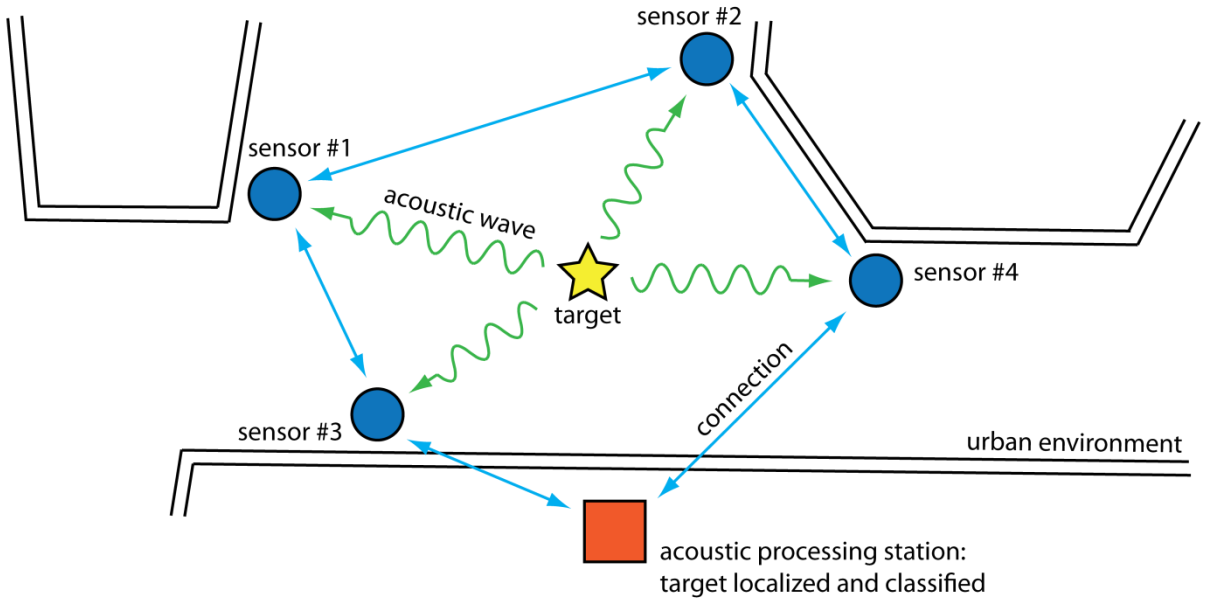
thesis

Localization and **Classification** using an **Acoustic Sensor Network**

experimental data processing
for urban acoustic surveillance



T.H. de Groot
2010

T.H. de Groot
Thales Nederland B.V.
Technische Universiteit Delft

Localization and **Classification** using an **Acoustic Sensor Network**

experimental data processing
for urban acoustic surveillance

Thesis

to obtain a degree

of **Master of Science**

in **Electrical Engineering**

at **Delft University of Technology**

Author: T.H. de Groot
Student number: 1269003
Division: MTS-Radar
Defence date: 1 July 2010

Committee members:
Prof. dr. A. Yarovoy
Dr. ir. G.J.M. Janssen
Dr. T.G. Savelyev
ir. E. Woudenberg

Summary

The Acoustic Sensor Network (ASN) has emerged as an important research area, because acoustic sensors can significantly increase situational awareness in many situations. Although little is currently known about acoustic surveillance, Thales Nederland is interested in the potential offered by an ASN in urban environments and specifically for classification. This is because the operational problem is not merely to detect targets, but also to localize and classify them in a robust way. Current radar implementations do not provide enough performance to classify targets in complex urban environments while now acoustic sensors are seen as an extra source of information. Therefore, a challenging project was initiated to investigate the potential and feasibility of a passive ASN to localize and classify targets.

Three different kind of targets were investigated for acoustic surveillance: guns (muzzle blast), vehicles (running piston engine) and humans (walking pedestrian). Can a passive ASN be deployed in urban environments to localize and classify them only by their emitted sound? It is a great challenge to cope with the received signals using a passive ASN in urban environment, because signals can - even within the same classes - differ significantly. This leads to a technical challenge when it comes to achieving robust localization and correct classification. Two project objectives were defined. Firstly, extract target information and use propagation models to localize the targets. Secondly, extract features which allow a classification method to discriminate between the different target classes. Experimental data processing had to be designed, implemented and evaluated with measured data for a performance indication.

To find target features and to investigate localization possibilities, extensive acoustic analysis is done on the three targets. The emitted energy of the gun was the dominant feature of the muzzle blast. The dominant features of the running piston engine were the harmonics. The walking pedestrian had characteristic time interval features between the footsteps.
An experimental framework was designed with a signal processor, localizer and classifier. The signal processor has to process the recorded signal in such a way that the localizer and classifier can use the result. The localizer is designed which can localize targets time-based and power-based. The feature extraction of the classifier provided discriminative features which allowed a classification method to discriminate between the classes.
The designed components are combined and experimentally implemented (proof of concept) with four microphones and tested for a system performance indication. The time-based localization performed well, but the power-based localization requires extensive calibration to perform proper. The dominant target features were extracted and allowed an experimental classification tree to discriminate between the classes.

Passive acoustic surveillance is possible, but the system performance depends very much on the operational situation (e.g. background noise). The performance mainly depends on the signal to noise ratio (SNR) and the SNR depends on the target class. The potential for localizing and classifying walking pedestrians is very low. Vehicle power-based localization and detection has potential, but good microphone hardware is required. Gun localization and classification has the highest potential and feasibility. Although there are some difficulties, throughout this project it became clear that acoustic sensors are able to provide extra information and features. This can be used to further increase the robustness and integrity of urban surveillance systems.

Acknowledgements

I want to thank God for being such a good Father to me. I want to thank Him for his love, generosity and forgiveness. More to the point, without God I would not have been able to finish this graduation project as I did. I want to thank Him for receiving the intelligence and skills which I could use in this project. Furthermore, I want to thank Him for giving me such a good acoustic surveillance system example in the form of the human.

Thales Nederland has provided me with a good working environment in which to do my graduation project. I also want to thank the organization for the (financial) scholarship. I want to thank Evert Woudenberg for being such a good mentor. Ronny Harmanny was very precise and critical, and I want to thank him for that attention to detail. I thank Joris van de Meerakker for his interest in the project and for participation in the discussions. Furthermore, I want to thank everyone who was interested in my project.

Great thanks goes to Delft University of Technology for all the education I have received here. With all the knowledge gathered I was able to do this master's graduation project. I want to thank my supervisor Alexander Yarovoy for his broad academic view. And finally I thank the committee members for reading this thesis and for asking me difficult questions and for being critical at my graduation.

I thank SV Doel Treffend in Delft for their hospitality and for permitting me to do some indoor gunshot measurements. JST-Waalsdorp in The Hague allowed me to record some shotgun sounds outside and so I would also like to thank them for their time and their hospitality.

Preface

Engineering is one of the greatest things there is. Using your intelligence and knowledge to arrive at a good solution is really challenging and exciting. I am glad that I was able to do a scientific academic graduation project, for myself, but also for everyone who is interested in my research.

Although I had an idea of what it would be like to do a graduation project, I was still at certain moments surprised. For example, I knew in advance that defining my project would be a difficult process, but I did not expect it would be so exhausting as it was. However, it was not a negative experience: I learned that defining a project is almost the same as learning. I first needed field knowledge and to do a literature survey before I could correctly set the boundaries of the project.

One of the crucial parts of any scientific master's graduation project is the construction of a scientific thesis. The road to a final thesis document, which is a pleasure to read for all who are interested, is very challenging. For instance, everyone has another opinion about the structure and content of the document. Although it is interesting and challenging to use the different feedback, I have managed to successfully complete this thesis.

I hope that every reader will read this thesis with pleasure and is able to learn about acoustic surveillance and can use this knowledge for his/her own project and/or product.

Contents

1 Introduction

Thales Nederland is interested in the potential offered by an Acoustic Sensor Network (ASN) for acoustic surveillance in urban environments. The current operational problem in complex urban environment is not merely to detect targets, but also to localize and classify them. The current radars do not provided enough performance to do this and now acoustic sensors are seen as an alternative way to significantly increase the situational awareness and especially the classification. Acoustic is very intuitive and can probably provide extra features in these complex urban environments. The Thales department in Delft, therefore, started a student project to investigate the potential and feasibility of an ASN. One of the first parts of the acoustic surveillance project was a graduation project which would investigate data processing for localization and classification by using an ASN.

The project started with a literature survey to investigate the current knowledge and solutions concerning acoustic surveillance. After that, an assumed operational situation was defined and three targets were chosen: guns (muzzle blast), vehicles (running piston engine) and humans (walking pedestrian). For the graduation project two objectives were defined. Firstly, provide target information with propagation models which allow a least square solver to estimate the target position. Secondly, provide target features which allow a classification method to estimate the target class. To achieve these objectives is a technical challenge, because the received signals at the passive acoustic sensors are very unpredictable in urban environments. An Experimental ASN (EASN), which could localize and classify target, has to be designed and implemented (proof of concept). The EASN should give an indication of the potential and feasibility of acoustic surveillance.

This chapter will discuss the results of the literature survey and it will explain the project. Sections 1.1 and 1.2 discuss the current status of ASNs. The current knowledge on acoustics (sound) is investigated in Sections 1.3 and 1.4. The localization and classification principles are outlined in Sections 1.5 and 1.6 respectively. Section 1.7 will outline the project objectives and the approach.

In papers and books various terms and words are used differently and are sometimes interchanged. This thesis aims to achieve good and clear separation to avoid such lack of clarity. For a correct understanding of the terms used in this thesis, the Glossary, Abbreviations and/or Symbols can consulted. The appendices contain useful information, for example on, experimental hardware analysis, wavelet analysis and classification methods.

1.1 Background

There is always a demand for better situational awareness in complex environments for medical, industrial, scientific, military, security and consumer purposes. For the operations, detection is not enough anymore, but also localization and classification is required for situation estimation. Correct situation information is a prerequisite to good decision making. The environments where extra situational awareness is demanded can vary from an empty deserts to extremely complex urban situations.

The applications can cover a large geographical area and require quick, accurate and reliable information. The fields of security (civil defence) and rescue services (police, fire brigade, ambulance) require good situational awareness in situations such as calamities, disasters and evacuation operations. Due to recent world events, which raise the fear of terrorist attacks, people are also constantly endeavouring to improve security. The applications are endless, but the point is that there is a high demand for efficient information extraction techniques.

A Sensor Network (SN) can help to extract information from a certain area. In a SN, area information is gathered by cooperative sensors, which are placed on location to provide the required data. Multiple sensors can provide significant advantages (such as accuracy) over single sensor systems. Because of the great potential, a lot of research is currently being done into SNs.
Typical SN implementations consist of a large number of cameras distributed in a certain area and connected to a central point. Thales Nederland investigates an assembly of different sensors, such as radar, video and sound, which are all connected wirelessly. This type of system is also known as a Wireless Multimedia Sensor Network (WMSN). Acoustic sensors can be a crucial part in such a WMSN, because many characteristics of an object can be inferred from the sound it generates. For example, acoustic sensors can help to aim the video sensors in a WMSN.

The Acoustic Sensor Network (ASN) has emerged as an important research area, because in many situations acoustic sensors can significantly increase situational awareness. Acoustics is also very intuitive for the human and can probably provide extra features of targets in urban environments. Although many aspects of an ASN have been studied and realized, there are still major technical challenges ahead that must be resolved before an ASN can be used in complex urban environments.
A critical aspect of most sensing systems is data processing. Much research is being done into ASN algorithms and different designs are suggested. However, it is still unclear whether localization and classification together in an ASN can be robust and reliable. Furthermore, can such an ASN be deployed in urban environments to localize and classify guns, vehicles and humans?

1.2 Acoustic Sensor Networks

First of all, this thesis will focus on sound which is an acoustic signal propagating through the air (gas). Thus the focus is on microphones, and not on seismic or underwater sensors. In some research areas, vibration is defined as an acoustic signal which propagates though the ground (solid). However acoustic signals can also travel through other material as well, like water (liquid) and in reality, microphones will unfortunately also measure a certain amount of vibration. This project will investigate acoustic sound recorded with microphones. Sound absorption in air is significantly less than vibration absorption in the ground [1] which allows the system to detect targets at a greater distance.

The ASN is becoming increasingly popular, because of its extended coverage and sensing capabilities. Characteristics can be extracted from sound signals and mostly sound cannot be easily damped or blocked, thus there are no illumination problems. There are many indoor and outdoor situations where an ASN can be used:

- Urban monitoring (individuals, vehicle traffic)
- Battlefield surveillance (sniper gunshot, vehicles)
- Boundary/perimeter protection (intruders, illegal hunting)
- Police and forensic evaluation (crime, homicide)
- Home surveillance (intruder security, baby/children nursing system)
- Medical disease detection (hospital, the elderly at home, disease spread)
- Census of native animal populations (habit monitoring)

Acoustic systems have a limited sensing range, due to propagation losses, and can be "fooled" by artificial sounds. The use of acoustic sensors in mission-critical applications is therefore limited. However, when acoustic sensors are combined with other kinds of sensors, like in a WMSN, the reliability of the total system can be increased.

1.2.1 Current implementations

ASNs come in many forms [2], [3] but there is generally a similarity: the systems are usually used for calamity detection and must only report unusual situations. Another typical thing is that the systems usually have to detect loud sounds and this is probably because of its feasibility. Besides the technical feasibility, the social importance plays a dominant role in acoustic research.
For acoustic detection, the most common target is the gunshot. Furthermore, there are also real implementations for gunshot detection and localization, such as the *Ears system* from *QinetiQ North America*, the *Boomerang* from *DARPA* and *BBN Technologies* or the *Gunshot Location System* from *ShotSpotter*. These systems are predominantly used by the army and government for sniper localization and to further increase situational awareness. Some can also detect explosions, whether this is due to the fact that the system cannot discriminate between the two or whether it is because an extra class has been added is unclear. The systems can be added to the soldier's equipment, mounted on vehicles or deployed on buildings or structures.
For detection of other targets, like for example the human cough, which can cause diseases, there are some small scale implementations, but they are mainly used for research purposes [4]. Although much research has been done into speech recognition, urban acoustic surveillance is a relatively new field of study which is why only infrequently used systems exist.

1.2.2 Thales' view

The Thales ASN design theory is to deploy many nodes, which all have a single passive acoustic sound sensor. In other words, each node has only one microphone as its sensing device. In this way, sensor nodes can be small and inexpensive. The passive sensor will be acoustically undetectable and "hidden" deployment is possible. Such nodes can be flexibly deployed in a random manner and in the future they may be deployed in hostile regions through air drops.

The advantage of having many sensors over having a few sensors are these: accuracy increase, less distance dependent and fault tolerant. Every sensor node has a certain sensing range. Since, in practice, the targets are moving in an area it is impossible to create a single "super" sensor that is always close to the target, which will severely impact the system performance. Deploying many sensors over a certain area will result in a probability increase that a sensor is close to its target.

In fact, it could be argued that many sensors distributed over an area could be seen as a sensor array with huge spatial separation. Multiple sensors at one node will also result in extra complexity, such as careful deployment and extra data processing. Thus, the huge spatial separation of many single sensors provides significant advantages over many sensors at one location.

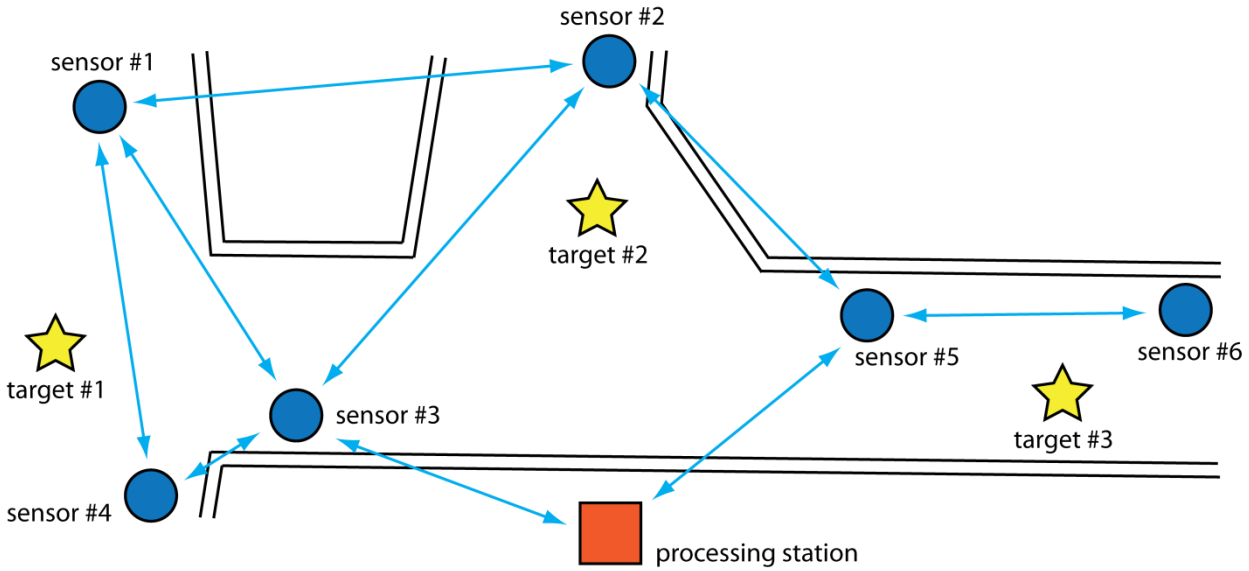


Figure 1-1: Illustration of ASN in urban environment

An illustration of a wireless ASN system is given in Figure 1-1. Thales Nederland wants that with such an ASN extra target features can be extracted. With these acoustic features, the situation can be better estimated and the situational awareness can be further increased.

1.3 Acoustic sound in urban environment

Sound is a well studied [5], [6], [7] topic and so this chapter will briefly discuss sound in urban environments, because it is crucial to investigate and predict how sound will propagate.

