Sound Source Localization and Distance Estimation in Open Environment using Simulation and AI

Master Thesis

Denis Rosset

University of Applied Sciences and Arts Western Switzerland

July 2023

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Supervisors:

Michael Mäder: Professor in computer science
Beat Wolf Professor: in computer science

Principals

Marc-Antoine Fénart: Professor in civil engineering
Gabriel Python: Scientific associate at Rosas

Acknowledgements

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.[**?**] [**?**]

Contents

1
Introduction

Within the framework of the research project ”NPR Teleoperation”, the engineers of the HEIA-FR have developed the first concept in Switzerland of a remote-controlled automated vehicle. However, teleoperation only makes sense if the vehicle is automated. There can be no teleoperation without automation (economic factors) just as there can be no automation without teleoperation (legal, technical, and social factors). ROSAS then created the Autovete (Automatisation de véhicules téléopérés) project, financed by HEIA-FR, to build up vehicle automation expertise. For a vehicle to be fully autonomous, the detection of other emergency vehicles is mandatory. To solve this issue, V2V (Vehicle-to-Vehicle) communication can be used but is not yet integrated into emergency vehicles. So, to be able to detect such a vehicle, two signals need to be processed: the sound of the emergency siren and the blinking lights of the vehicle. The first use case of this project focuses only on sound source distance estimation and localization.

To understand if sound source estimation and localization could work for emergency detection, a simpler use case has been created for this project. It is the detection of excessively noisy vehicles on the street. The goal is to measure the sound level of the passing vehicles and compare it with the legal limits. If a vehicle exceeds the limit, the system can record its license plate and report it to the authorities. This way, the system can help reduce noise pollution and improve road safety. To implement this use case, the system requires a microphone array, a camera, and a processing unit. The microphone array captures the sound signals from different directions and sends them to the processing unit. The processing unit applies a sound source localization algorithm to estimate the direction and distance of the sound source. The camera captures the image of the vehicle and performs license plate recognition. Big improvements in sound source localization with the help of machine learning are being made1 and can be used to reliably localize the origin of a sound using one or more microphone arrays (multiple microphones operating in tandem).

A non-negligible problem is that the number of real-world datasets with moving sources in an open environment is limited. A solution is to create the datasets in realistic sound propagation simulation. To validate and use the model, it should also be tested to see how it reacts against adversarial attacks, understand how it can be used in a real environment and limit the attack vector.

1.1 Motivation

1.1.1 Objectives

1.1.1.1 Objective n°1 Dataset according to the baseline

The first objective is to construct a dataset that is coherent with the project's baseline. The dataset should contain the target variable, features, and necessary pre-processing steps such as missing data imputation, data normalization, and feature engineering. This dataset will help create and understand the problem.

1.1.1.2 Objective n°2 Model for better sound source localization and distance estimation

The project should use a neural network model to detect the origin of a sound using a microphone array. The neural network should be trained using the dataset created in objective 3.1 and should be able to accurately localize the sound source. The trained neural network model should be evaluated in a real environment to see how it performs. It should also be evaluated to see how dependable it is in localizing sound sources and how it can be improved.

1.1.1.3 Objective n°3 The model should be tested to see how it reacts to attacks

The trained neural network model needs to be evaluated by testing it on data that has been modified in some way, such as by adding or removing noise, or by modifying the sound source. The model could also be tested against various types of attacks, such as masking, time-warping, and frequency-shifting.

1.1.2 Challenges

1.1.2.1 Challenge n°1: Realistic datasets

One of the main challenges in this project is to construct realistic datasets that accurately capture the data in a real-world scenario. The lack of open-source datasets that contain moving sound sources in open-loop environments can make this difficult. To overcome this challenge, the project should use realistic simulations of sound propagation to produce suitable datasets.

1.1.2.2 Challenge n°2: Robust models

Another challenge is to create a neural network model that can accurately detect the origin of a sound using a microphone array. The trained model should be highly accurate and robust enough to resist adversarial attacks. To assess the model's robustness, the model should be tested on data that has been modified in some way, such as by adding or removing noise, or by modifying the sound source. The model should also be able to accurately localize the sound source despite the attack.

1.1.2.3 Challenge n°3: Adequate evaluation metrics

The last challenge is to create an evaluation metric that adequately reflects the model's performance. The evaluation metric should take into account the accuracy of sound source localization in a real environment as well as its ability to resist adversarial attacks. The metric should also be able to capture how well the model can provide reliable results in a variety of environments.

