FACULTY OF ENGINEERING

Sound Classification System

Final year project

EERI 474

by:

Christoff Smit 25881469

North West University - Potchefstroom Campus

Supervisor: Mr. Andreas Alberts (BEng)

November 8, 2020

Abstract

Executive Summary

Table of Contents

List of Figures

List of Tables

List of Terms

sound localization - estimating the position of a sound source relative to the microphone (or an array of microphones) by measuring the direction of and distance to the sound.

point of origin - epicentral position of the sound source (relative to the mic array)

List of Acronyms

MFCC's - Mel Frequency Ceptral Coefficients
CV - Computer Vision
MIR - Music Information Retrieval

1 Introduction

1.1 Problem Statement

Develop a sound classifier (to be used in an existing security system) which is capable of accurately identifying the sources of certain sounds.

1.2 Background

This project entails the classification of sounds (in terms of their respective sources). The product is meant to be mounted on top of an existing security system developed by another engineer, Mr. XX.

The system is currently capable of performing sound localization using a 3-piece microphone array (arranged in triangular formation), driven by software written in the Julia programming language. The author now has to introduce classification capabilities to the system.

1.3 Proposed Solution

The author recommends achieving the project goal by applying machine learning algorithms to the frequency content of specified sound samples.

Frequency spectral analysis should be performed on these samples and the results used to build sample libraries with which to train, and test, the classifier.

1.4 Project Objectives

1.4.1 Primary Objective

The development of highly-accurate sound classification capabilities to add to the existing system.

1.4.2 Secondary Objective(s)

- Continue adding to the library of sounds that can be uniquely distinguished by the system. Increase the size of the general population for the training- and test data sets.
- Label identified sound sources as being either friendly, neutral or a threat.
- Deploy the classifier as an intuitive security surveillance system by implementing reactive security features such as: setting off alarms, contacting authorities, etc.
- Perform tests on various neural network algorithms, choose the most applicable/realistic option and possibly improve upon it.

1.5 Project Scope

1.5.1 Deliverables

The software required for the system to perform sound classification based on audio captured by the microphone array.

The system should be able to distinguishing between the following sounds:

- voices
- footsteps
- dogs barking
- gunshots
- shattering glass
- a motor engine
- alarms
- whistles

1.5.2 Cost

The project does not involve significant financial cost as it largely entails the development of software by the author.

However, possible expenses at a later stage might include:

- Purchasing sound samples from online databases with which to train the classifier.
- Purchasing fuel to pick up the existing system from Mr. XX to apply the classification software and test its accuracy.

Note that the project budget is set at R2500.

1.5.3 Risk Analysis

Technical Risks:

- ssssssssssss
- ssssssssssss

Project Risks:

- ssssssssssss
- ssssssssssss

1.5.4 Resources

1.5.5 Feasibility

Although most modern security systems rely heavily on both visual *and* audio monitoring for threat detection, the latter carries more weight in conditions where video surveillance proves a challenging task to accomplish due to obstructed view, bad lighting, etc.

Audio surveillance is admittedly bound to its own set of limitations, but the equipment used is usually of a more robust nature and less expensive than that of its visual counterpart.

The classification of sounds (in addition to locating their points of origin) is a very attractive feature to introduce to the system. It would enable the recognition of a sound source as a possible threat, and could perform security actions accordingly, e.g. set off an alarm, contact authorities, etc.

1.5.6 Testability

The testability of this system is ensured by applying proper machine learning principles in dividing the available database of sound samples into training- and test sets of appropriate size.

1.5.7 Limitations

The following limitations are imposed on the classifier:

- The largest possible limiting factor on the classifier is the amount of training provided to the neural network responsible for classification. If the system is not sufficiently trained, it can't be expected to accurately classify sounds.
- Another limitation is the quality of microphones being used to monitor the vicinity. Poor microphones would lead to worse classification accuracy and a smaller possible area that can be successfully monitored.
- Noise (e.g. weather, traffic, animals) also imposes limitations on the classification accuracy as it distorts the sound waves to be measured.