Acoustic sound is a longitudinal wave, and the main properties which are present are propagation loss and propagation time. Put simply, it demands power and time for the wave to propagate from one point to another.
There are many other phenomena which result in signal change. A common situation is when an acoustic wave is travelling from one medium to the other. This results in reflection and transmission. Other dominant phenomena are: refraction, diffraction, scattering, interference and the Doppler effect. These effects mainly depend on the frequency and the medium, and therefore, on the sound spectrum and position of the target.
The urban environment can be very complex and it can have many situations: shadow regions, multipaths, multiple targets, environmental noise, etc. To stress the complexity, other possible effects are: wind (turbulence), precipitation (rain, hail, snow), non-uniform temperature and humidity, animals (insects, birds), mankind (barriers like buildings, sound-producing objects).

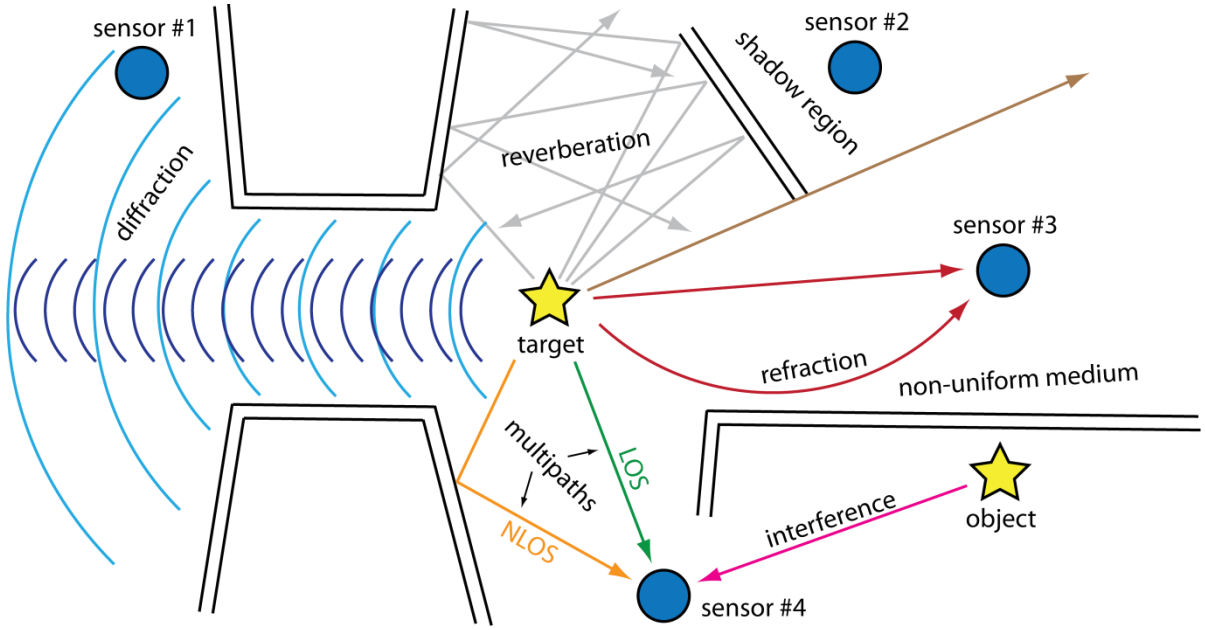


Figure 1-2: Acoustic complexity in urban environment

To show the urban sound complexity an illustration of some effects is given in Figure 1-2. With all the mentioned environment effects, it is hard to imagine that an ASN deployed in urban environment can be robust. Furthermore, it shows the current limits and/or major challenges for acoustic sensing in operational use. Shadow region, multipath and interference are simple principles, but can be seen as the most dominant and they can be very problematic.

1.4 Acoustic sound emitted by targets

In urban environment, multiple objects of arbitrary types will transmit certain sounds, but this thesis focuses on three types of targets: guns, vehicles and humans. Although the word 'target' may suggest that it needs to be eliminated, this is not necessary the case. In this thesis they are called 'target', because the system objective is to localize and classify them. Each of these targets has been studied in previous research and produces its own specific kinds of sounds. Although within the same class the target signals could differ significantly from each other, there are sometimes some common factors. The challenges of this thesis are firstly, to extract information for target localization and secondly, to extract the discriminative features to perform correct classification.

1.4.1 Gun sounds

Gunshots have also been well studied [8] and there are three main acoustic gun sounds: muzzle blast, mechanical action and shock wave. The mechanical action is generally much quieter than the muzzle blast and the shock wave, so this signal is only present if the microphone is located close to the firearm.

A conventional firearm uses an explosive charge to propel the bullet out of the gun barrel. Most of the acoustic energy is emitted in the direction the gun barrel is pointing. Some handguns and rifles can be equipped with an acoustic suppressor to reduce the sound signal of the muzzle blast.

The shock wave is only present when the bullet is going supersonic. The supersonic projectile's passage through the air launches an acoustic shock wave propagating outward from the bullet's path. The shock wave expands in a conical fashion behind the bullet, with the wave front propagation moving outwards at the speed of sound.

1.4.2 Vehicle sounds

Some studies have been done into acoustic vehicle detection [9]. Normally, vehicles under normal operating conditions produce unique acoustic signatures. The signature of these vehicles varies depending on the vehicle type, but also on the vehicle dynamics, such as engine speed, load and road surface. Vehicles of the same class under similar operating conditions generate similar acoustic signatures that can be used for classification. More precisely the sounds from the vehicles are caused by [10]: the rotational parts, the vibrations in the engine, the friction between the tires and the pavement, the wind effects, the gears and the fans.

It has been suggested that the important features of vehicles which are useful for classification lie in the range of 25 to 400 Hz [11]. It is also said that military vehicles have strong harmonic signatures that can be used to classify them. Unfortunately, most civilian vehicles do not have acoustic characteristics suitable for classification.

1.4.3 Human sounds

People can make many different sounds, and most of these, like talking, coughing, screaming, sneezing, snoring, laughing, groaning and crying derive from the mouth. Many of these sounds can be used to increase situational awareness. Speech is the most extensively investigated area, but more from the point of view of studying language and identifying words. Studies have also been done for medical purposes [4], [12].

Although the mouth makes most of the human sounds, the rest of the body also produces some sounds. A footstep is a sound source that has also been investigated [1], [13]. The acoustic signature of a footstep can be used for human recognition, and even identification is suggested. When a human is walking, the sound characteristics of a person's footsteps are determined by three dominant conditions: footwear (sneakers, bare foot, etc.), ground surface (concrete, wood, etc.) and gait (individual motion, speed, etc.). The main difficulty in recognition derives from the change in these three conditions. A footstep signal mainly consists of two predominant characteristics. Firstly, the (striking) force normal to the supporting surface and secondly, the (friction, sliding) tangential force.

1.4.4 Feature extraction

Feature extraction is the challenging aspect required for target classification. Feature extraction is equivalent to making an acoustic class fingerprint of the recorded signal. The crucial task when creating a successful classification is to construct signatures built from characteristic features that enable discrimination between classes. The accuracy and the distinctiveness of the provided features determines the classification performance. The feature extraction from the acoustic signals emitted by targets, which can be hampered by many environmental factors, makes feature extraction difficult.

Many feature extraction methods are suggested, and sometimes the term feature extraction is used vaguely. Generally the extraction is based on time models, Fourier transforms or wavelet analysis. There are many different designs, and even face recognition techniques are suggested [14]. Usually a high sampling frequency is required for good feature extraction, but it has also been suggested that sparse sampling around 10Hz could be sufficient for classification [15]. Although sparse sampling is energy efficient, other problems arise.

Thus there are many algorithms, but some are relatively complex and computationally exhausting and so those methods are just not effective or good. Amazingly, with gunshot detection it has been suggested that difficult techniques should be avoided and that just the absolute value of the signal should be considered [16].

1.5 Localization

When a target is transmitting an acoustic signal and the signal is received at different sensors, target localization is possible. After the correct information has been extracted from the nodes and the localization equations (which are based on the acoustic wave propagation) have been constructed, localization is merely a mathematical problem in which the target position is the unknown.

Target position estimation is relative to the node positions and positions can be defined in multiple ways. For example, the nodes positions can be calculated relative to other positions. In a three-dimensional universe, localization can be an estimation of the three distance position components of the target.

There are two methods available for passive position estimation: power-based localization and time-based localization. Power-based localization is based on propagation loss and time-based localization is based on propagation time. Both methods require Line-of-Sight signals for correct position estimation. Therefore, both are sensitive to: shadows regions, multi paths and other signal effects.

The localization techniques are sometimes known as received signal strength and Time Difference of Arrival (TDOA). TDOA is considered to be more robust and is the technique that is most frequently used [17], but for stationary signals TDOA localization is impossible, because there is no clear beginning or ending to the signal. Sources with stationary signals can be localized with received power, but require extensive environmental calibration and are most often used with tracking algorithms [18], [19].

The mathematical solver method, which is an active research field in itself, is needed to estimate the positions with the equations and the observed information. The idea with least square solvers is to find a solution (target position), which best fits the measured data.

1.6 Classification

Classification is the process of deciding to which class a target belongs. Before something can be classified, features have to be extracted. The classifier uses the extracted features to discriminate between classes: a classifier transforms feature input into class output. The classifier should use the discriminative features to accomplish its goal. Some classification methods are discussed in Appendix E.

With classification, pre-knowledge is needed. The system cannot recognize something if it has never cognized it before. Pre-knowledge can, for example, come from a training set of feature vectors, in which every training feature vector is assigned to a class.

All classification methods depend on the provided features. In other words, the discriminative characteristics need to be in the features and there is a classification limit with certain extracted features. For example, the Bayes error rate, which is a challenge to compute, can provide the lowest achievable error rate for a given classification problem [20].

There are many ways to create a classifier, and it is a field of study in itself. Neural Networks is a well known method and has many different learning variants [21]. System learning sounds very exciting, but in almost all cases it is just an estimation of the classifier parameters. In system learning there are two approaches: supervised learning and unsupervised learning. In supervised learning the system knows what the (output) class should be for each training example. In unsupervised learning, which is also sometimes known as clustering, the system does not know the class and must construct clusters independently.

There are many designs for feature extraction and classification. Due to the related difficulties and complexity, many acoustic classification systems use system learning methods to arrive at a classification system. For example, a Fast Fourier Transform is chosen for feature extraction and the total result is given to a kind of Neural Network. Although this requires little design, the classification performance is unpredictable in new situations which were not included in the training set. Furthermore, this approach provides little insight into the classification system.

In binary classifiers, the output can only have two values. Multiple classifiers can also be combined into one classifier tree. For example, binary classification trees can be used for multi-class classification problems [2], [22].

1.7 Graduation project

This graduation project is initiated to assess the potential, feasibility and reliability of an ASN for surveillance purposes in urban environments. Because of the great demand and the political sensitivity, most of the research is done by companies and military which do not publish their work. As Section 1.3 discussed, the acoustic complexity is huge and everything indications that this is a tough topic. Much work is required to realize an ASN, but little has so far been done which is why the aim in this graduation project is to investigate data processing for acoustic localization and classification of targets. A difficulty for this project is the current limited knowledge on ASN. At this stage, it is too difficult to investigate acoustic surveillance in realistic complex situations and therefore a simpler operational situation is assumed. Although this situation is simpler, this will show the potential and feasibility of ASN in urban environment.

1.7.1 Assumed operational situation

The data processing will be designed for a certain operational situation. The operational situation is the circumstances in which the system should work. The Experimental ASN consists of four nodes, each with an omni-directional microphone and a processing station. All the data is simultaneously recorded and available at the processing station, without including the need to involve network intelligence. In other words, the nodes' system clocks are synchronized. The three-dimensional node positions are also available at the processing station.

The EASN will be deployed in an urban environment which will mainly contain air as a transport medium for sound. Another assumption is that when a target is present, all four nodes will be closer or at a distance of ten meters from the non-moving target and all the four sensors will receive a Line-of-Sight signal from the target. This is a reasonable operational situation for the first EASN and an example of such an operational situation is sketched in Figure 1-3.

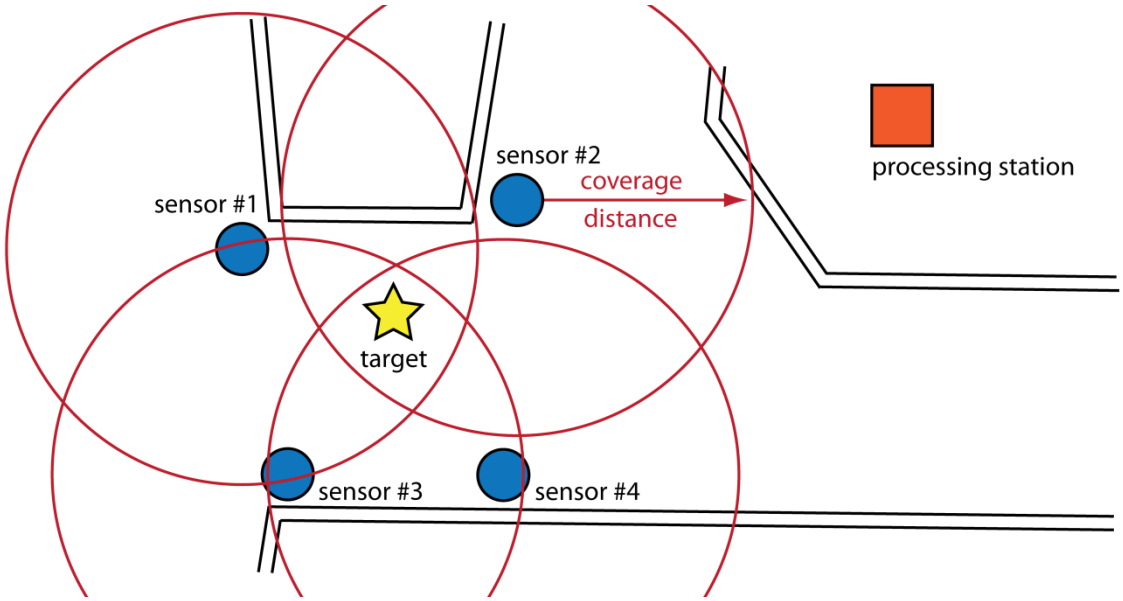


Figure 1-3: An operational situation

The following ground target classes and the following sounds may be present in the covered area:

- Gun (muzzle blast)
- Vehicle (running piston engine)
- Human (walking pedestrian)

Although a pedestrian moves, individual footsteps do not. It is assumed that at a certain time at most one target is present in the covered area. At the current research stage it is too early to be very specific about signal to noise ratio (SNR) requirements, however the following rule of thumb will be used: in practice an SNR of minimally 20dB is necessary for robust processing.

1.7.2 Project objectives

Although some target classes are presumed to exist, it is still a great challenge to cope with the received signals in a passive ASN, because signals can - even within the same classes - differ significantly. Furthermore, the target has absolute no desire to be localized and classified. This problem makes it very challenging to achieve robust localization and to achieve correct classification. The goal of this project is to design a part of the EASN data processing which should do the following with the recorded signals and node positions:

1. Provide target information with propagation models, which allows a least square solver to estimate the target position, for localization purposes.
2. Provide target features, which allows a classification method to discriminate between the different target classes, for classification purposes.

The main focus of this project is on extracting the right information and constructing equations for localization and extracting the optimal features which will allow discrimination for classification. The relevant information needs to be separated (filtered) from the irrelevant information. The aim is to use discriminative physical features, in other words, features which are explainable and understandable. The challenge is that this should be achieved with only the recorded sound signals from a passive EASN deployed in the urban environment, which will alter the signal, and that nothing is known about the target. To test the result and to further investigate the potential of ASN in urban environments, the EASN will be designed and implemented (proof of concept). The EASN will combine localization and classification of the three selected targets.
There exist many topics, methods and solutions to cope with the project problems. However to set the project boundaries, the following topics have been excluded from the project:

- Hardware design and technology
- Network-communication, energy efficiency and data-fusion
- Self-localization and time-synchronization
- Acoustic environment and target modelling
- Least square, classification and tracking methods

Thus, for example, the construction of a mathematical solver or a classification method is considered to be beyond the scope of this project. The above topics are active research areas in themselves and will therefore not be considered in this thesis.

1.7.3 Project approach

To accomplish the project objectives, the project has to be based on an effective approach. First the current knowledge of acoustics, localization, classification is studied to investigate the general possibilities. This introduction chapter and some parts in the appendices are based on this study. Next the acoustic sound signals of the selected targets are measured and investigated to find out how the localization and classification of these target can be achieved with sound. The investigation will clarify what kind of information and features can be extracted from the target signals. Then with this knowledge, methods will be designed to extract the correct information for localization and to extract the features required for classification for a certain operational situation. After the methods have been designed, they will be tested and evaluated with the experimental equipment in an experimental environment to obtain a performance indication.

Many methods and solutions can be used with different properties, but not all are suitable for the project goal. The following four criteria are desirable in the experimental selection process:

- The method *complexity* must be low. The method implementation must not be too difficult.
- The method has been used before and therefore it is possible to say how well it performs. This criteria is called: *previously proven*.
- What is the *effectiveness* of the method? Is the result in proportion to the required computational power and time.
- Is it likely that the method will perform well in the system? It is wise to choose the method with the most *potential*.

The above criteria form the foundation for the designer.

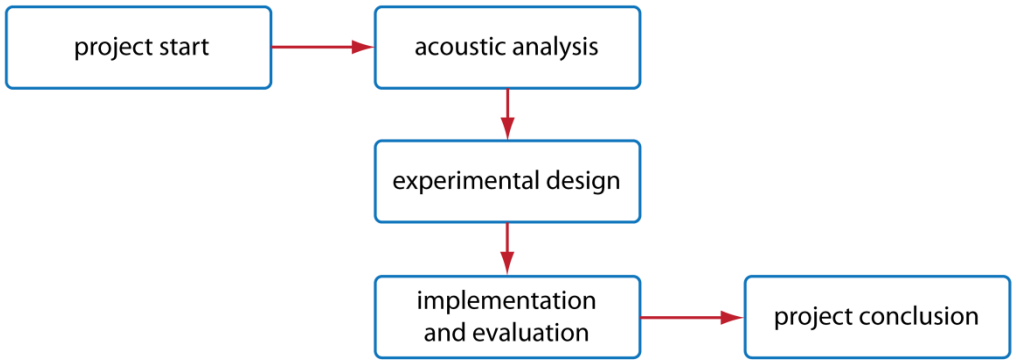


Figure 1-4: Project approach

The project approach is summarized in Figure 1-4. This thesis has a similar mapping to the project approach. This chapter has discussed the project. Chapter 2 will discuss the acoustic analysis of the considered target. In Chapter 3 the high-level system components are designed for component distinctions. Chapters 4, 5 and 6 will design the system components. The designed components will be implemented in Chapter 7 and they will be evaluated in Chapter 8. Chapter 9 outlines the conclusion and gives the results and recommendations.