1.2 Structure of the thesis

- **Chapter 2: Background**: This chapter provides an overview of the background knowledge necessary to understand the project, such as the theory of machine learning and sound propagation, and a brief introduction of sound source localization.
- **Chapter 3: Methodology**: This chapter explains the methodology used in the project, such as creating a dataset, creating a simulation, building a neural network model, and evaluating it using an appropriate metric.
- **Chapter 4: Setup**: This chapter describes the work done on the project, such as the simulations used to create the dataset, the neural network model used, and the evaluation metrics used.
- **Chapter 5: Results**: This chapter presents the results of the project, such as the performance of the neural network model and the evaluation metrics.
- **Chapter 6: Conclusion and future work**: This chapter concludes this project with a discussion of the main results and a summary of the key findings.

¡¡

2
Background and Literature

This chapter introduces technical concepts and background used in the conceptualized solution of the thesis. It also explains the state of the art of the different technologies used in the thesis and the current state of research in sound source localization and distance estimation.

2.1 Sound Propagation

Sound propagation is the physical process by which sound waves propagate in a given environment. The strength of the sound wave depends on various factors, including the frequency, environment, and distance from the sound source. These factors make an accurate identification and localization of a sound source difficult; thus, a more accurate and robust sound source localization system is needed.

2.1.1 Realistic sound propagation in simulations

2.1.2 Microsoft Project Acoustics

Microsoft Project Acoustics is a sound propagation engine that simulates the propagation of sound waves in a given environment. It is used in various applications, including video games, virtual reality, and augmented reality. It simulates wave effects like obstruction, reverberation, and occlusion in complex 3D scenes without requiring zone markup or raytracing. It works similarly to a raytracing engine but is precomputed and optimized for real-time performance.

2.1.3 Sound Propagation in game-engine

2.2 Sound Source Localization

Sound Source Localization (SSL) is the process of determining the position of a sound source. It usually uses a microphone array that captures the sound signals from multiple directions. SSL is used in various applications, such as speech recognition, robot navigation, surveillance, and security. In this thesis, SSL is used to estimate the distance and direction of a sound source to detect excessively noisy vehicles.

2.2.1 Spectrograms for sound visualization

Spectrograms are a visual representation of the frequency content of a sound signal. They are often used in sound source localization to identify the direction of a sound source. The spectrogram is a two-dimensional representation of the frequency content of a sound signal (figure 2.1).

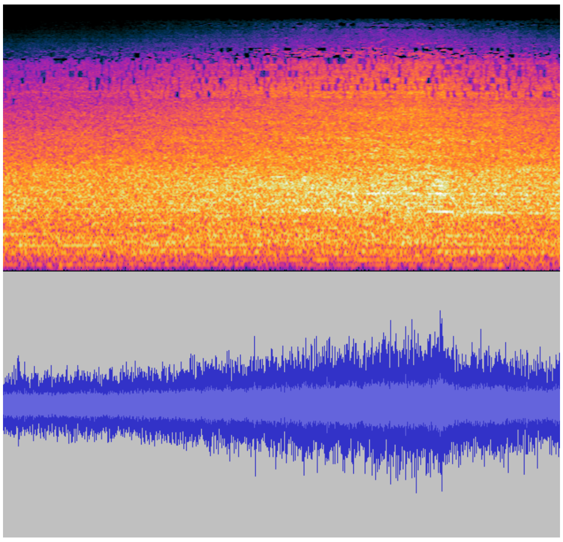

Figure 2.1: Spectrogram of a sound signal

The x-axis represents time, and the y-axis represents frequency. The intensity of the color at each point in the spectrogram represents the amplitude of the frequency component. A matrix of spectrograms allows the representation of multiple channels, such as the ones recorded by a microphone array. On that matrix, each spectrogram represents the frequency content of a single channel (figure 2.2).

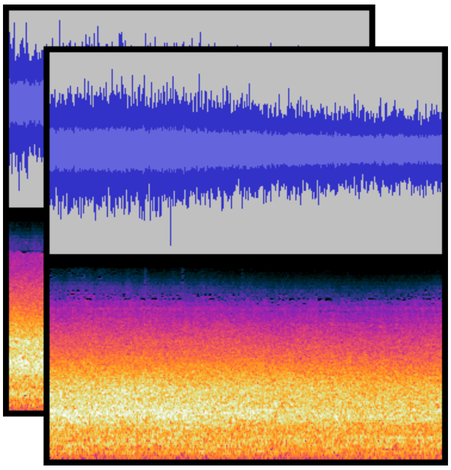

Figure 2.2: Dual channel spectrogram matrix of a sound signal

Looking at the frequency content of the sound signal allows us to identify the time delta of a recorded sound by using a multi-channel spectrogram. The bright spot on the spectrogram will indicate a jump in the amplitude and determine the start time of the record of a loud sound. By comparing this time with the other channel, we can find the direction of the sound source by comparing the sound signal's time delta with the other channels' time delta (figure 2.3).