1.5.8 Safety

The nature of this product is such that it does not inflict any physical harm to its surroundings and therefore does not pose safety threats to the user.

1.6 Project Planning

1.6.1 Schedule

Figure 1.1 shows the time-schedule for the project in the form of a Gantt chart.

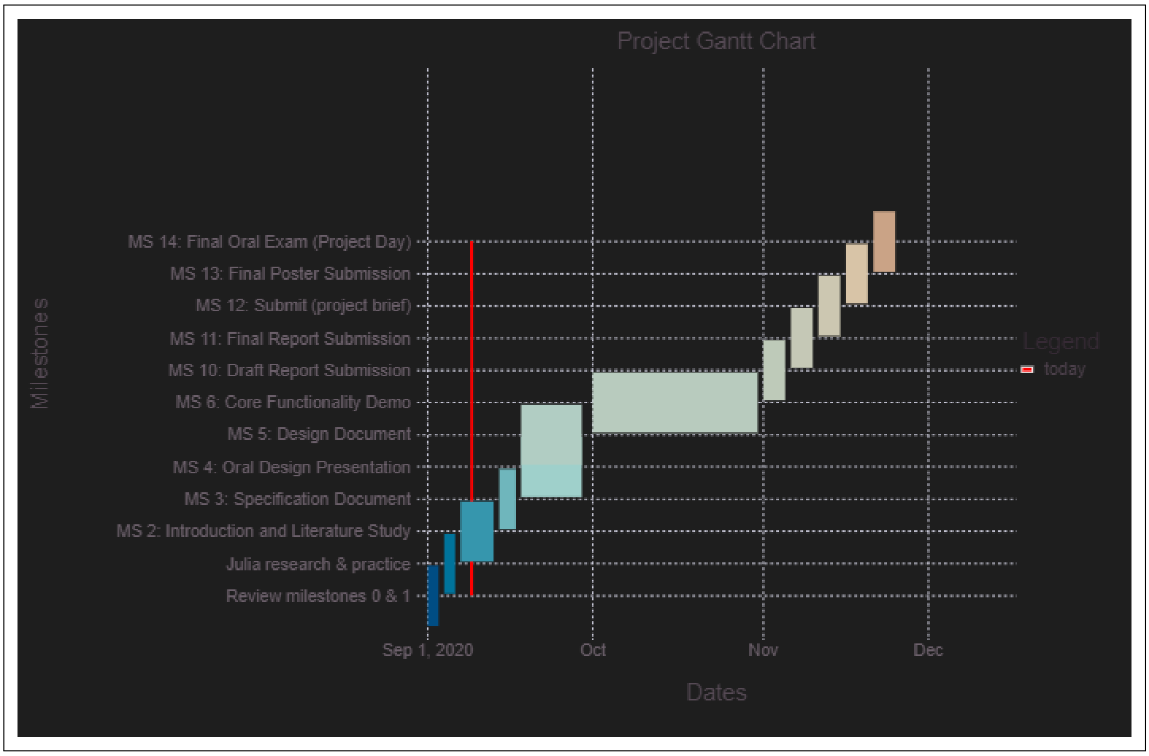
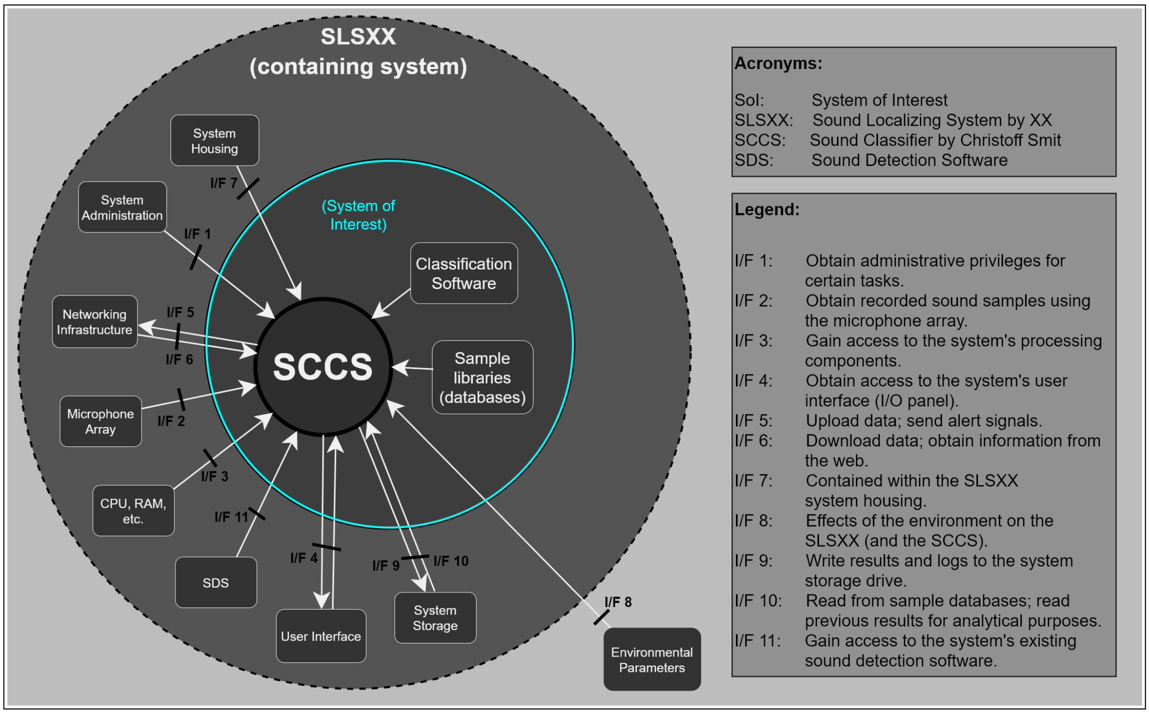