2 Acoustic Analysis

Measurements of the acoustic environment and targets are required for an open-minded design approach. Hardware study is considered beyond the scope of this thesis, but a short experimental hardware and environment analysis is provided in Appendix A. The main focus of this chapter is the investigation of the target sounds and this is needed to clarify what kind of information and features can be extracted to perform localization and classification. Can signals received at acoustic sensors be used to estimate the target position and to extract characteristic target features?

This chapter will review the results of the acoustic measurement from the three targets: guns, vehicles and humans. The measurement results of the gun muzzle blast is discussed in Section 2.1. The running piston engine will be acoustically investigated in Section 2.2. Section 2.3 investigates the walking pedestrian.

2.1 Gun muzzle blast

For the gunshot measurements, two shooting clubs were asked if gunshots sound recordings could be made. The first one allowed measurements indoor, but the second allowed measurements outside. It appeared during the measurements, that a gunshot sound was too loud for the experimental hardware to be placed near by the handgun/rifle. Therefore the microphones had to be placed far away from the target and on the microphone a yellow ear-damper was placed to attenuate the high frequencies. Further experimental hardware study is provided in the Appendix A. This section will discuss the gun muzzle blast result.

The first location was at S.V. Doel Treffend in Delft. An indoor shooting range of 25 meters was available on 17 May 2010. Due to the mapping of the shooting range and because the microphones had to be placed far from the target, the microphones had to be placed on the shooting range. The measurement setup is given in Figure 2-1.

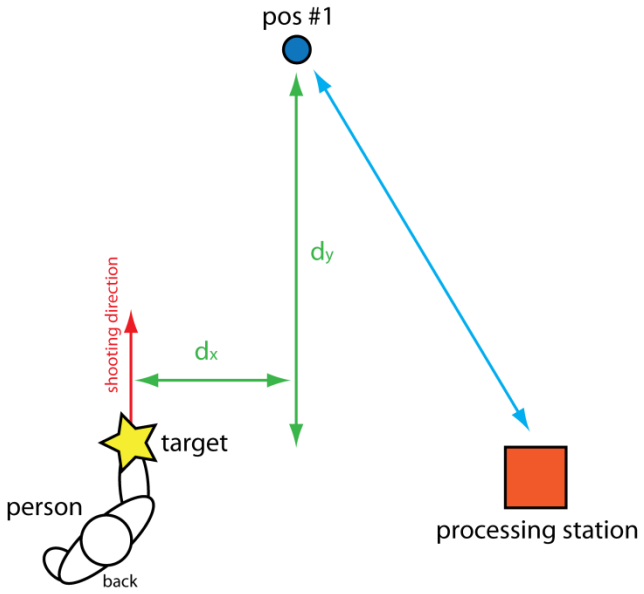


Figure 2-1: Measurement setup for recording gun muzzle blast in Delft

At the microphone position two microphones were placed, both at a height of 1 meter. The second microphone has an ear-damper. The height of the handgun was approximately 1.5 meters, distance d_x was approximately 2 meters and the distance d_y was approximately either 10 or 21 meters.

At Doel Treffend two handguns were used. The first one was a .22 caliber (5.7mm) revolver, the second was a 9mm pistol. Recordings at d_y equal to 10 meters were taken of multiple .22 caliber shots. The standard recording is given in Figure 2-2 and the ear-damp version is shown in Figure 2-3. As the recordings show, the non-damped microphone provides a strange and probably incorrect recording due to the limits (clipping) of the experimental hardware. With the ear-damper the signal is damped (mainly the high frequencies), but the experimental hardware could record a normal signal almost without this clipping (still a little bit at 0.0145 second).

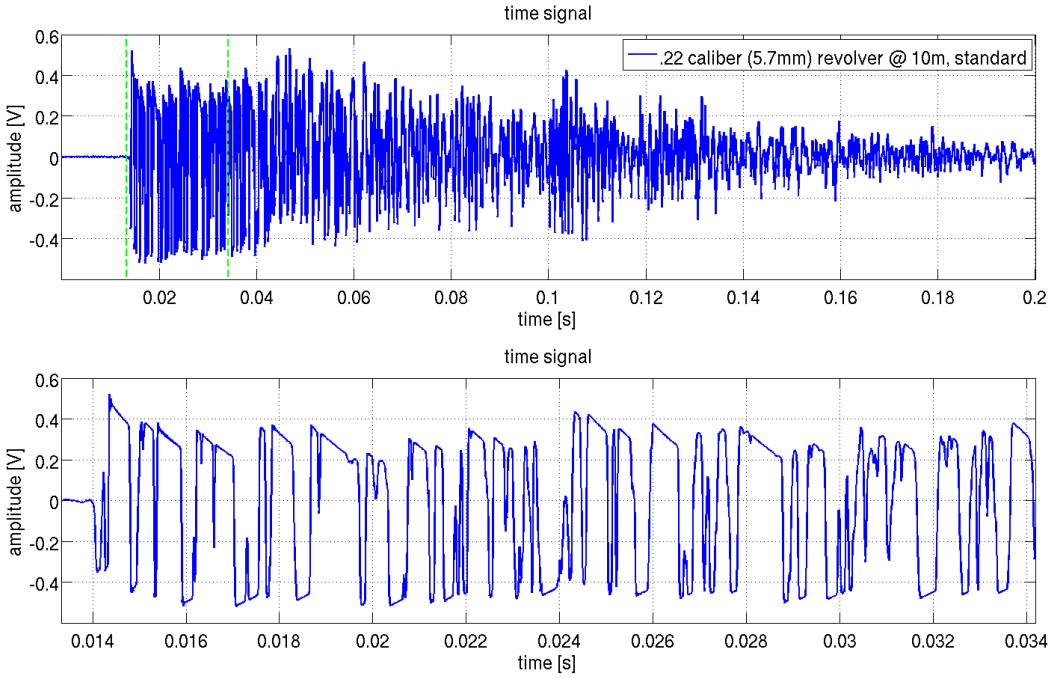


Figure 2-2: Revolver standard measurement at 10 meter in Delft, green dotted line shows zoomed area

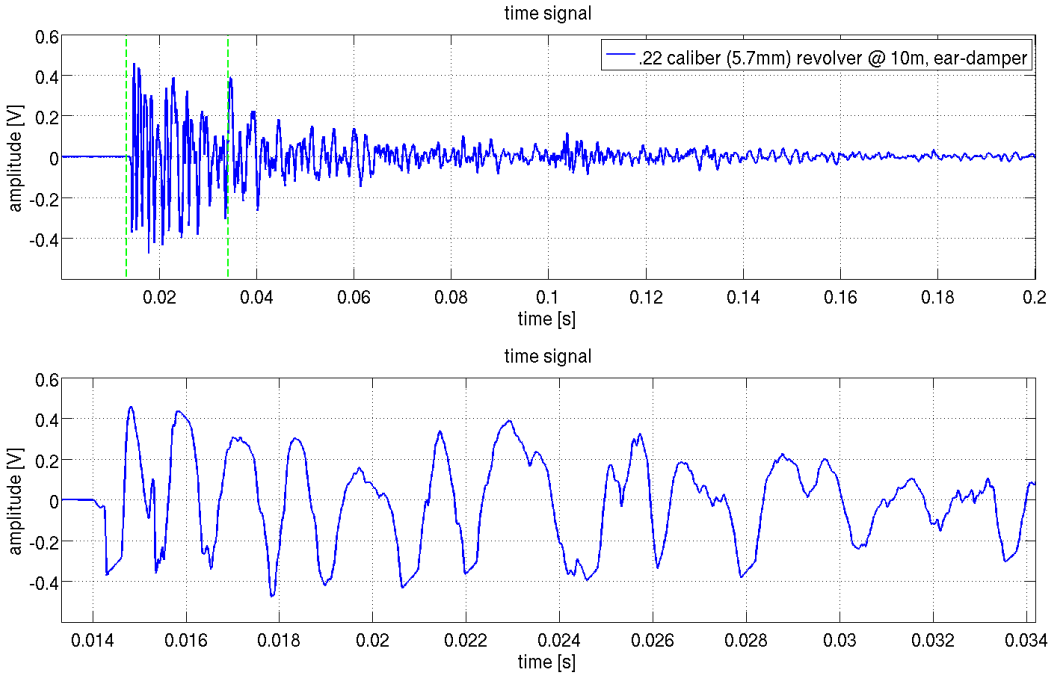


Figure 2-3: Revolver ear-damp measurement at 10 meter in Delft, green dotted line shows zoomed area

An similar measurement was taken at 21 meters with the same .22 caliber. Note that the shooting range was approximately 25 meters long. Although the distance was larger, clipping was still present with the standard measurement. The ear-damper result is given in Figure 2-4. With this measurements it is still difficult to see the original signal and the reflections (reverberation).

Another point to notice is that an explosion of a muzzle blast consists of two parts, the first part the primer and the second part the gun powder to propel the bullet. Whether these two sounds can be seen in these recordings is difficult to say. For now, it is certain that the signal energy is lower at a greater distance.

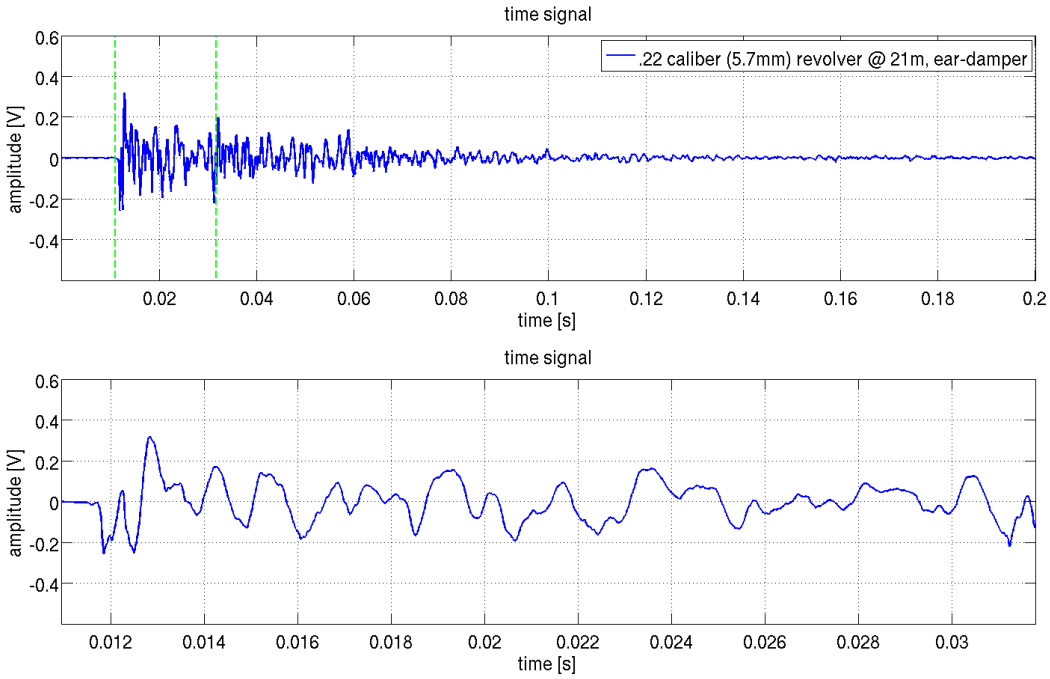


Figure 2-4: Revolver ear-damp measurement at 21 meter in Delft, green dotted line shows zoomed area

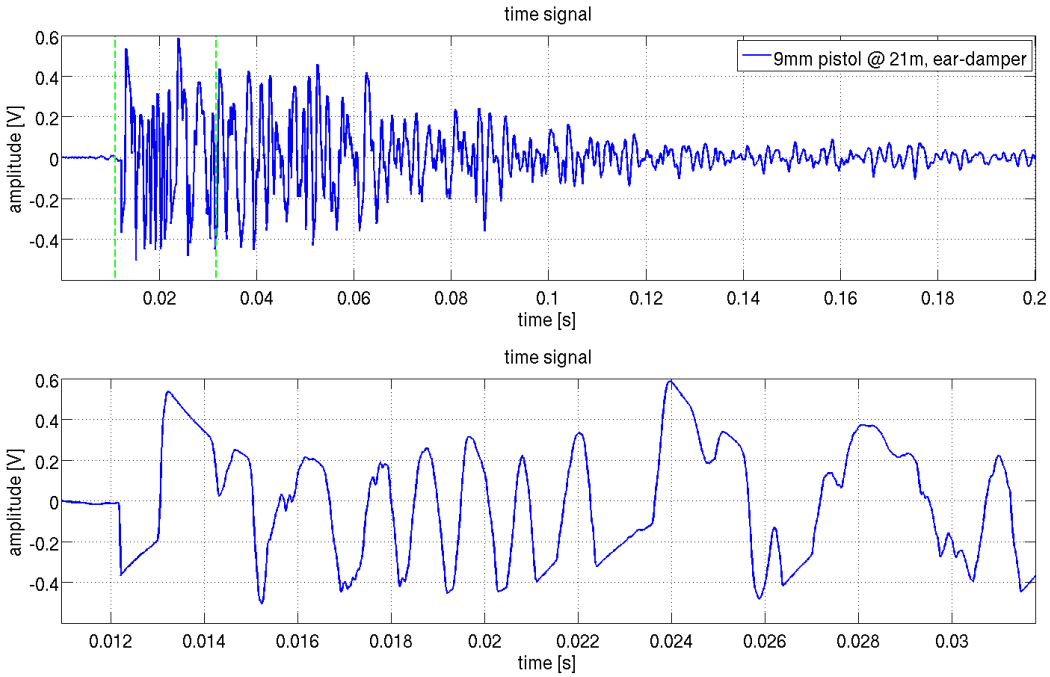


Figure 2-5: Pistol ear-damp measurement at 21 meter in Delft, green dotted line shows zoomed area

After the .22 revolver, the 9mm pistol provided a much louder sound. The ear-damped recording at 21 meter is shown in Figure 2-5. Even with the ear-damper at 21 meter in Figure 2-5 the signal was too strong for the experimental hardware. The three ear-damped signals with a spectrogram are given in Figure 2-6 and Figure 2-7. A spectrogram better shows the (wide-band) reflections.

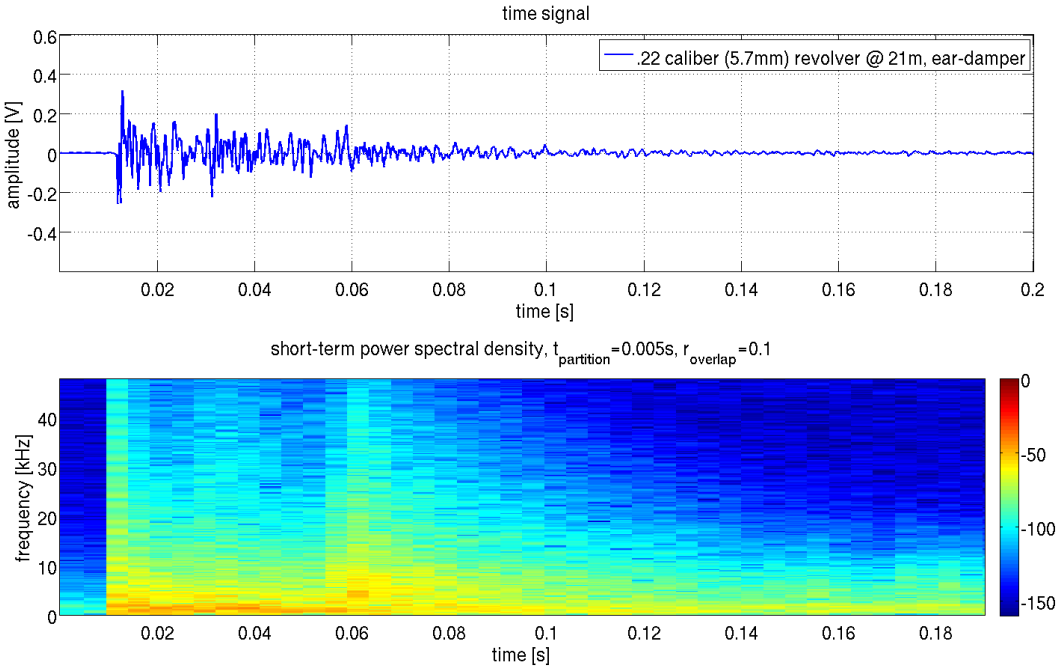


Figure 2-6: Revolver ear-damp measurement at 21 meter in Delft

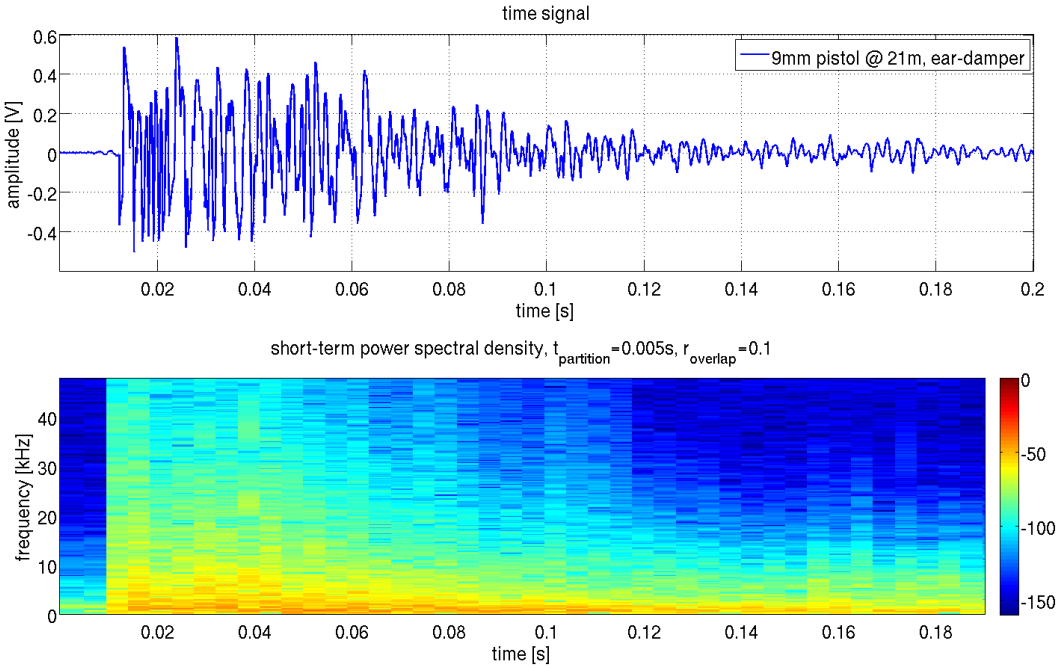


Figure 2-7: Pistol ear-damp measurement at 21 meter in Delft

Although some recordings shows a strange effect (tight slope), which can be a sort of clipping, the recordings sound still normal for the human ear. Thus, when the recordings were played after the experiment, the recordings seemed to be normal. It seems that the human ear is not very sensitive for these high frequencies or that the human ear is also limited in a certain way to notice this. The ear-damped recordings sound similar as for the people around the gunshot with (other kind of) ear-dampers.