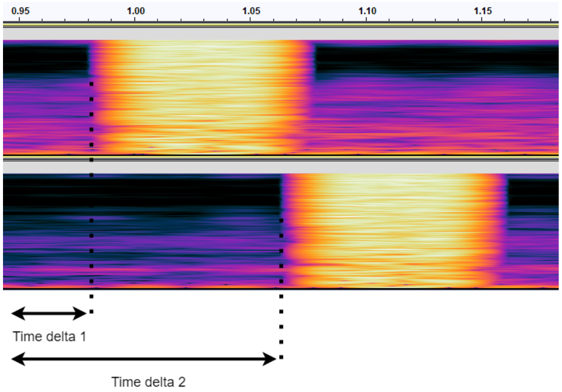

Figure 2.3: Spectrogram of two sound signals with time their delta

Since we know the distance between the microphones, we can determine the direction of the sound.

2.2.2 Origin of sound using two microphones

Admitting the following setup (figure 2.4), if the time delta 1 is greater than the time delta 2 of the other channels (setup 1), the sound source is closer to microphone 2. If the time delta 1 equals the time delta 2 (setup 3), the sound source is at the same distance to both microphones. If the time delta 2 is greater than the time delta 1 (setup 2), the sound source is closer to the microphone 1.

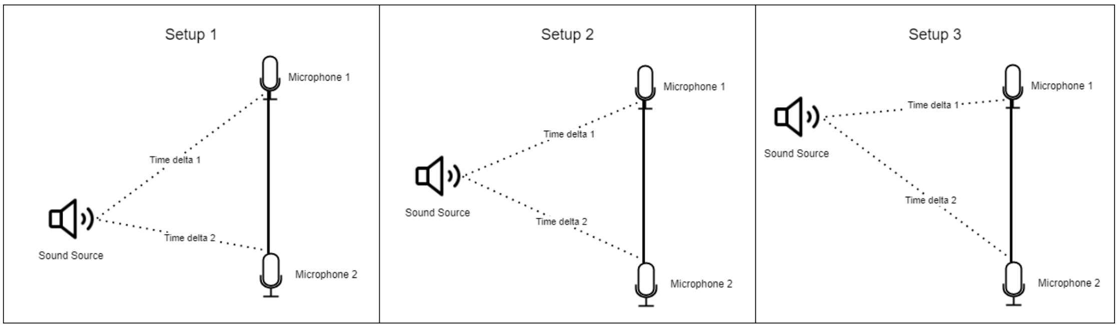

Figure 2.4: Sound source localization setup

This concept can be formalized and is better explained in [2]. Once the delay between the two microphones is known, the equation allows us to find the direction of the sound source by using trigonometric calculations. As in the figure 2.5, considering point M as the sound source and point A and B as microphones, the distance between the two microphones is d and the time delta between the two microphones is Δt, the angle α can be calculated.

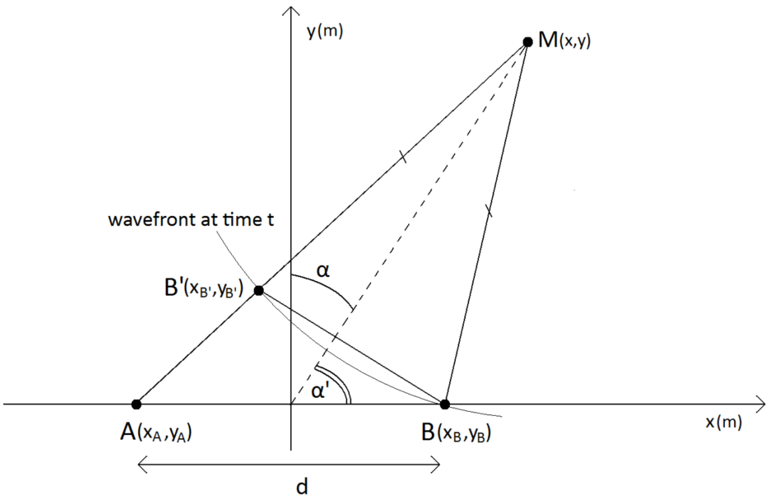

Figure 2.5: Equation formalization. Original image from [2]