Figure 1.1: Project Gantt chart

1.7 Preliminary Conceptual Design

Figure 1.2 shows the preliminary context diagram for the project.



Acronyms:

SoI:	System of Interest
SLSXX:	Sound Localizing System by XX
SCCS:	Sound Classifier by Christoff Smit
SDS:	Sound Detection Software

Legend:

I/F 1:	Obtain administrative privileges for certain tasks.
I/F 2:	Obtain recorded sound samples using the microphone array.
I/F 3:	Gain access to the system's processing components.
I/F 4:	Obtain access to the system's user interface (I/O panel).
I/F 5:	Upload data; send alert signals.
I/F 6:	Download data; obtain information from the web.
I/F 7:	Contained within the SLSXX system housing.
I/F 8:	Effects of the environment on the SLSXX (and the SCCS).
I/F 9:	Write results and logs to the system storage drive.
I/F 10:	Read from sample databases; read previous results for analytical purposes.
I/F 11:	Gain access to the system's existing sound detection software.

Figure 1.2: System Context Diagram (preliminary)

Figure 1.3 shows the preliminary physical architecture of the system.

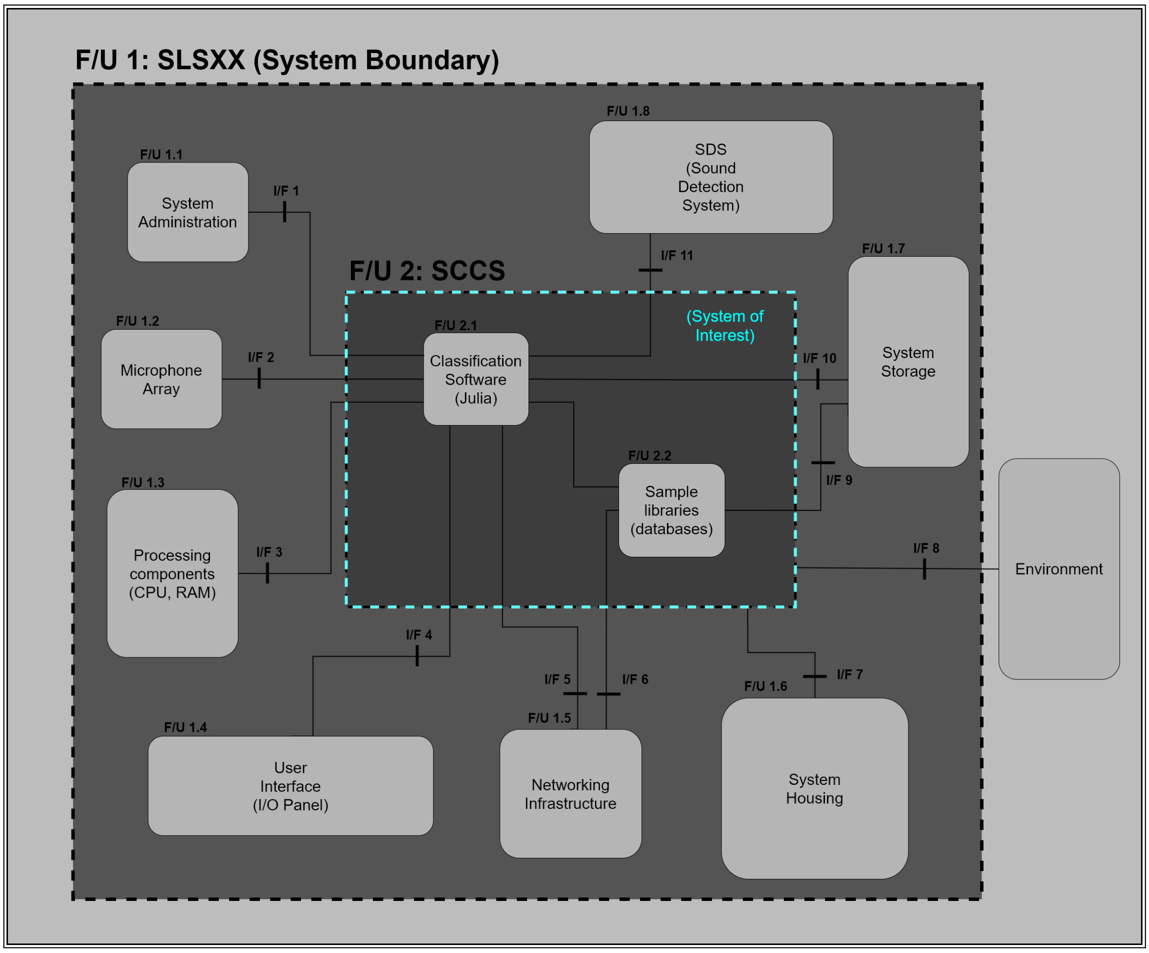


Figure 1.3: System Physical Architecture (preliminary)

Figure 1.4 shows the complete life cycle of the system.

