temporal aspects of auditory processing and the understanding of spiking neuron dynamics contribute to many uncertainties in the feasibility of the project. Moreover, the training of Spiking Neural Networks in order to attain a good accuracy is difficult due to the inherent non-linearity and sparsity of spikes.

Regarding uncertainties about the data, dealing with Youtube, means some audio may not be in the right format (a 5s video where we extract a 10s audio for example), and some might not even be available (users deleting their video), which might lead to unbalanced data.



### 5.4.2 Looking for a working model

Many model for sound classification already exist. In particular, for AudioSet, there are already multiple pre-trained deep neural network, for example YamNet. However, it is more difficult to find open source SNNs, as they are less widespread than other types of neural networks.

### 5.4.3 Estimate computation time

Importation of data and training are the most consuming task. We don't know yet how long a training will be. As for the importation of audio, we estimate a little more than 10 seconds on average for the importation of a single audio sample without parallelization, we expect to import between 1.000 and 10.000 audio samples.

## 6 Theoretical study

### 6.1 Theoretical study on SNN

There are multiple archetypes of spiking, each model as its specificity, and while it seems interesting to restraint ourselves to the LIF (Leaky Fire and Integrate) which seems the more interesting, it may be relevant to analyze some types of model to evaluate their efficiency. [?]

We will need to choose multiple parameters regarding our neural network, like the encoding, the type of training.

Neural coding schemes are used to convert input pixels into spikes that are transmitted to the excitatory neurons. There are four main neural coding schemes: rate coding, time-to-first spike coding, phase coding, and burst coding.[?]

There are also multiple ways to train a SNN:

- Shadow training: A non-spiking ANN is trained and converted into an SNN by interpreting the activations as a firing rate or spike time.
- Backpropagation using spikes: The SNN is natively trained using error backpropagation, typically through time as is done with sequential models.
- Local learning rules: Weight updates are a function of signals that are spatially and temporally local to the weight, rather than from a global signal as in error backpropagation.

## 7 Matching with available resources

## 7.1 Work time

Schedule			
Task	Duration	When	Who
<b>Data collection/extraction</b> From csv/Youtube to mp3	1-2 days	22/11-24/11	Loai
<b>Data preprocessing</b> cleaning/enhancement/ augmentation	3-4 days	24/11-28/11	
<b>Models implementation</b> understanding snnTorch basics	1-2 days	24/11-25/11	
finding pretrained/creating and tuning models	4-5 days	26/11-01/12	
<b>Training/tuning</b> training	multiple weeks		
extracting/synthetizing info	multiple weeks		
<b>Working on the DOWs and presentation</b> DOWs writing: continous work	2 hours x 2 per week		
slides for oral presentation	2-3 days	27/11-29/11	

## 7.2 Computer resources

We will work with the following:

- Hardwares: 2 laptops
- Softwares: Google Colab (primarily), Kaggle

## 7.3 Human resources

We will work in parallel most of the time, each one focusing on a different aspect to gain time when it is possible.

## 8 Appendix

### 8.1 Data

#### Data Set used for the project

- Google AudioSet : <https://research.google.com/audioset/>

### 8.2 Existing models

- Sparsity through Spiking Convolutional Neural Network (SCNN) for Audio Classification at the Edge : <https://github.com/CongSheng/SpikingConvNN-AudioClassification/tree/main>
- Multilayer Spiking Neural Network for audio samples classification using SpiNNaker : <https://github.com/jpdominguez/Multilayer-SNN-for-audio-samples-classification-using-SpiNNaker>
- StereoSpike: Depth Learning With a Spiking Neural Network for image classification (CNN) : <https://github.com/urancon/stereospike> - <https://ieeexplore.ieee.org/document/9969606>
- Spiking Neural Networks Trained via Proxy (shadow training): <https://ieeexplore.ieee.org/document/9810220>
- VGG model trained on Google's AudioSet dataset : <https://github.com/MaryamEbr/Audio-Classification-with-AudioSet-and-VGGish/tree/main> - <https://research.google/pubs/pub45611/>
- Rethink CNNs for Audio Classification : <https://github.com/kamalesh0406/Audio-Classification/tree/master> - <https://arxiv.org/abs/2007.11154>
- An LSTM model for audio classification we made in class of Deep Learning : [https://colab.research.google.com/drive/1kJ4hFfh3J13\\_nrtCMRrPoE\\_eI15xUQWC?usp=sharing](https://colab.research.google.com/drive/1kJ4hFfh3J13_nrtCMRrPoE_eI15xUQWC?usp=sharing)
- BindsNET, a Python package for training and simulating spiking neural networks : <https://github.com/BindsNET/bindsnet>
- NEST, a simulator for spiking neural network models that focuses on the dynamics, size and structure of neural systems rather than on the exact morphology of individual neurons : <https://www.nest-simulator.org/>

### 8.3 Code

- A collection of scripts to analyze, prepare for and download Google's Audioset : <https://github.com/bakhtos/GoogleAudioSetScripts>

### 8.4 Links

## References

- [1] J. K. ESHRAGHIAN, M. WARD, E. NEFTCI, X. WANG, G. LENZ, G. DWIVEDI, M. BEN-NAMOUN, D. S. JEONG, AND W. D. LU, *Training Spiking Neural Networks Using Lessons From Deep Learning*. <https://arxiv.org/abs/2109.12894>, 2021.