The second shooting location was at JST-Waalsdorp in The Hague, which allowed outdoor measurements. The date was 19 May 2010, the sky was blue, the sun was shining and gunshots were fired with a shotgun caliber 12 under/over. The location allowed a more flexible deployment of the microphones, and three positions were defined as Figure 2-8 shows.

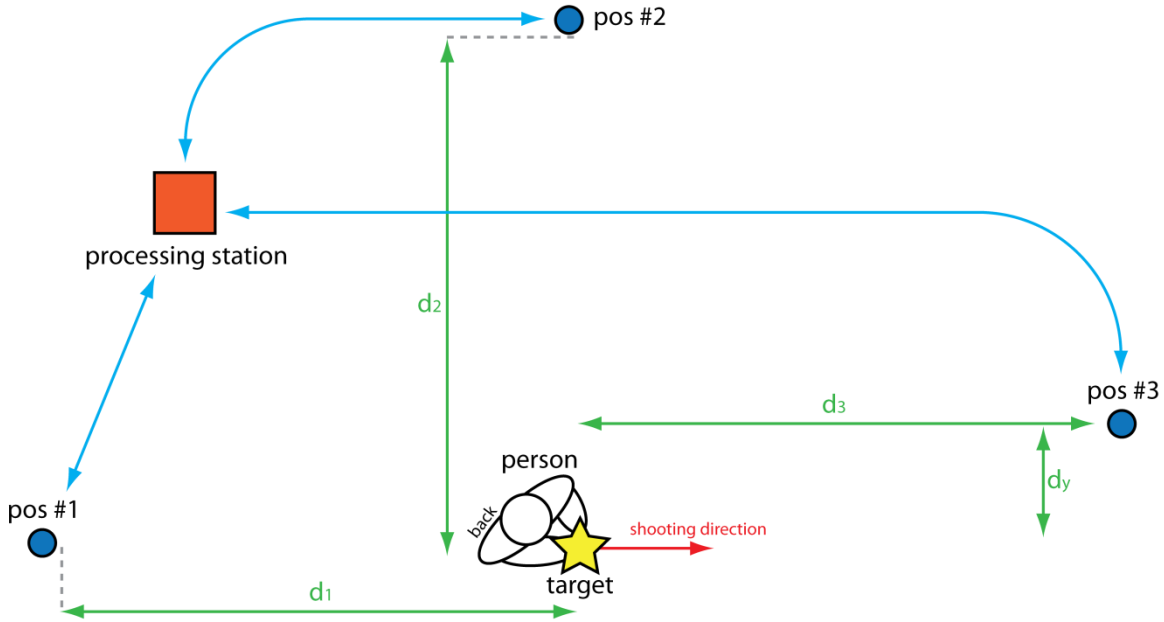


Figure 2-8: Measurement setup for recording gun muzzle blast in The Hague

The distances d_1, d_2 and d_3 are 10 meter and the distance d_y is 1 meter. The same two microphones were placed at 1 meter height.

The received signal at the back was the weakest and the received signal at the standard microphone is given in Figure 2-9. Also at this location and with this gun, at a distance of 10 meters the signal was still very strong. Figure 2-10 shows the ear-damp version of the same signal. The result is almost a short impulse, and the reflections are more separable then in the indoor location.

Figure 2-11 shows a received ear-damp shotgun signal at the left of the shooter. The ear-damped version at the left shows already a little clipping in the beginning (0.01 second) of the signal.

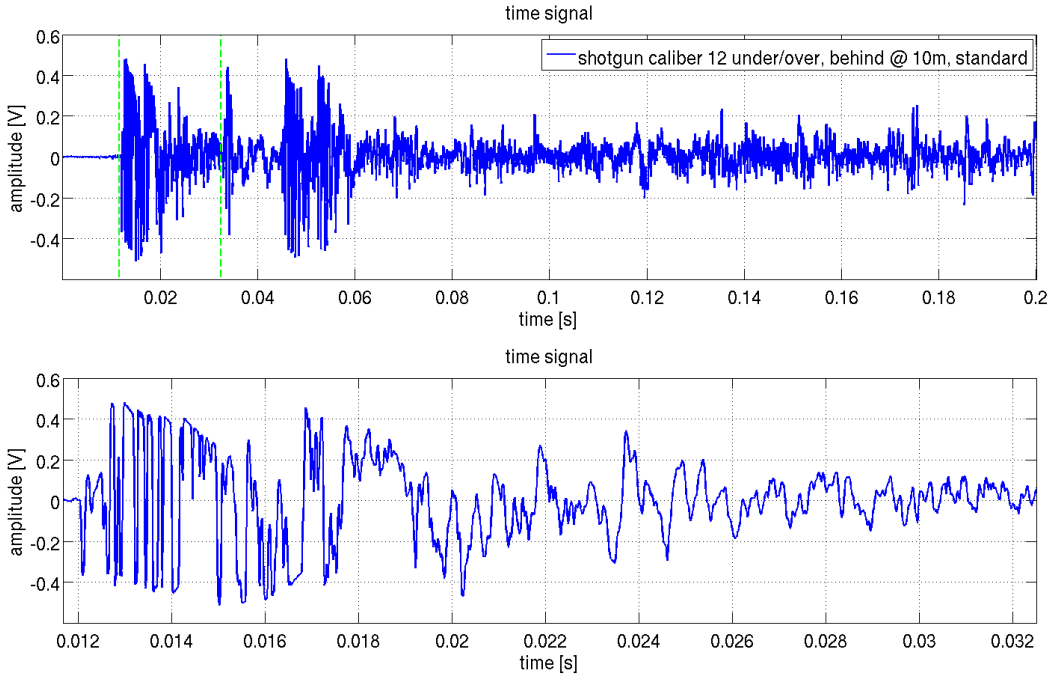


Figure 2-9: Shotgun standard measurement behind at 10 meter in The Hague

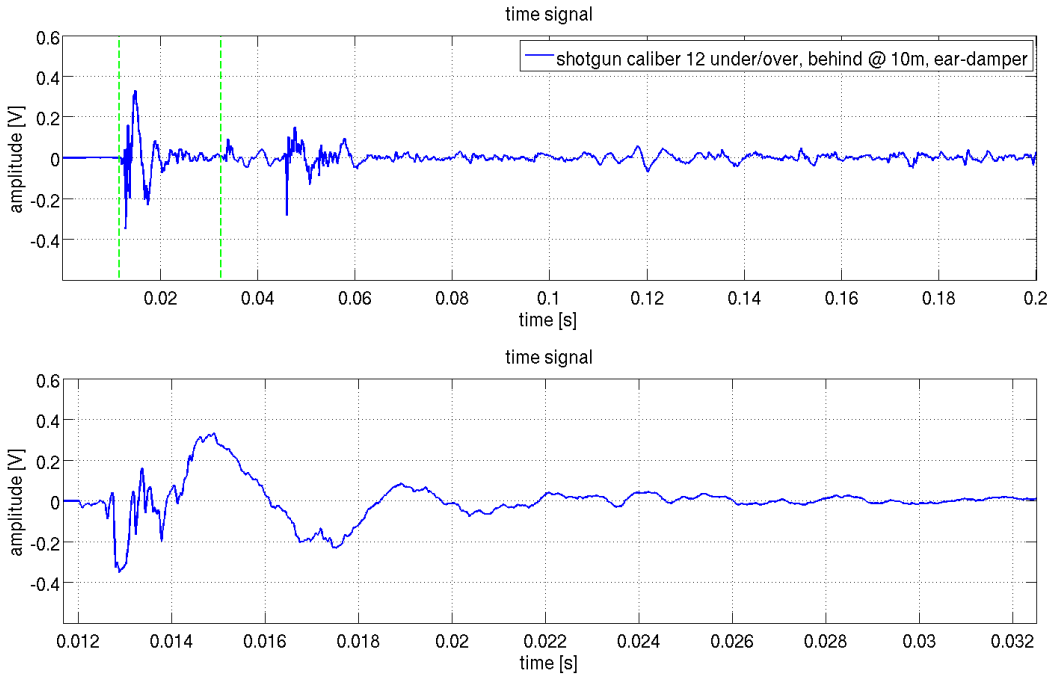


Figure 2-10: Shotgun ear-damp measurement behind at 10 meter in The Hague

At different positions the reflections are different and this is also the case when the microphone is placed in front of the gun. Figure 2-12 shows the received signal at microphone position three and it further shows the limits of the used microphones for these kind of distances. It becomes

clear that the received power at the front of the gun is the strongest, and behind the gun the received power was the lowest.

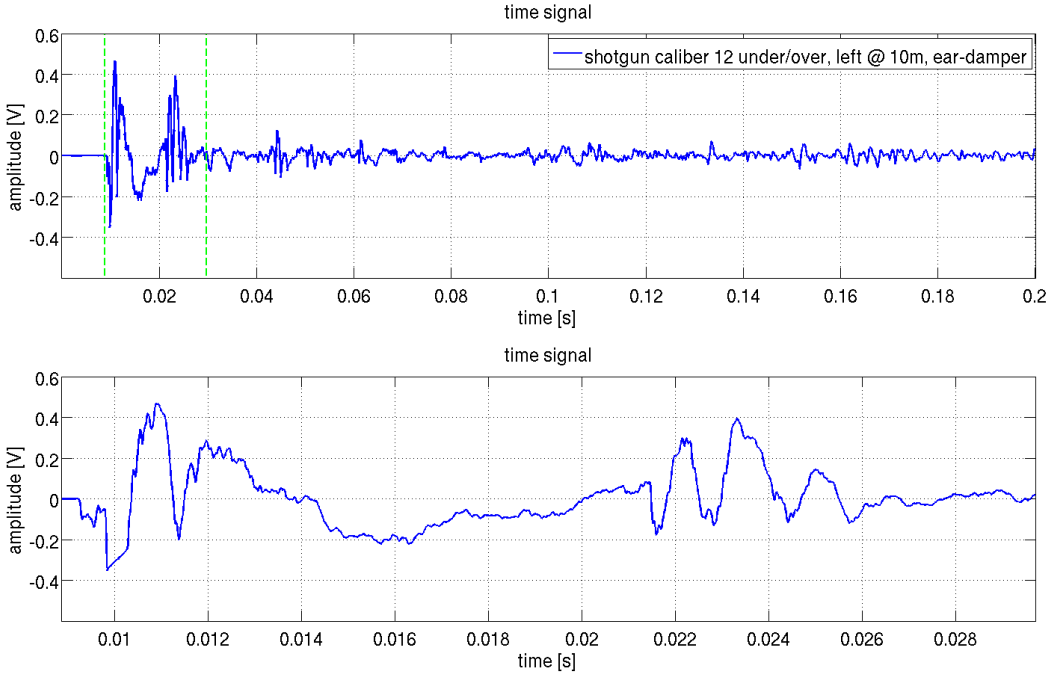


Figure 2-11: Shotgun ear-damp measurement left at 10 meter in The Hague

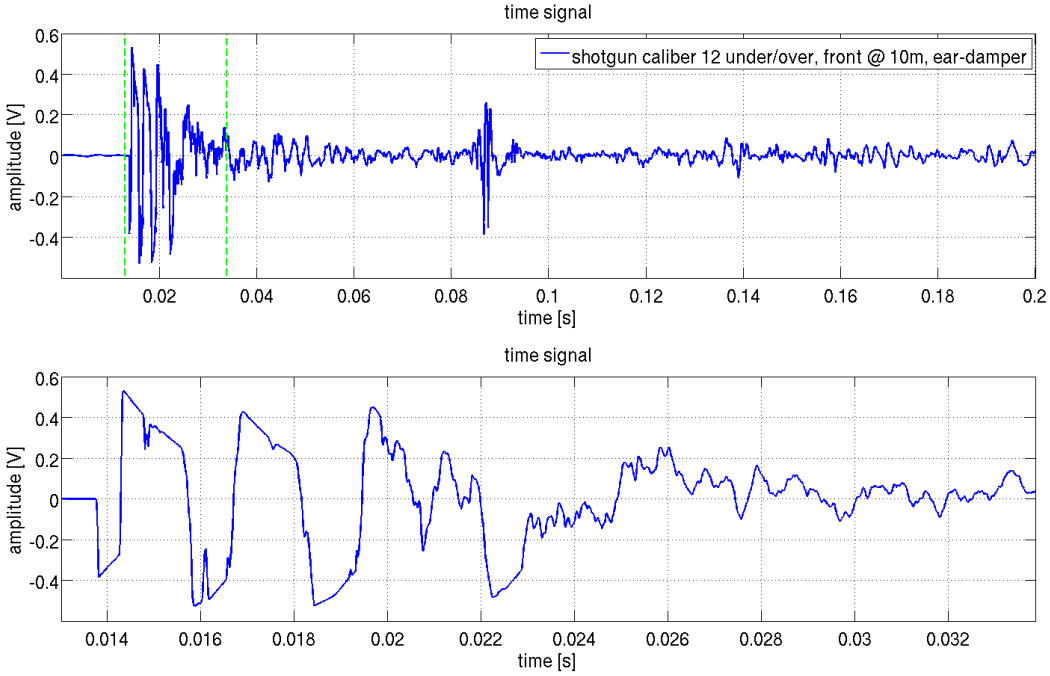


Figure 2-12: Shotgun ear-damp measurement front at 10 meter in The Hague

The outdoor environment allows a better analysis of the reflections. The spectrogram can give a good insight and in Figure 2-13, Figure 2-14 and Figure 2-15 the ear-damp signals are given with a spectrogram. Due to different microphone positions, the reflections are different and this is because the environment was not perfectly flat and some fences were placed at the location.