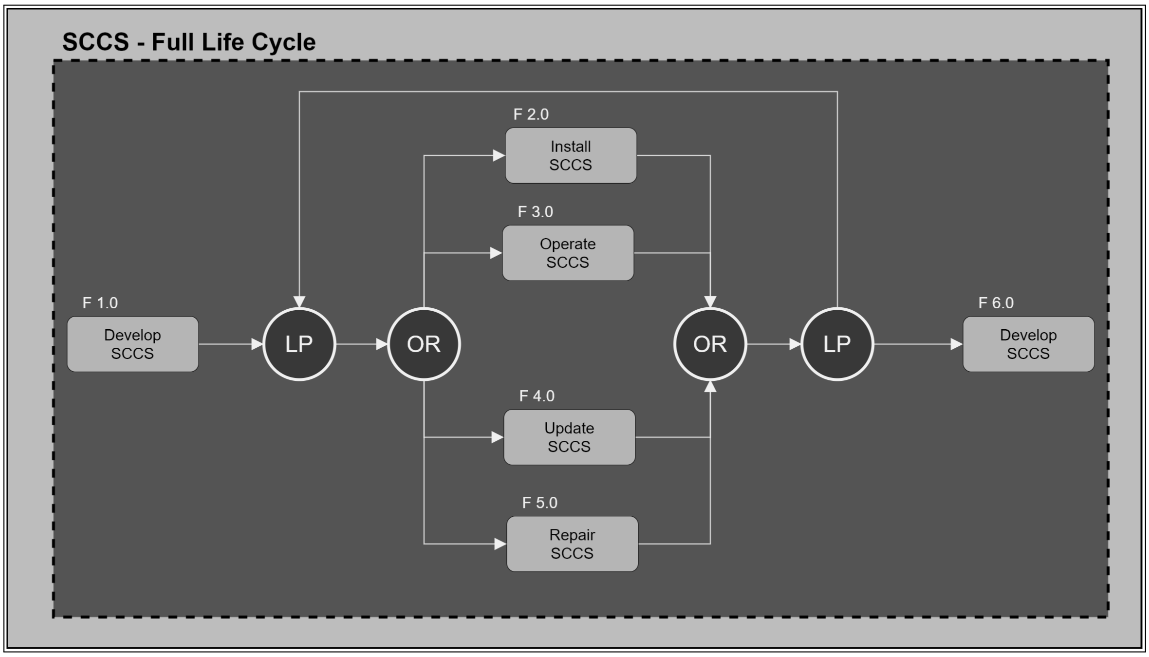


Figure 1.4: Full System Life-Cycle

Figure 1.5 illustrates the *Operate* function in Figure 1.4

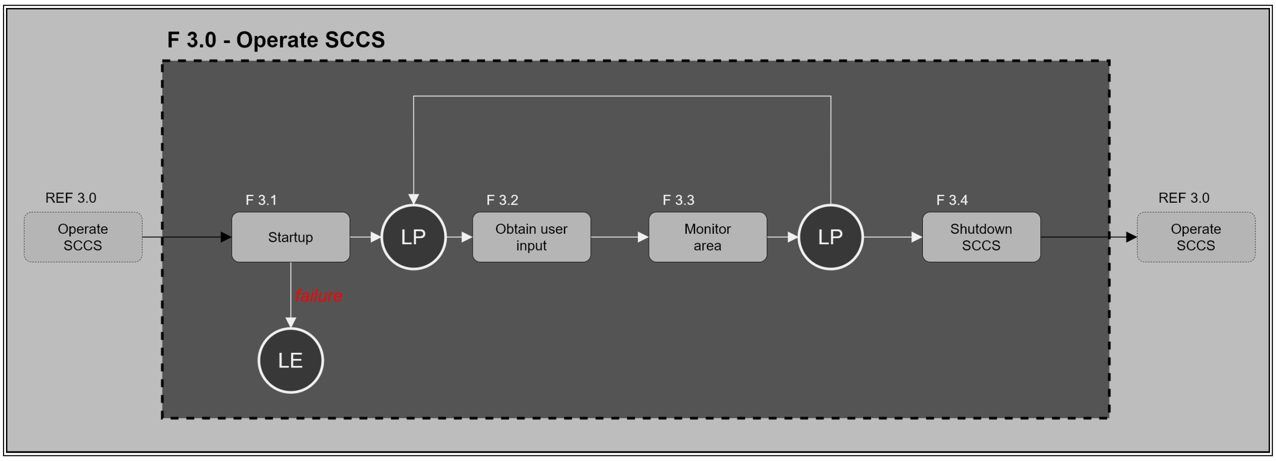


Figure 1.5: Function 3.0 - Operate SCCS

2 Literature Study

This chapter entails an in-depth literature study of the project scope. A large part of the chapter is devoted to a technology survey which provides background on the functional units (main project components) identified in Figure 1.3.