By looking at the graphic, the following equation can be found:

$$AB' = AM - B'M \tag{2.1}$$

With Pythagorean theorem:

$$AM = \sqrt{(X_a - X)^2 + (Y_a - Y)^2} \tag{2.2}$$

$$BM = \sqrt{(X_b - X)^2 + (Y_b - Y)^2} \tag{2.3}$$

The two microphones have the same Y coordinate, so $Y_a = Y_b = Y$ and $Y_a - Y_b = 0$ and $X_a = -X_B$ The equation becomes:

$$y = \pm\sqrt{\frac{AB'^2}{4} - x_B^2 + x^2(\frac{4 \cdot x_B^2}{AB'^2} - 1)} \tag{2.4}$$

This setup shows that two microphones are enough to determine the direction of a sound source.

2.3 Neural networks

Neural networks are machine learning algorithms based on biological neurons used to solve various problems, including image recognition, speech recognition, and natural language processing. Neural networks learn from provided data to solve a problem without explicitly programming the solution. Many domains, like self-driving cars, facial recognition, and medical imaging, achieve better results using neural network models.

A neural network (figure 2.6) is composed of multiple neurons (the circles) that are organized in layers and connected to the neurons in the previous and next layers.

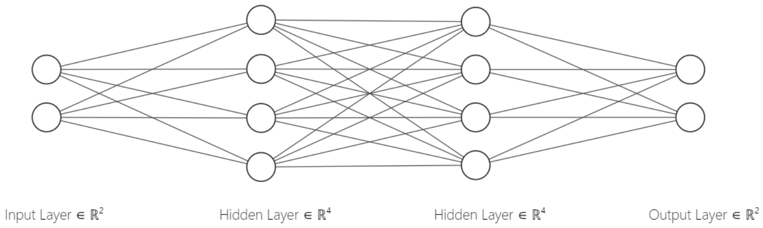

Figure 2.6: Neural network

Deep neural networks are a type of neural network composed of multiple layers of neurons. They are trained on a large dataset of images and then used to classify new images. There are countless architectures [1] and implementations of neural networks, but they all share the same basic principles.

2.3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are neural networks used for image recognition. They use convolutional layers to extract features from images. These features are then fed into fully connected layers to perform classification.

2.3.2 Convolutional Neural Networks for source localization

CNNs are mainly used to classify images but can also classify sounds. They are used in this thesis to classify the spectrograms of the sound signals recorded by the microphone array. The spectrograms are converted into images and fed into the CNN. The CNN then outputs a probability distribution over the possible classes. The output of the CNN gives the probability of the sound source coming from a specific direction.

2.4 Related Work

2.4.1

3 Conception

3.1 Sound Propagation Simulation

3.2 Dataset Creation

3.2.1 Embedded system setup

3.2.2 Data recording and storage

3.3 Neural Network for Sound Source Localization

3.4 Adversarial Attack conception

4 Setup

4.1 Simulation model creation

4.1.1 Microsoft Project plugin

4.1.2 Unity plugin

4.2 Dataset creation

4.2.1 Microphone installation

4.2.2 Embedded system setup and access

4.2.3 Data transmission

4.2.4 Data recording and storage

4.3 Neural Network for Sound Source Localization

4.3.1 Dataset preparation

4.3.2 Neural Network architecture

4.3.3 Training

4.4 Adversarial Attack

4.4.1 Fast Gradient Signed Method implementation

5
Results

5.1 Sound Propagation Simulation

5.1.1 Dataset Augmentation

5.2 Neural Network model results

5.3 Adversarial Attack results

5.3.1 Adversarial Attack mitigation

6

Conclusions and Future Work

6.1 Conclusions

6.1.1 Specification Fulfillment

6.2 Future Work

6.2.1 Sound Propagation Simulation

6.2.2 Sound Propagation Simulation bachelor's thesis

A thesis named *SimSound3D is*

6.2.3 Advanced adversarial attack

6.2.3.1 Patch attack

6.2.3.2 Targeted attack

6.2.4 Dataset publication

6.2.4.1 Dataset annotation

6.2.5 Loxo Ears model

A
Appendix

A.1

List of Tables

List of Figures

Bibliography

[1] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E. Alsaadi. A survey of deep neural network architectures and their applications. Neurocomputing*, 234:11–26, 2017.*

[2] Carlos Fernández Scola and María Dolores Bolaños Ortega. Direction of arrival estimation : A two microphones approach. 2010.