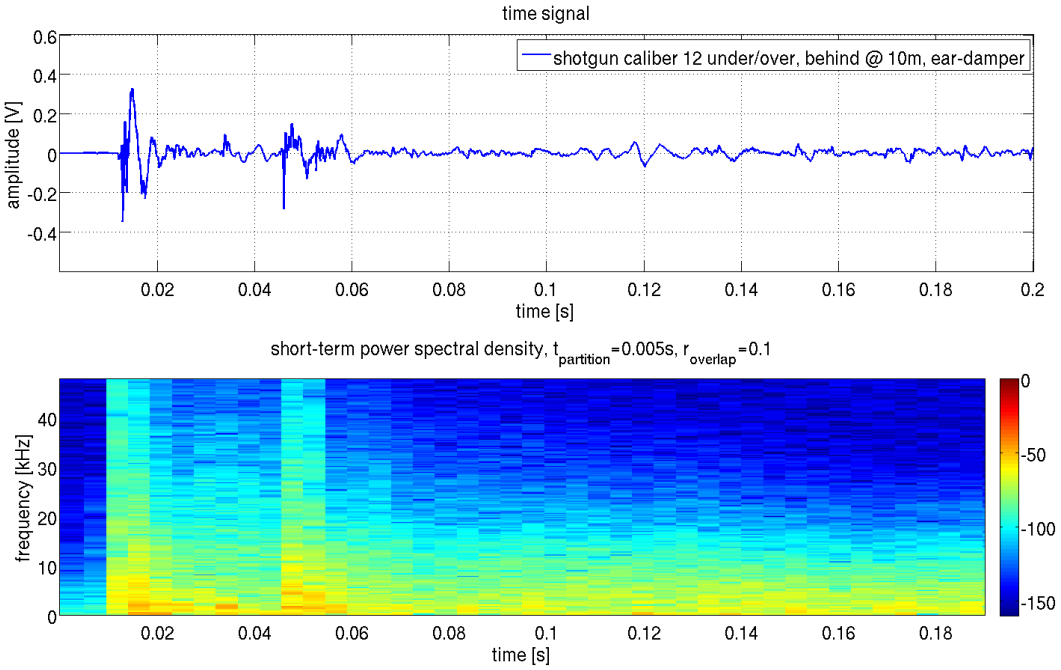


Figure 2-13: Shotgun ear-damp measurement behind at 10 meter in The Hague

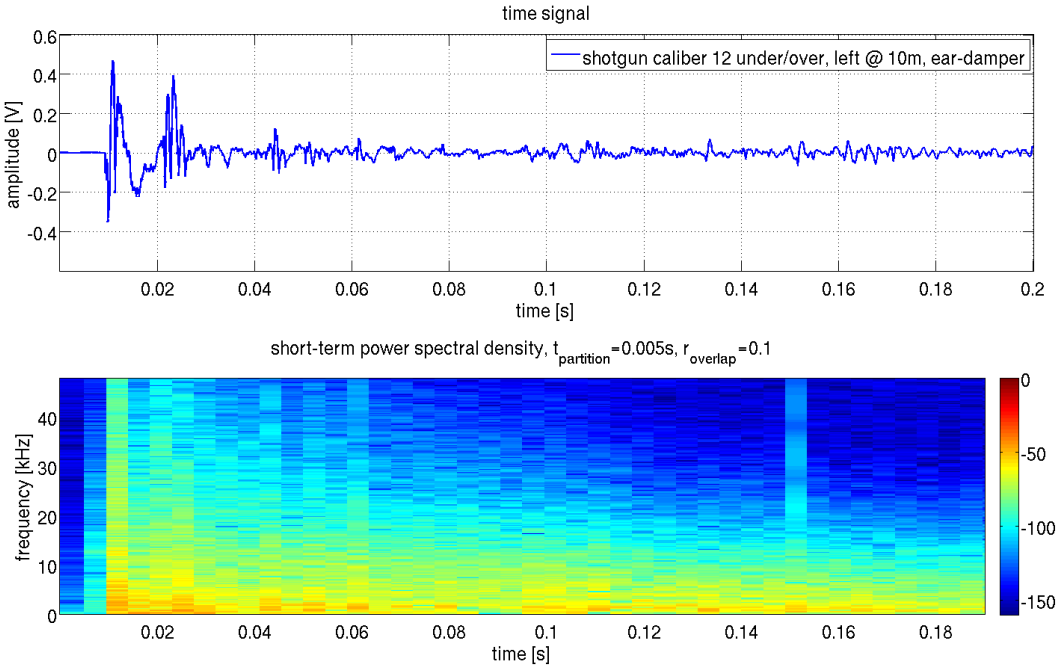


Figure 2-14: Shotgun ear-damp measurement left at 10 meter in The Hague

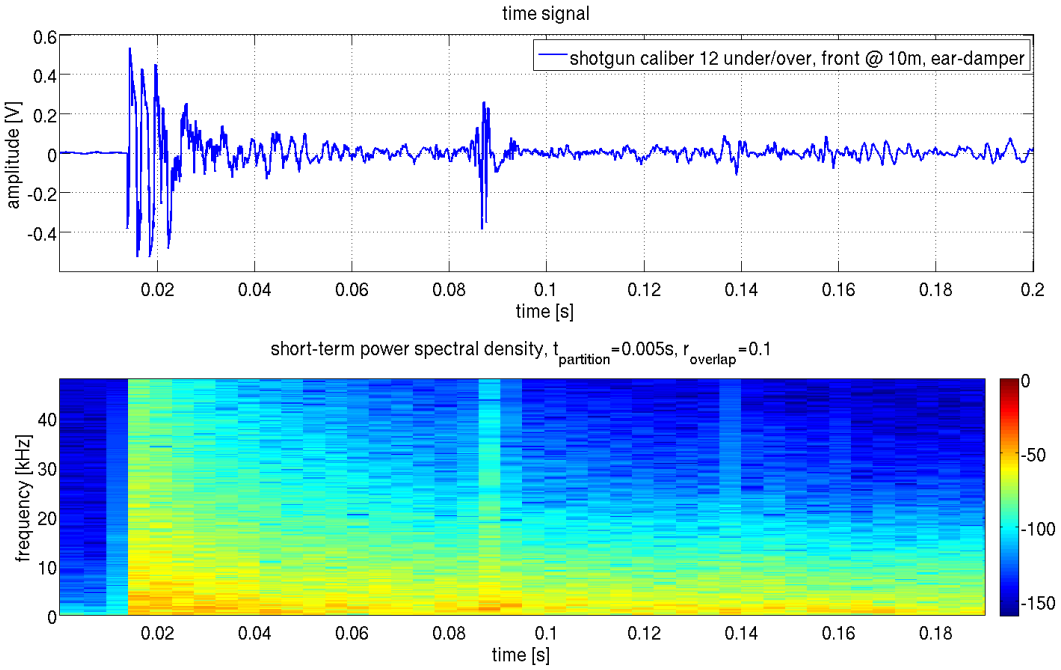


Figure 2-15: Shotgun ear-damp measurement front at 10 meter in The Hague

For the spectrograms two parameters are defined, first the $t_{partition}$ which is the partition size in time and second the partition overlap ratio $r_{overlap}$. For every partition the power spectral density is determined and the overlap ratio determines how much a partition overlaps with the previous partition.

With these indoor and outdoor gunshot measurements, four conclusions can be made:

1. The received signal is angle dependent
2. The emitted energy is very high
3. The emitted signal is short in time, but also wide-band in frequency
4. With the same ammunition and angle, the gunshot sounds are very consistent

In previous research [8], which also studies the shock wave, has a similar conclusion for the acoustic muzzle blast.

2.2 Vehicle running piston engine

This section will discuss measurements of a Renault Laguna with a 2.2 liter engine and four cylinders at the Thales Delft parking area. The measurement date is 30 March 2010, the wind speed was 5 m/s and it rained a little.

The vehicle measurement setup for a running engine is shown in Figure 2-16. It shows that there are three different microphones positions. The distance d_i was measured without the vertical height. At every position two microphones were placed, the first microphone at 1 meter height and the second at 0.5 meter height.

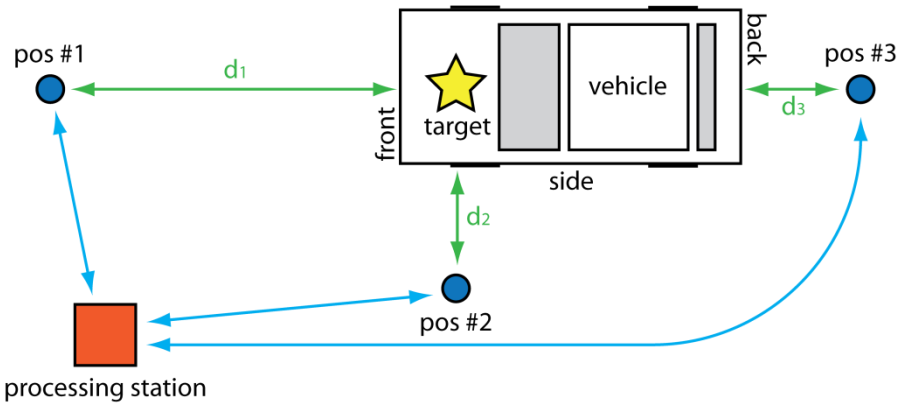


Figure 2-16: Measurement setup for recording vehicle running engine

The sample duration T_s was 3 seconds and the sample frequency F_s was 96kHz. At the beginning of the measurement the environment without a running engine was measured. The vehicle has an engine speed of 1000 Revolutions Per Minute (RPM) when it is running stationary.

The three distances were set to 1 meter and the difference between the different angle positions was investigated. There was no significant difference noticed between the different microphone locations at the same distance to the vehicle, as Figure 2-17 shows with the Power Spectral Density (PSD). However, the highest received power was at the front of the vehicle, which is explainable due to the opening for the air intake. At the back of the vehicle, the signal power was a little bit lower, which is probably due to damping of the vehicle itself and the extra engine distance, but a special signal of the exhaust was not noticed. There was also no significant difference noticed between the two different heights.

Note that the amplitudes of the power spectrum, which is (at least in this thesis) not the same as PSD, with a wide spectrum (until 48kHz) are aggregated, which result in that the PSD is averaged/smoothed over the frequencies. Power aggregation is the power summation of multiple bins and place the result in a wider new bin. If power aggregation is applied in this section, it is done with a factor of 500.

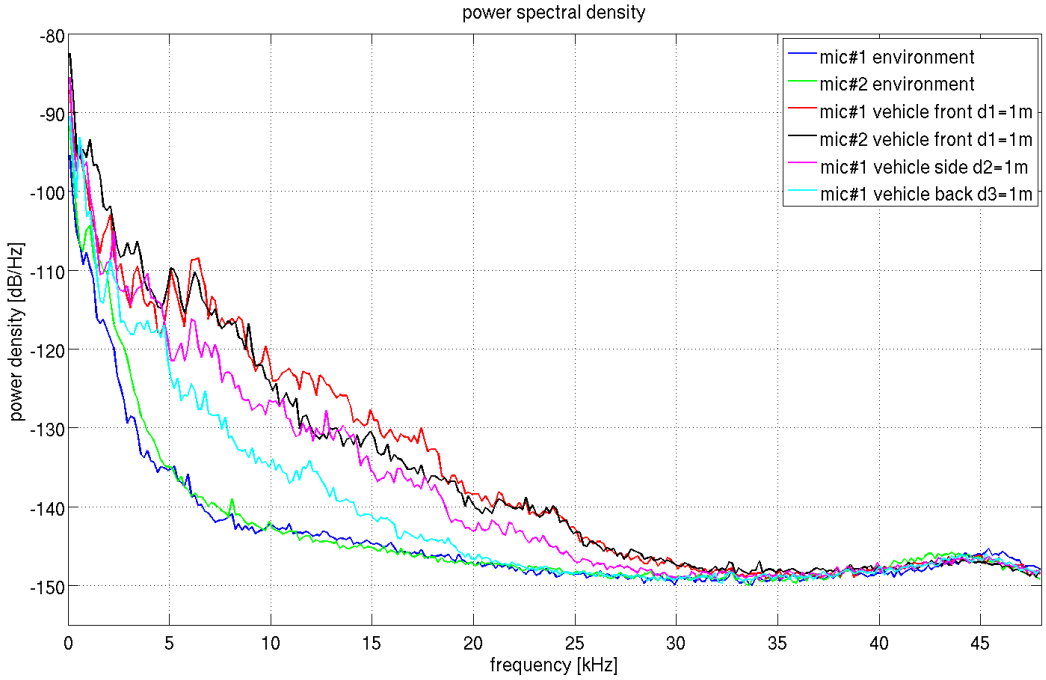


Figure 2-17: PSD of vehicle with different microphone positions.

The second series of measurement was for the investigation of different distances. The distance d_1 was set to 1, 2, 4, 6 and 10 meter. Figure 2-18 and Figure 2-19 show the PSD at different distances of two situations, with an engine running at 1000 RPM and 2000RPM respectively. As the figures show, the received power with 2000RPM is much higher than with 1000RPM.

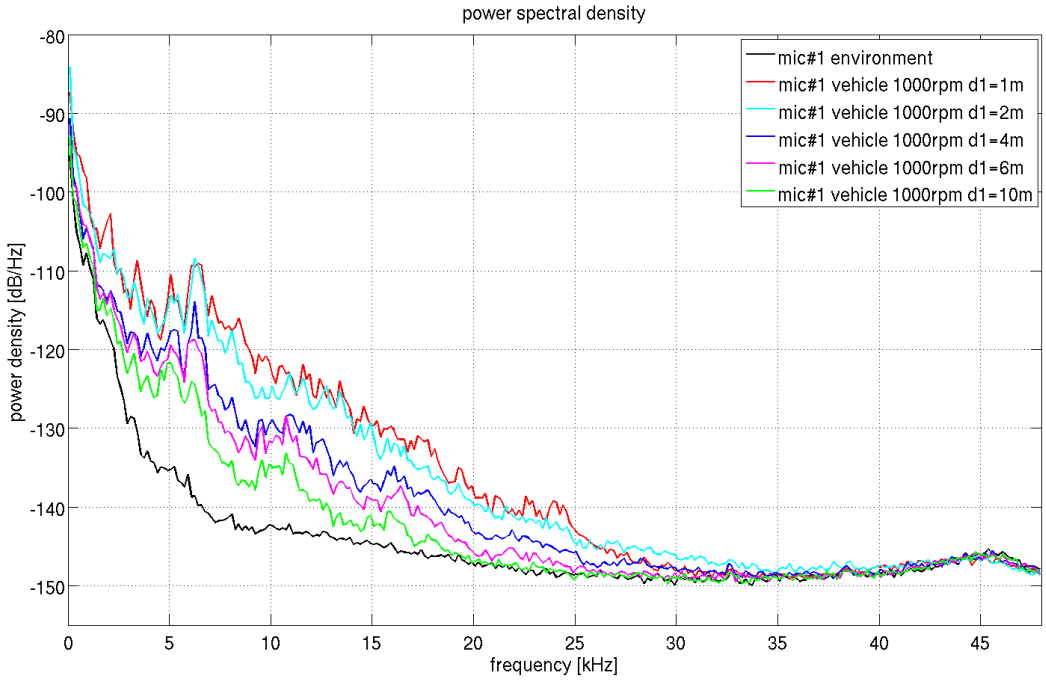


Figure 2-18: PSD of vehicle with 1000RPM

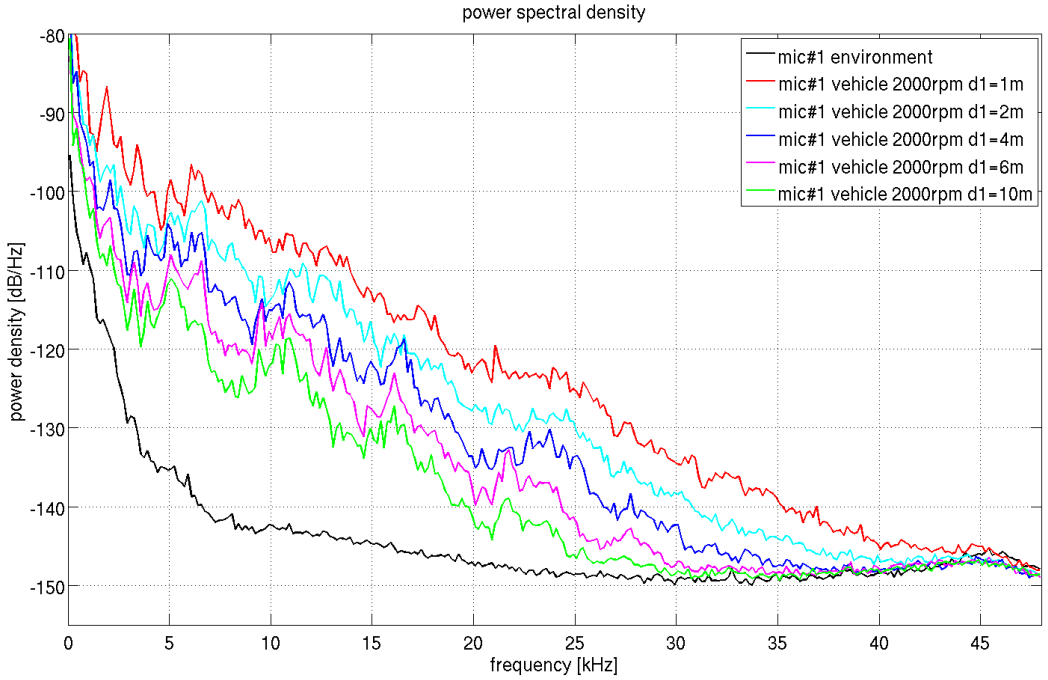


Figure 2-19: PSD of vehicle with 2000RPM

Note that the previous two figures also give a indication for power-based localization performance, which will be further discussed in Chapter 4. The overall received energy is due to the vibration of the vehicle and higher harmonics of the vehicle fundamental frequency. Figure 2-20 shows the PSD with no power aggregation of the low frequency band 0Hz - 300Hz.

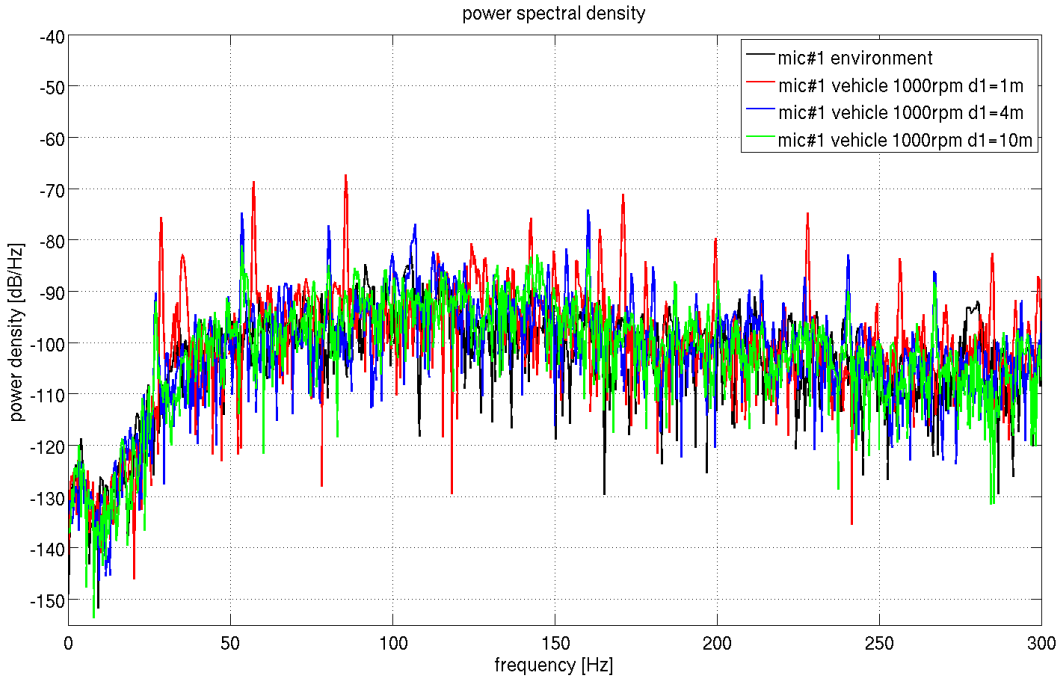


Figure 2-20: PSD at low frequencies of vehicle with 1000RPM

It is known that with a four-stroke engine every cylinder sparks every two revolutions. Thus, with four cylinders two explosions will occur at every revolution. In other words, with 1000RPM and two explosions every revolution, the frequency of explosions would be around 33Hz. This fundamental frequency is seen in Figure 2-20. The same principle is true for the 2000RPM measurements. The fundamental frequency of around 66Hz is present in Figure 2-21 with 2000RPM measurements. Due to the fact that the driver controlled the engine speed with his feet for three seconds, caused the wide peaks at distances 1 meter and 10 meters.