2.1 Introduction

The detection and interpretation of sounds have been a part of humans' daily lives for thousands of years. Automatic detection and analysis of certain sounds may therefore be an appealing ability to introduce to a machine.

Sound may be recorded in a given number of channels, using a microphone with a specified bit-depth. The bit-depth indicates the amount of frequencies which may be uniquely identified by the sensor (microphone).

2.2 General process (for audio machine learning)

Figure 2.1 shows the WAV file recording of a dog barking 5 times over a period of 4 seconds.

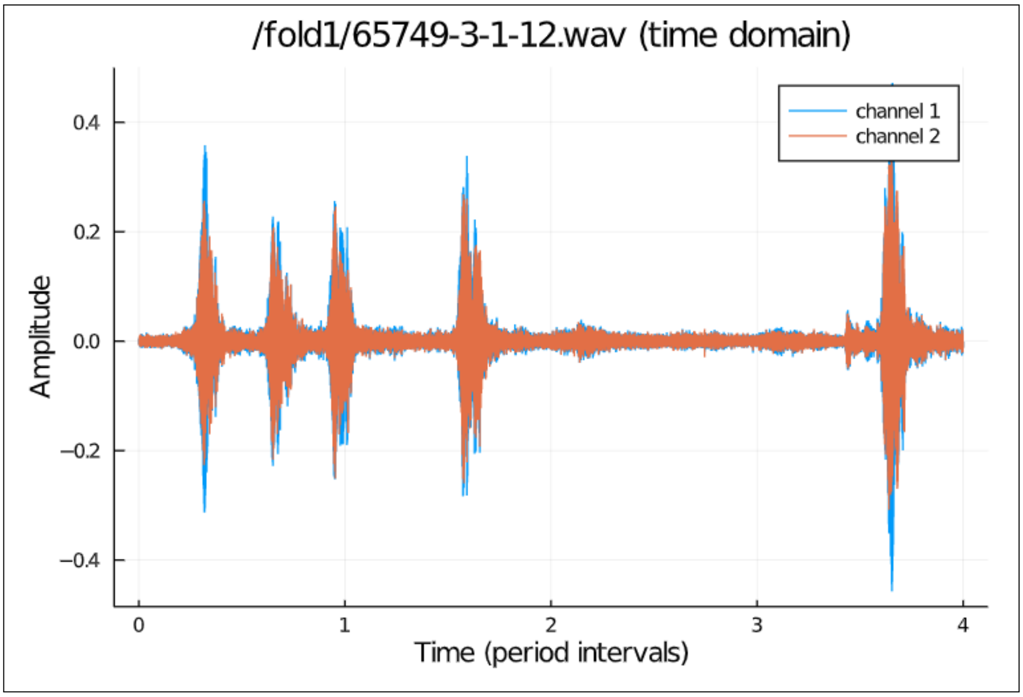


Figure 2.1: 2-channel audio recording of a dog barking

The (Fast) Fourier transform is determined for the entire sample, providing an instantaneous ("snap-shot") illustration of the recording's total frequency content:

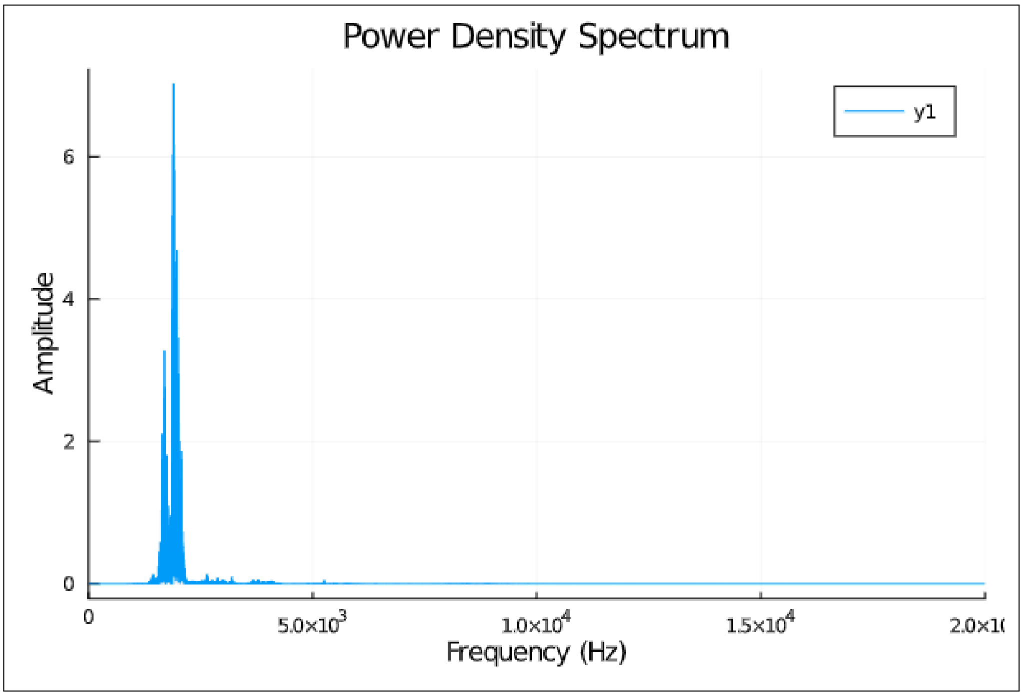


Figure 2.2: Periodogram (FFT) of the audio recording

It becomes apparent that the dog is mainly barking at a specific frequency around 2 kHz.

Note that only frequencies up to 20 kHz are considered. This is chosen in accordance with the average human hearing range, but also with the Nyquist frequency of the microphone in mind. The microphones used maintain a sampling frequency of 44.1 kHz. Therefore, according to Nyquist's theorem, the highest frequency that each (of the three) microphones may accurately sample is $44.1/2 = 22.05$ kHz.

However, calculating the Fourier transform of the entire sound sample neglects the element of time, which is a very important characteristic in audio recording.

Consequently, the STFT (short-time Fourier transform) is calculated by determining the above-mentioned FFT at a few successive instances in time to produce the spectrogram in Figure 2.3.

Finally, the (logarithmic) Mel filterbank is applied to the signal, producing the two-dimensional array of MFCC's in Figure 2.4 that serves as input to the neural network. Note that 13 coefficients are considered here with time windows of 25 ms and a step size of 10 ms.