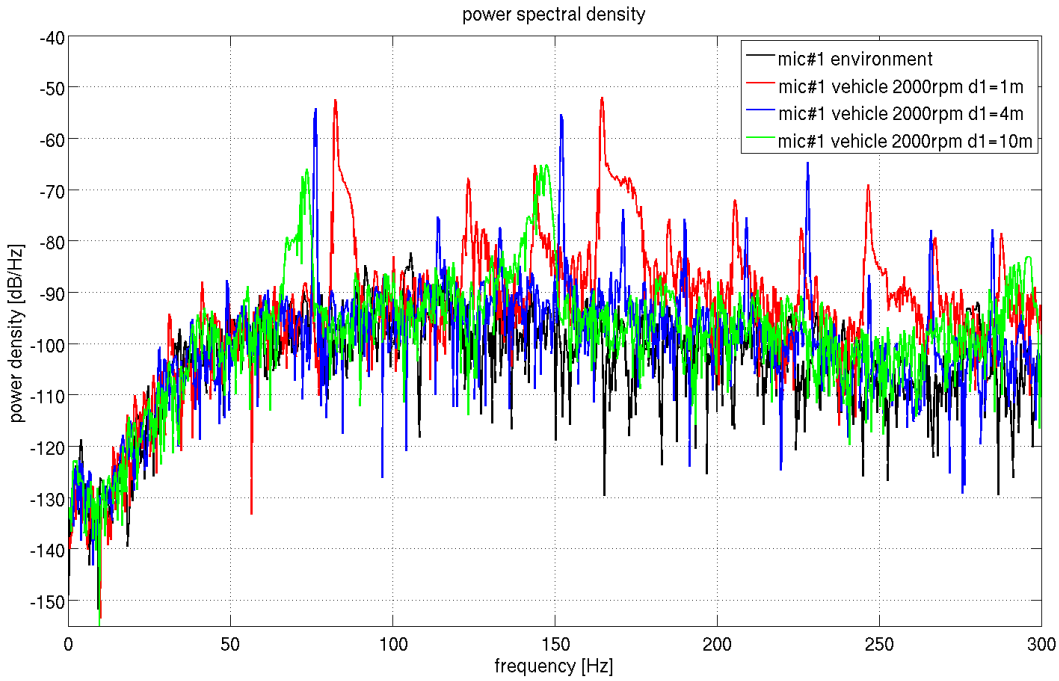


Figure 2-21: PSD at low frequencies of vehicle with 2000RPM

In another measurement, the driver was asked to play 10 seconds with the engine RPM. Figure 2-22 shows a spectrogram of a recording at 2 meters. The plot shows that the harmonics of the vehicle are changing over time, due to the changing engine speed. It also shows that the emitted power changes, when the RPM is changed.

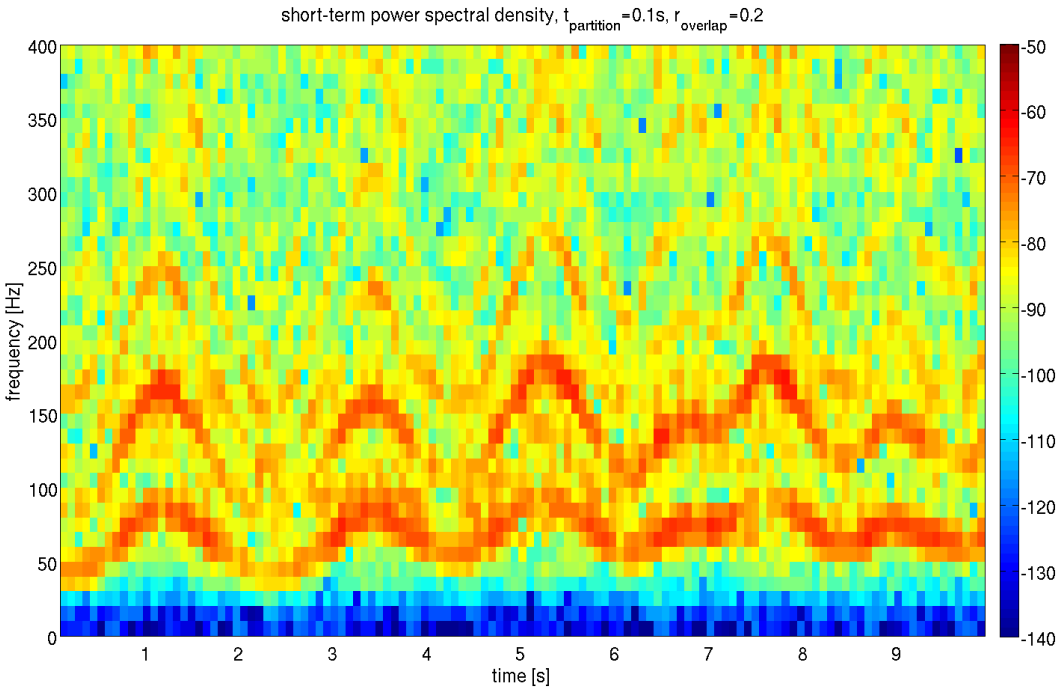


Figure 2-22: Short-term PSD at low frequencies of vehicle with playing RPM

Similar measurements as above were done with a Honda Civic Shuttle with a four cylinder piston engine. The results were similar, because the harmonics were also present and were also linked with the RPM. Thus the following conclusions are made:

1. The emitted vehicle signal is a wide-band signal
2. The first harmonic around 50Hz is (understandable) linked to the vehicle RPM

The vehicle measurements confirm that the main vehicles features lie in the low (below 400Hz) frequencies as concluded in previous research [11]. Further vehicle discussing and feature extraction will follow in Section 5.2.

2.3 Human walking pedestrian

This section will discuss measurements from a walking pedestrian at the Thales Delft parking area on 30 March 2010. The wind speed was 5 m/s and it rained a little.

The human measurement setup for a walking pedestrian is shown in Figure 2-23. There is one microphone position and at this position two microphones are placed. Microphone one at a height of 1 meter and the second microphone at 0.5 meter. The pedestrian walked at a distance d_1 of 1,2,4 and 6 meter. Three pedestrians were investigated, where the second and third was the same person but with other shoes. The sample frequency F_s was equal to 96kHz. The distance d_1 was measured without considering the microphone height.

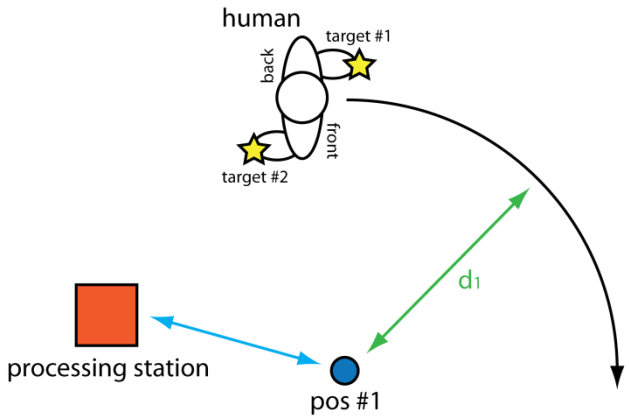


Figure 2-23: Measurement setup for recording human footstep

Figure 2-24 shows the result of the first walking pedestrian at one meter. The signal has a very high bandwidth, which is explainable by the fact that a hitting force on the ground produces a clap. A bit of sand was present on the stone surface and therefore a subtle sound was present due to sand friction. The signal power of the footsteps are not high, but at 4 meters it is even less as Figure 2-25 shows. Note the scale of the instantaneous power amplitude of the plots.

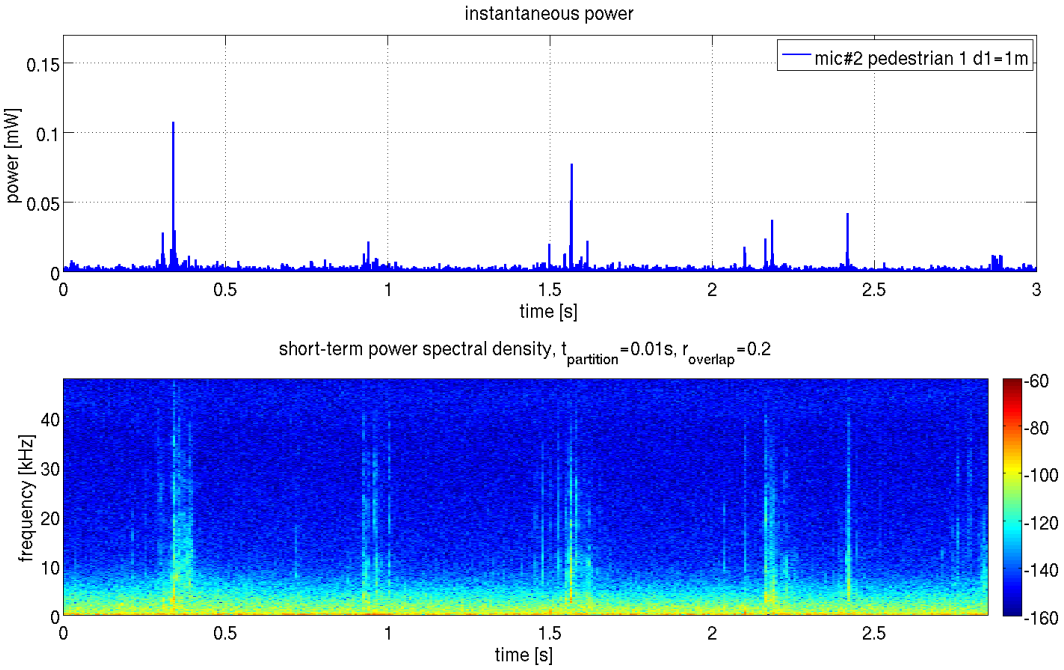


Figure 2-24: Instantaneous power and Short-term PSD of pedestrian 1 at d1=1m

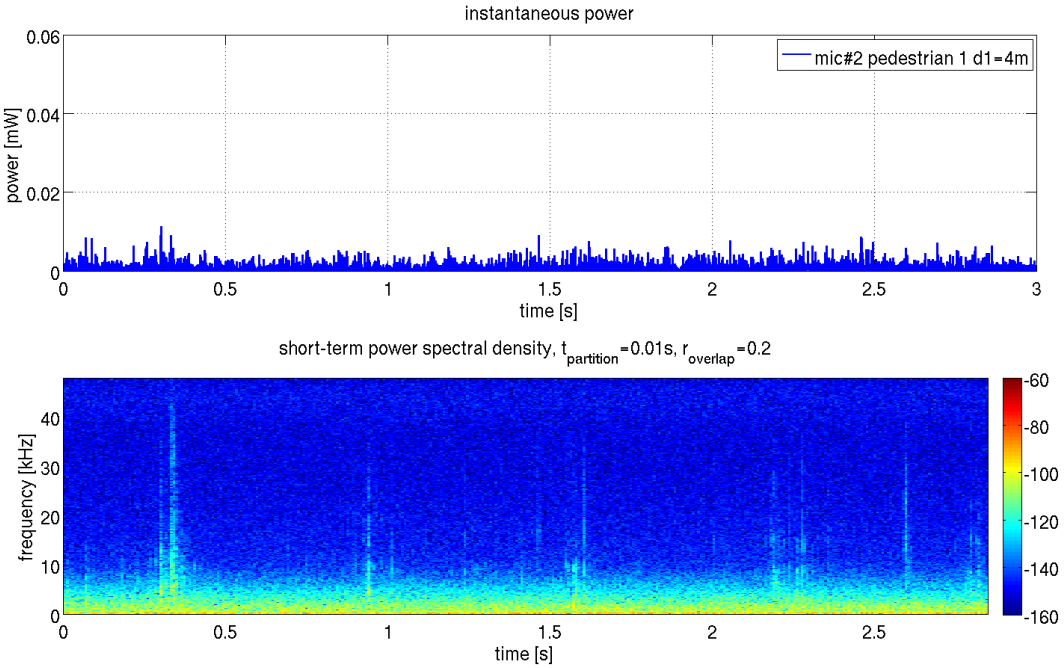


Figure 2-25: Instantaneous power and Short-term PSD of pedestrian 1 at d1=4m

Another walking pedestrian is shown in Figure 2-26. The person had other kind of shoes when compared with the first pedestrian, which resulted in a low SNR at already 1 meter. The third pedestrian walked with shoes which were especially selected for their high sound level for this

experiment. Figure 2-27 and Figure 2-28 shows that the SNR was high relative to the other situations.

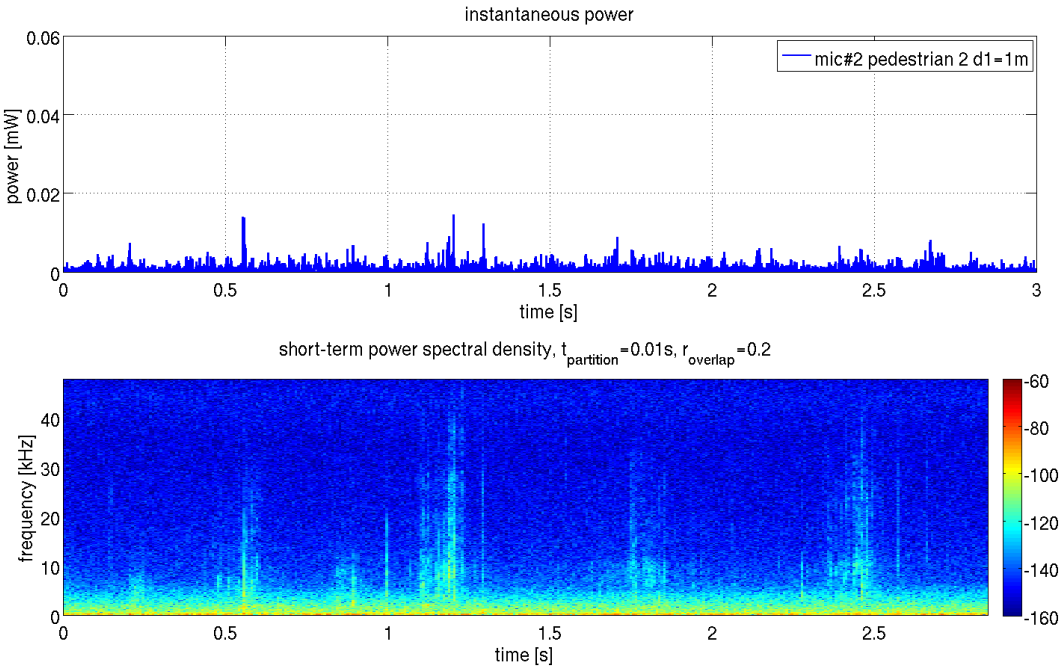


Figure 2-26: Instantaneous power and Short-term PSD of pedestrian 2 at d1=1m

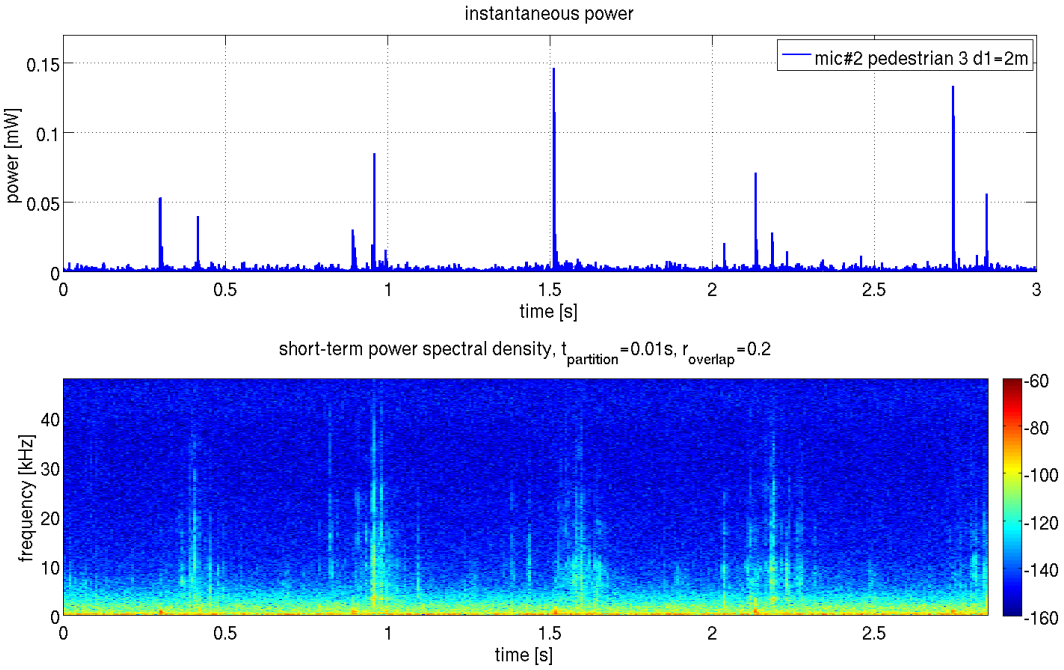


Figure 2-27: Instantaneous power and Short-term PSD of pedestrian 3 at d1=2m

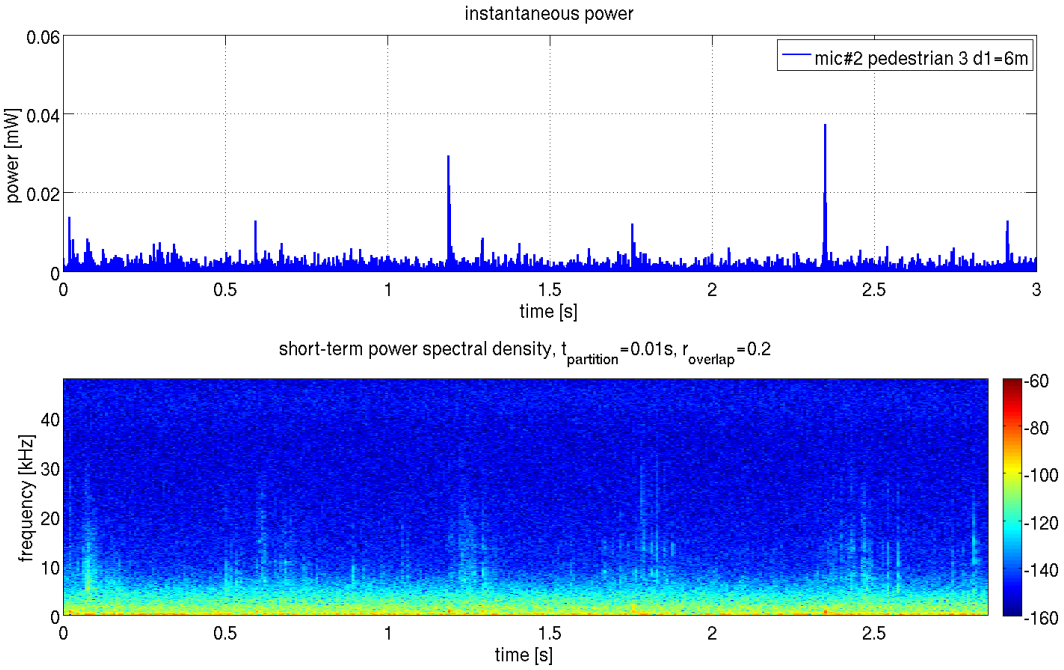


Figure 2-28: Instantaneous power and Short-term PSD of pedestrian 3 at d1=6m

As the previous result shows, the received signal is very dependent on the footwear and the received power is not very promising for footstep classification at these distances. For better insight into the signal the best ground, shoes and environment were selected for optimal measurement conditions. Figure 2-29 shows the third pedestrian walking indoor of the Thales Delft department in the SI Lab after the air-conditioning was shut down. This shows the best result which was accomplished.

The indoor signal gives some insight in the received signal. For every footstep two separate signals were received: the first from the heel bone and the second from the metatarsus (between the toes and mid-foot). Another conclusion is that the left and right footstep are not identical. Furthermore, a left or right footstep is not necessary identical to a previous left or right footstep.

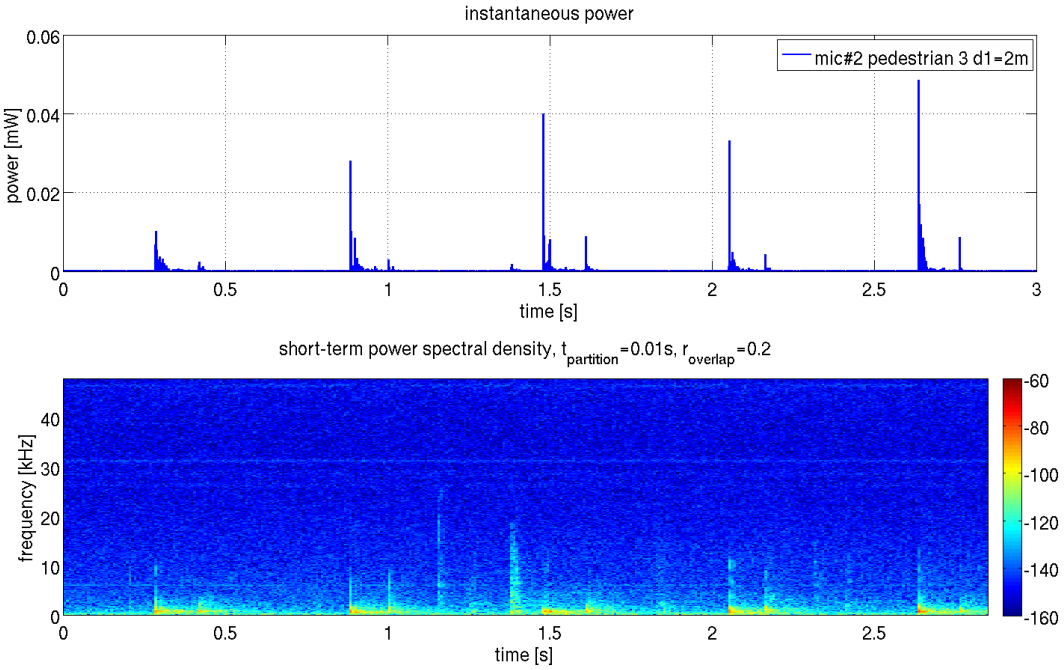


Figure 2-29: Instantaneous power and Short-term PSD of pedestrian 3 at d1=2m indoor

The pedestrian measurements results show that footsteps barely produce sound. The outdoor measurements provided an SNR between 5dB and 30dB at short distances (until 4 meter). The indoor measurements provide an SNR around 40dB in the low frequencies. The pedestrian footstep conclusion with these measurements are:

1. Signal depends on gait, shoes and ground
2. Received signal power is low and unpredictable
3. Striking force signal is the strongest and wide-band

The signal dependency on walking condition, footwear and floor is also stressed in previous research [13]. Further discussion and feature extraction for walking pedestrian will follow in Section 5.3.

3 System design

A good engineering approach is to start on a high level and descend slowly to the smaller sub-systems. In this chapter, an system framework on a high level is designed where the different tasks are outlined. The major challenge for this data processing is to cope with all kind of different signals at the input and to filter only the relevant information.
The Experimental ASN should localize and classify a target in the covered area using the recorded data. This means that the system has two tasks and therefore two separate sub-systems are needed. A third sub-system is constructed to process the raw recorded signal. The global overview of the system is shown in Figure 3-1.

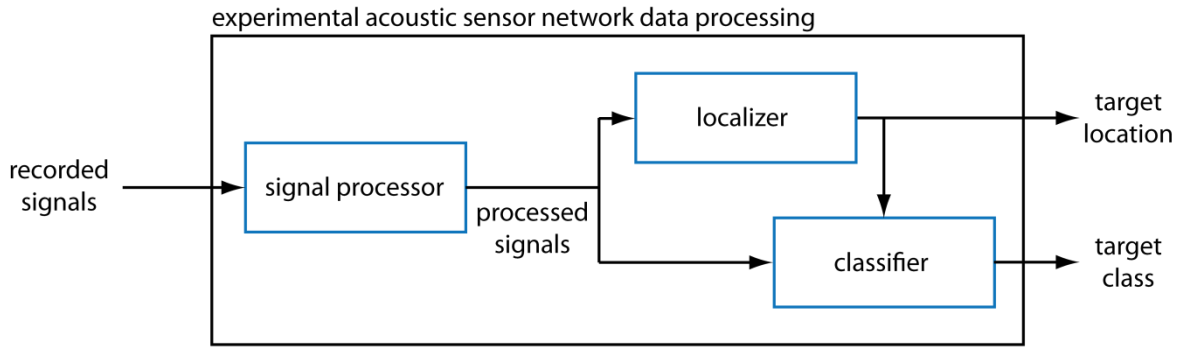


Figure 3-1: Experimental ASN system overview

First the *signal processor* (SP) will process the *recorded signals* into the *processed signals*. The *signal processor* will provide the *processed signals* in partitions which allows the *localizer* and *classifier* to process the partitions separately. The signal processing is taken apart, because it seemed that both the localizer and the classifier needed the same signal processing. The *localizer* will first estimate the *target location* and then the *classifier* will estimate the *target class*. Thus, the *processed signals* are first used to localize the target and after position estimation, the target is classified. This approach allows the classifier to use the *target location*. Another approach could be to first classify and then to localize. Than the target localization is improved at the cost of target classification performance, because a more specialized localization can be performed, but with a less specialized classification. Yet another approach could be localize, classify and then again localize, but this approach is becoming similar to tracking, which was excluded from the project. This system approach will also be more justifiable when the reader is studying the next chapters.

This chapter will discuss the initial localizer design is in Section 3.1. In Section 3.2 the classification initial design is discussed. The signal processor will be designed after the localizer and classifier in Chapter 6, because before the signal processor can be designed it must be clear what kind of signals are required by the localizer and classifier.

3.1 Localization

The localizer will estimate the target position and has three different components. First, the *information extraction* (IE) is trying to extract information for localization purposes from the processed signals. Second, the *propagation model* is the pre-knowledge of the system about how acoustic waves propagate. Third, with the model and the extracted information a *target location* can be estimated by the *location estimator*. The overview of the localizer is shown in Figure 3-2. The main focus of this project is the information extraction, but also on the propagation model.

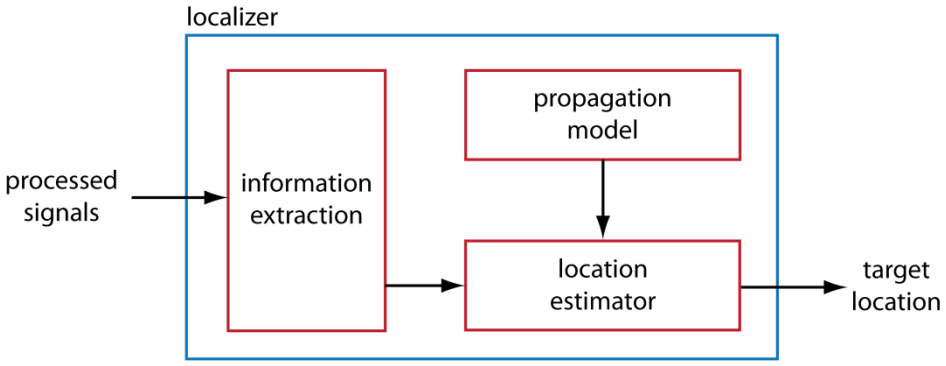


Figure 3-2: Localizer overview

Propagation model describes the movement of the acoustic wave and how the properties (such as energy) change over time and/or space. In other words, these models (equations) describe how acoustic waves normally propagate.

Information extraction should try to estimate the current properties of the wave. In other words, what is the current status of the acoustic wave at the different nodes? The IE process is actually depending on the propagation model, because the propagation model tells which gaps need to be filled so the *location estimator* can determine the target position.

There are of course different formats possible, but this mapping is done for its clarity. In Chapter 4 the *information extraction* and the *propagation model* will be further designed.

3.2 Classification

The classifier will estimate the target class and has a similar format as the localizer. The *feature extraction* (FE) discovers the current target features and the class knowledge provides the general class features. Another noticeable thing is that Figure 3-3 shows that a target location can be used at the FE. The *class estimator* will receive the features and class knowledge and decide which class is detected: *gun*, *vehicle*, *human* or *no target*.

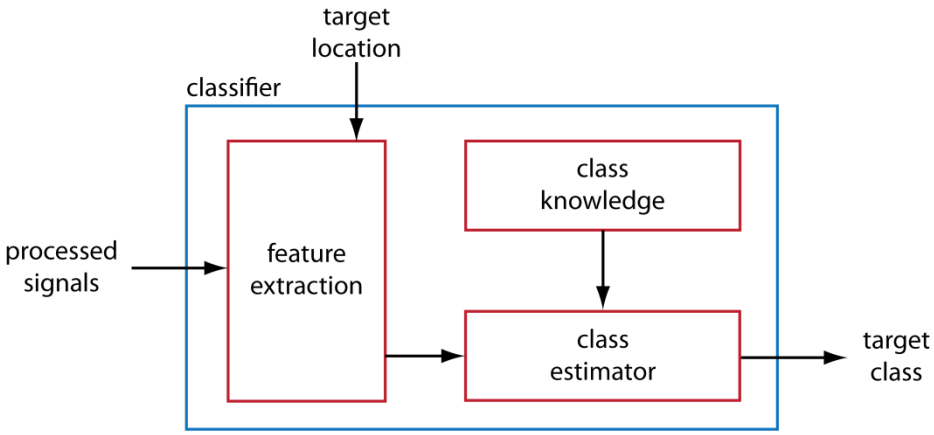


Figure 3-3: Classifier overview

Feature extraction is a process of making characteristic values of the signal and provide them in a feature vector. In other words, FE is the construction of an acoustic fingerprint or sound signature. The great challenge of feature extraction is to provide features which allow to discriminate between the classes. Thus, the features should have similarities if the target signal is from the same class.

Class knowledge is needed, for recognizing a target. How to gather this pre-knowledge has some overlap with the classification method choice. Target models could be used as class knowledge, but target modelling was excluded from this project. On the other hand, creating general target knowledge can have some overlap with modelling. Either way, it is an extensive challenge for the designer to provide solid pre-knowledge.

To clarify, the feature extraction is the main focus of this project. Chapter 5 will further design the *feature extraction* for the three targets classes.

4 Localizer design

Localization is the estimation of the target position. There are two approaches for localization with acoustics: based on propagation time and based on propagation loss. Normally the localization on time is more robust then localization on power. However, sometimes localization based on time is impossible, and therefore, localization based on power will be used for rough position estimation. Figure 4-1 shows the detailed overview of the localizer.

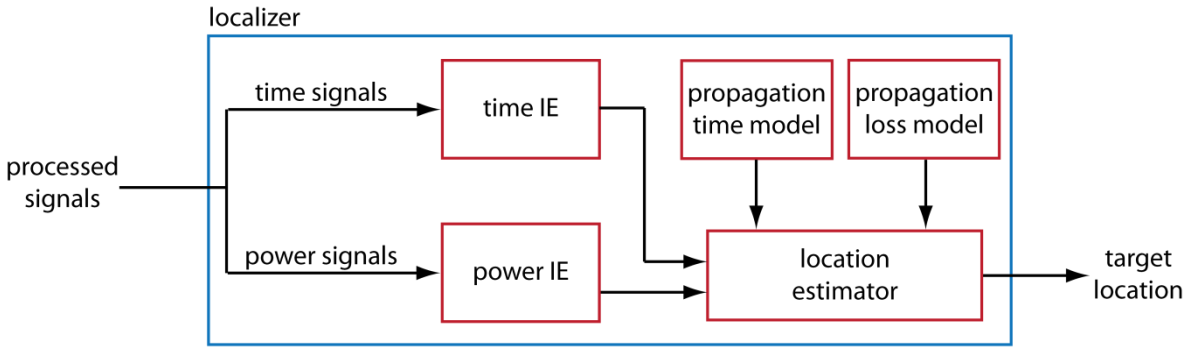


Figure 4-1: Localizer detailed overview

Two approaches are possible for position estimation, and therefore, two IE and two propagation models are needed. Also note that two different kinds of *processed signals*, which will be provided by the *signal processor*, are needed: *time signals* and *power signals*. The time and power based methods have their own processes. The last part of the localizer, the *location estimator*, will combine the different methods and estimate the target location.

The system will be based on Euclidean geometry and the observed distance between the target and node i at time t is given by:

$$d_i(t) = \left\| p_i(t) - p_t(t - T_{ti}) \right\| \qquad (1)$$

Where $p_t(t)$ and $p_i(t)$ are the positions of the target and node i respectively at time t and T_{ti} is the time duration of a signal to travel from the target to node i.

The *propagation time model* will be constructed in Section 4.1 and the *propagation loss model* will be constructed in Section 4.2. The *information extraction* for time and power-based localization will be designed in Sections 4.3 and 4.4 respectively. A theoretical discussion about time and power-based localization performance can be found in Section 7.3.

4.1 Propagation time model

By estimating the received time of the target signal at each node and using the propagation time model, the target position can be estimated. The speed of sound in air is mainly dependent on temperature and is approximately [5]:

$$c_{air} = 331.29 + 0.607 \cdot T_{air} \quad (2)$$

Where T_{air} is the ambient air temperature in °Celsius. More fundamental equations exist, but for this research the above formula is sufficient. When the temperature T_{air} is -20°C, +20°C or +40°C the approximately sound speed c_{air} is 319m/s, 343m/s or 356m/s respectively. Therefore, the Time of Arrival (TOA) at node i of the k^{th} target signal is written as:

$$T_{OAi}(k) = T_{OE}(k) + \frac{d_i(k)}{c_{air}} + T_{ei}(k) \quad (3)$$

Where $T_{OAi}(k)$ is the arrival time at node i of the k^{th} signal, $T_{OE}(k)$ is the time that the k^{th} signal was emitted, $d_i(k)$ is the target distance at Time of Emission (TOE), and T_{ei} is the time error due to modelling error and observation noise.

Due to the propagation time, the target signal will be received at different times at different nodes. In TOA localization the emitted and received times are known to calculate target distances, and with enough nodes to determine the target position. However, in the considered situation nothing is known about the emit time, and therefore no distance between the target and receiver can be calculated, like in TOA. TDOA solves this by, as the name implies, subtracting the received times from each other and therefore creating time difference information.

With TDOA good time measurements can be damaged. This happens when a good measurement is combined (subtract) with a bad measurement. Therefore, not the standard TDOA is used, but a TOA approach were the TOE is unknown. This result is a similar problem, but in comparison to TDOA it has one additional equation and one additional unknown. This approach has the advantage that no measurement is damaged by a bad measurement and that every measurement can be analyzed in a clean separate way. The situation of the known and unknown values and the assumptions are outlined in Table 1.

Known	Assume	Unknown
$T_{OAi}(k)$ $p_i(k)$	c_{air} is known and constant T_{ei} is zero mean and normal distributed	$T_{OE}(k)$ $p_t(k)$

Table 1: Known, unknown and assumption for time-based localization

For three dimensional localization a minimum of four equations with four $T_{OAi}(k)$ measurements are required. Note that the goal of the localizer is to estimate $p_t(t)$ and not $T_{OE}(k)$. A time offset in the $T_{OAi}(k)$ will not affect $p_t(t)$, because target position estimation is based on the time information relative to each other.

4.2 Propagation loss model

Target position estimation can be done by measuring the received power at each node and using the propagation loss [1] model. The power received at node i is written as:

$$P_{Ri}(f,t) = G_i(f) \cdot P_0(f, t - T_{ti}) \cdot A_\alpha(d_i(t), \alpha_a(f)) \cdot A(d_i(t)) + P_{ei}(f,t) \quad (4)$$

Where $P_{Ri}(f,t)$ is the received power at node i as a function of frequency and time, $P_0(f,t)$ is the received power at reference distance d_0 as a function of frequency and time. $G_i(f)$ is the microphone gain of the node i. $P_{ei}(f,t)$ is the power error, which contain effects of modelling error and observation noise. $A_\alpha(d, \alpha_a)$ is the sound absorption functions and $\alpha_a(f)$ is the air absorption coefficient. $A(d)$ is the attenuation function due to geometric spreading of sound and is independent of frequency:

$$A(d) = \left(\frac{d_0}{d}\right)^2 \quad (5)$$

Atmospheric absorption of sound is caused by friction losses in the transmission medium [23]. The result of sound attenuation $A_\alpha(d, \alpha_a)$ by atmospheric absorption is presented as a function of a distance d and the air absorption coefficient α_a in dB/m:

$$A_\alpha(d, \alpha_a) = 10^{-\left(\frac{\alpha_a \cdot (d - d_0)}{10}\right)} \quad (6)$$

The maximum value of absorption (in dB/m) in air at room temperature over all humidity for frequencies up to 50 kHz is given by [24]:

$$\alpha_a(f) = 3.3 \cdot 10^{-5} \cdot f \quad (7)$$

Higher frequencies are thus attenuated more than lower frequencies. More precise studies and plots of atmospheric absorption exist [25], but is considered not relevant for the current project investigation. Furthermore, for a good absorption estimation, professional equipment is needed to estimate the relative humanity and atmospheric pressure.

The transfer function $P_R(f)/P_0(f)$ with the $\alpha_a(f)$ of (7) and d_0 equal to one is given in Figure 4-2.