2.3 Background

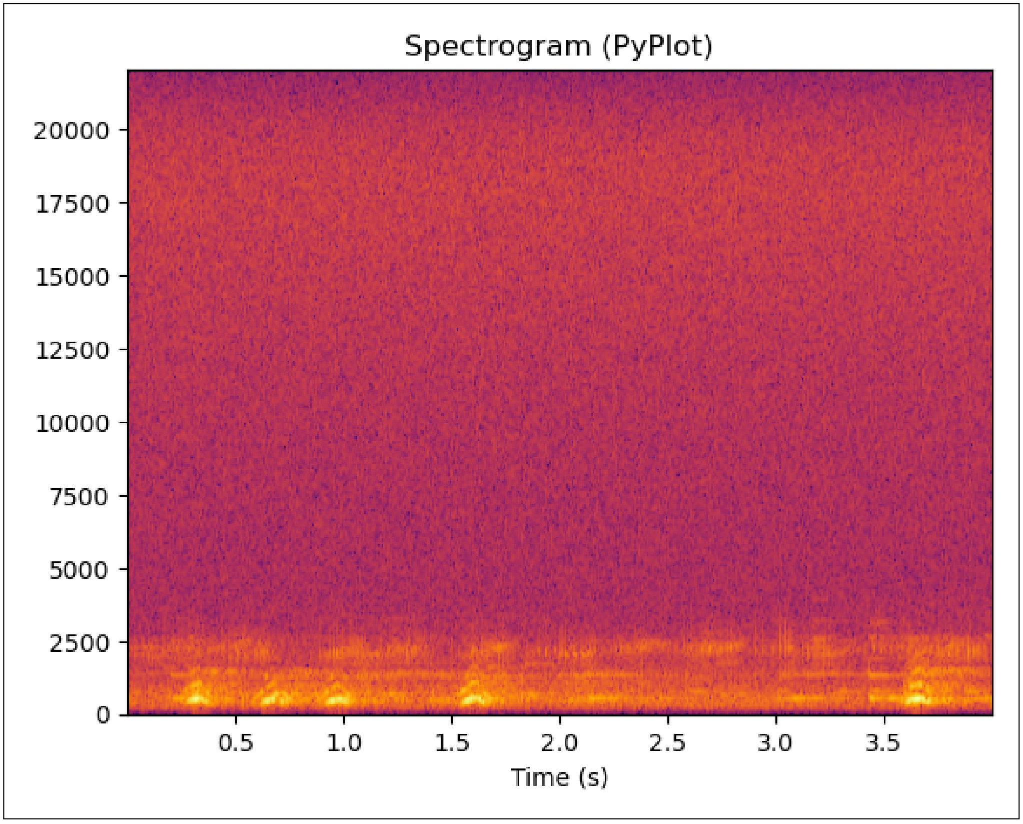


Figure 2.3: STFT (NFFT=512, noverlap=128)

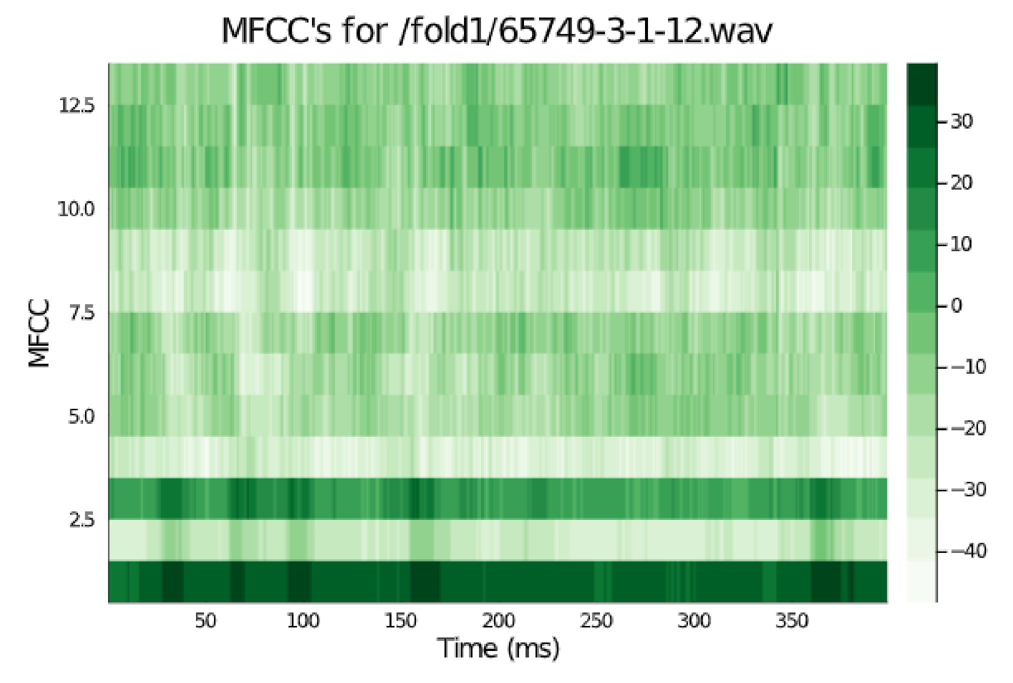


Figure 2.4: MFCC'c heatmap (N=13)

2.4 Technology Survey

2.4.1 Trade-Off Studies

3 System Design

4 Detail Design and Implementation

5 Testing and Evaluation

6 Conclusion & Recommendations

Christoff Smit - November 8, 2020

Appendices