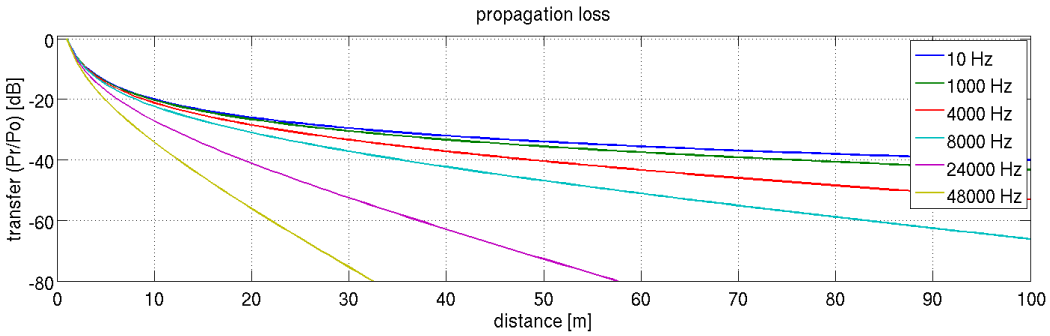


Figure 4-2: Propagation loss versus distance with different frequency

With the propagation loss formula it is known that every frequency component has a different absorption. Sometimes [19] the atmospheric absorption function is left out to reduce complexity. However, the use of different frequencies can result in better accuracy, because the absorption function can be included in the calculation, but also add in complexity. Furthermore, the use of multiple frequencies also allows, when multiple targets have different frequencies in their emitted signal, multi-target power-based localization.

The geometric spreading and atmospheric absorption are a function of distance, which can be used for localization. The situation of the known and unknown values are outlined in Table 2. The known values can be directly measured and the unknown have to be calculated using some assumptions.

Known	Assume	Unknown
$P_{Ri}(f,t)$ $p_i(t)$	$t - T_{ti} \approx t$ $P_{ei}(f,t)$ is zero mean and normal distributed $G_i(f)$ is known or equal over all microphones	$P_0(f,t)$ $p_t(t)$

Table 2: Known, unknown and assumption for power-based localization

It is required to assume that $t - T_{ti} \approx t$, because otherwise with each added $P_{ri}(f,t)$ a new unknown is added to the problem. It is assumed that T_{ti} is small in comparison to the time for a target to change its position.

If the dimension of $P_0(f,t)$ is one, and the dimension of $p_t(t)$ is three, then the system has to calculate four values. Thus, four equations and four $P_{ri}(f,t)$ measurements from four nodes are needed. When more frequency components are used, also more values have to be solved. However, more frequency components do not provide extra dimensions, but only better accuracy of the target distance. For example, for two nodes with three frequency components, the system has enough equations to "solve" the six unknowns, but this will not provide a proper target position because extra dimensions are needed.

Similar as with time-based localization, the localizer goal is to estimate $p_t(t)$ and not $P_0(f,t)$. Therefore, a constant in the microphone gain $G_i(f)$ or in the reference distance d_0 will not affect the localization performance.

4.3 Time information extraction

With the *time signals* provided by the signal processor, time information will be provided by the *time information extraction* to the *localizer estimator*. Thus, the Time IE will estimate T_{OAi}. The construction of the time signals by the time signal processor will be discussed in Section 6.1. The Time IE will first search for peaks, which indicate the time of arrival, in the time signal $TSP_{ip}[t]$ which is a function of time. An algorithm to search for peaks, which are local maxima, in a signal is given in Appendix B. Which peak to select, which will be used to estimate the T_{OAi}, is done by giving every peak a mark. The peak with the highest mark in partition p will be selected:

$$T_{OAi}(p) = t_{peak\_i}\left[\arg\left[\max_k[W_{peak\_i\_k}]\right]\right] \tag{8}$$

Where $t_{peak\_i}[k]$ is the time of peak k at node i and $W_{peak\_i\_k}$ is the mark of peak k at node i. Example weight functions are given in Appendix B. If the marking is done only proportional to the peak height, the maximum peak is selected. However, in Chapter 8, it will become clear that a more complex method is needed to select the right peaks. Furthermore, in Chapter 8 this Time IE process will be discussed further.

With threshold Tr_{TIE} it is decided if the selected peak is relevant:

$$T_{OAip\_relevant} = \begin{cases} 1 & \text{if} \quad y_{peak\_i}\left[\arg\left[\max_k[mark_{peak\_i\_k}]\right]\right] \geq Tr_{TIE} \\ 0 & \text{else} \end{cases} \tag{9}$$

Where y_{peak_i} is the height of the peak. The location estimator will only use relevant information. In other words, the target position estimation is only started when the extracted information is relevant and thus a correct target position can be expected.

4.4 Power information extraction

The *power information extraction* will use the *power signals* to provide power information to the *localizer estimator*. In other words, the Power IE is going to estimate $P_{Rip}[f]$. The power signal processor, which will be discussed in Section 6.2, will provide for every partition p and node i a power signal $PSP_{ip}[f]$ which is a function of frequency. The power IE will use $PSP_{ip}[f]$ to estimate $P_{Ri}[f, p]$:

$$P_{Ri}[f, p] = PSP_{ipa}[f_a] \tag{10}$$

Where $PSP_{ipa}[f_a]$ is the power aggregated version of $PSP_{ip}[f]$. The Power IE process can also decide to ignore certain frequencies. How the $P_{Ri}[f, p]$ is constructed has some freedom, because it is irrelevant what the estimated transmitted signal will be for the estimated location. In Section 6.2 three kinds of power signals are constructed and all three can be used at the Power IE.

The Power IE will also determine if the extracted information is relevant:

$$P_{Rip\_relevant} = \begin{cases} 1 & \text{if} \quad \sum_f W_{PIE}[f] \cdot PSP_{ip}[f] \geq Tr_{PIE} \\ 0 & \text{else} \end{cases} \tag{11}$$

Where Tr_{PIE} is the Power IE threshold. The decision can depend on the different frequencies if the weight function $W_{PIE}[f]$ is chosen non-uniform. Only when the extracted information is relevant and thus a correct target position can be expected, the target position estimation is started.

4.5 Location estimator

The *location estimator* will combine all the results and process it into one target position. It was assumed in Section 1.7.1 that at most one target is present in the covered area, thus there is no need for multiple target position estimation, which can be done parallel processing. The location estimator is specified in more detail in Figure 4-3.

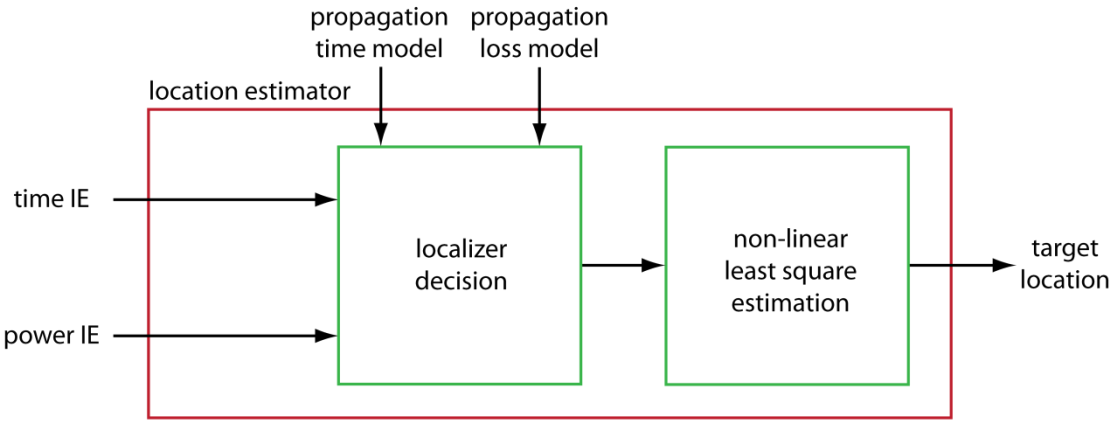


Figure 4-3: Location estimator

The first step is to decide which extracted information will be used for the position estimation by using the $P_{Rip\_relevant}$ and $T_{OAip\_relevant}$. If a stationary sound is emitted by the target, for example a running vehicle engine, there is no time information available (no clear signal begin and also much ambiguity) and the position must be determined with the power-based localization. When a non-stationary sound is emitted by the target, for example with a gunshot, the position can be estimated with time-based localization.

In the *localizer decision* also the time and power information can be weighted. In other words, it can be decided which measurements must be more or less dominant in the mathematical position estimation. After this, the *non-linear least square estimation* will calculate the target position which fits the best with the extracted information.

The time error for a certain received signal, using the propagation time formula given in Section 4.1, is given by:

$$T_{ei}(d_i) = T_{OAi} - T_{OE} - \frac{d_i}{c_{air}} \tag{12}$$

For every node measurement with time based localization, the non-linear least square estimator will do the following:

$$\arg\min_{[\widehat{T_{OE}}\,;\,\hat{d}_i]} \left\| W_{LE\_T}(T_{OAi}) \cdot T_{ei}(d_i) \right\|^2 \tag{13}$$

Thus, with time-based localization the $[\widehat{T_{OE}}\,;\,\hat{d}_i]$ is estimated which fits with the extracted information. With the estimated target distance $\hat{d}_i$ which gives the minimum error, the estimated target position $\hat{p}_t$ can be calculated.

With the propagation loss formula of Section 4.2 the power error for a certain fixed time moment is given by:

$$P_{ei}(f, d_i) = P_{R_i}(f) - G_i(f) \cdot P_0(f) \cdot A_\alpha(d_i, \alpha_a(f)) \cdot A(d_i) \qquad (14)$$

The non-linear solving estimator will do the following for every node measurement and every frequency with power-based localization:

$$\arg\min_{[\widehat{P_0(f)}\,;\,\hat{d}_i]} \left\| W_{LE\_P}(P_{ip}) \cdot W_{LE\_f}(f) \cdot P_{ei}(f, d_i) \right\|^2 \qquad (15)$$

Thus the solver will find the argument $[\widehat{P_0(f)}\,;\,\hat{d}_i]$ which will minimize the squared error. As already discussed: although the goal of the least square solver is to find $\widehat{T_{OE}}$ and $\widehat{P_0(f)}$, this is not the goal of the localizer.

The weights, which are included in (13) and (15), have effect, because the least square finder will search for the overall least error. In other words, equations with higher weights using the same error $P_{ei}(f, d_i)$ value, provide a higher equation value and thus the least square finder will better minimize the $P_{ei}(f, d_i)$ of equations with higher weights. Time-based localization can be weighted according to the time of arrival and power-based according to received power and frequency.

As Section 4.1 and 4.2 discussed, for both power and time-based localization at least four measurements are needed. In other words, four $P_{Rip}[f]$ or four $T_{OAi}[k]$ are needed for three-dimensional localizing. Of course the use of more measurements is possible or a combination of time and power measurements can be made. This can, if the weights are chosen properly, result in a higher position accuracy and/or integrity.

5 Classifier design

Classifying is the estimation of the class of the target. Three target classes are chosen for the current EASN in Section 1.7: gun, vehicle and human. Three targets with three different associated signals: running piston engine, walking pedestrian and muzzle blast. All targets require their own feature extraction, because the targets produce very different sounds. The detailed overview of the classifier is shown in Figure 5-1.

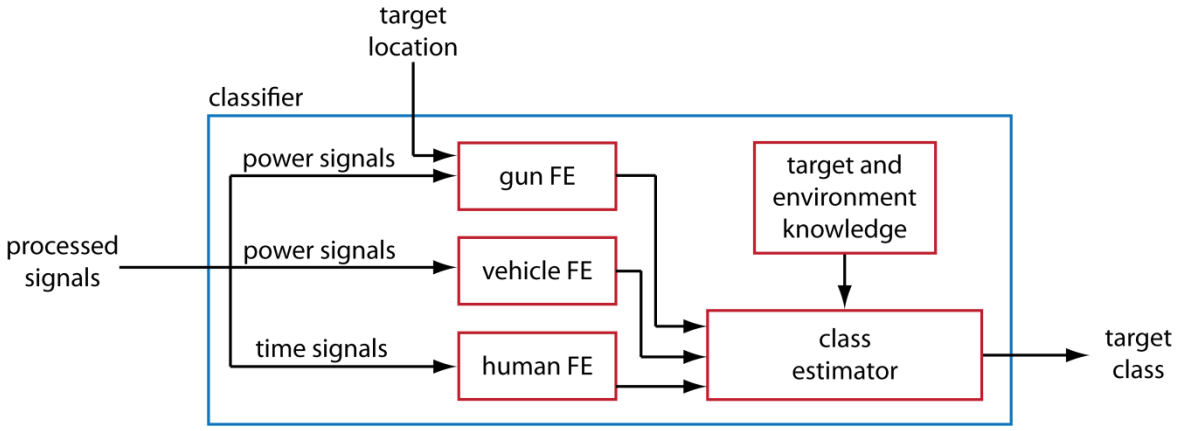


Figure 5-1: Classifier detailed overview

The *feature extraction* processes will use different processed signals. The Vehicle FE and the Gun FE will use the *power* signals to provide features. Only the Human FE will use time signals from the SP. The Gun FE will also use the target location to correctly extract gun features. The mapping of this classifier design will be more justified, when the feature extraction sections are studied.

The feature extraction of the signal for guns, vehicles and humans will be outlined in Sections 5.1, 5.2 and 5.3 respectively. A theoretical discussion about feature extraction and classification performance can be found in Section 7.4.

5.1 Gun feature extraction

Muzzle blast characteristics need to be extracted to recognize a gun. Guns emit signals with much energy due to an explosion, but the explosion sound differs over the different guns which is discussed in Section 2.1. Gun characteristics will be extracted by using the power signals and the target position.

The aggregated power spectrum of the received signal at one meter will be used as a characteristic gun feature. Due to the atmosphere and target distance the signal is attenuated and therefore the atmosphere can be considered as a channel which has filtered the signal. When the target distance is known, a channel estimation can be constructed. The channel and the received power spectrum can be used to estimated the emitted power spectrum. The channel will be constructed with the propagation loss equation as discussed in Section 4.2. With this equation the channel is given by:

$$P_{R_i}(f,t)/P_0(f,t-T_{ti}) = G_i(f)\cdot A_\alpha(d_i(t),\alpha_a(f))\cdot A(d_i(t)) + P_{ei}(f,t) \quad (16)$$

For the channel compensation method $P_{ei}(f,t)$ is not needed and will be further assumed zero. The system can use $P_{R_i}(f,t)$ to estimate the received power at d_0 equal to one meter:

$$P_0(f,t-T_{ti}) = P_{Ri}(f,t)/[G_i(f)\cdot A_\alpha(d_i(t),\alpha_a(f))\cdot A(d_i(t))] \quad (17)$$

As the formula shows, the system needs to know the target distance for channel compensation. The target distance can be estimated in a previous process stage. This $P_0(f,t-T_{ti})$, which is the received spectrum at one meter, will be the gun feature (at node i with partition p received from the signal processor):

$$F_{gun\_i}[p] = [P_0(f,t-T_{ti})] \quad (18)$$

The length of $F_{gun\_i}[p]$ depends on $P_0(f,t-T_{ti})$, however to decrease the length, power aggregation can be applied or certain frequencies can be filtered. Power aggregation is the power summation of multiple bins and place the result in a wider new bin.

The fundamental problem with this approach is: the system does not know which frequencies are emitted, and therefore, it does not know which frequencies it has to compensate. It would be undesirable to amplify frequency components which are not emitted by the target. This can be solved by adding a threshold. However, a big advantage with channel compensation is that the transmitted power spectrum can be used as a feature for classification. This spectrum can show if the emitted power is high enough and possibly if the signal is wide-band.

Unfortunately, this approach does not allow gun feature extraction when no target position is estimated. However, if this is the case, is not very relevant, because normally a gunshot has so much energy that a target position should have been determined.

5.2 Vehicle feature extraction

Vehicle feature extraction is a challenging task, because of the variety of vehicles. Mostly a vehicle is a grumbling and vibrating object due to a running engine. The Vehicle FE process will use the power signals received from the Power SP.

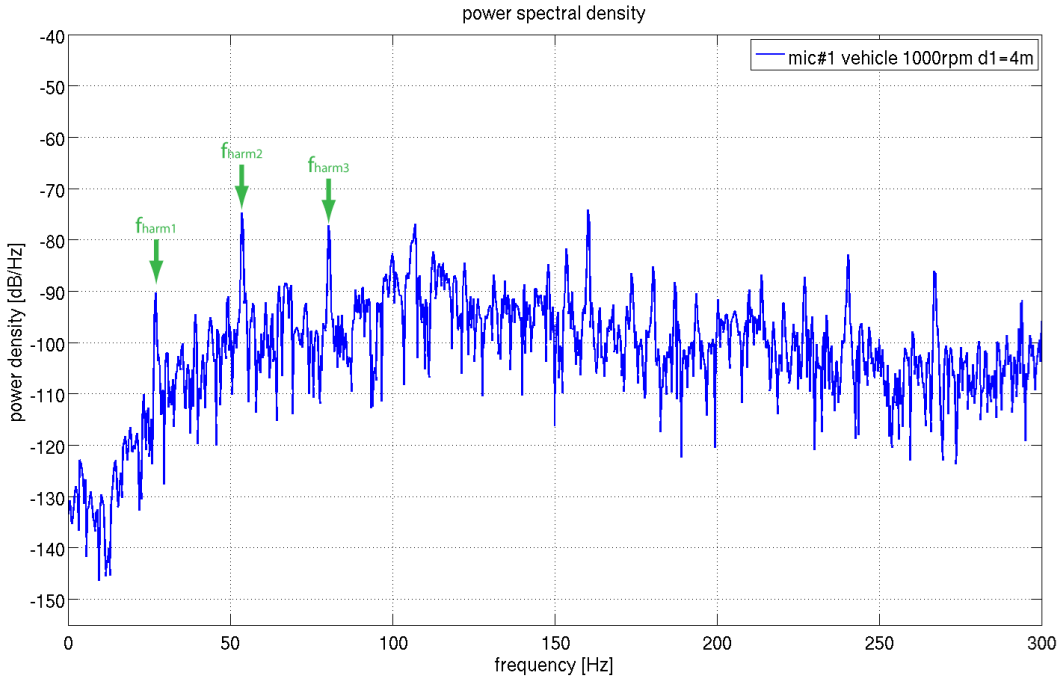


Figure 5-2: Example of the first three harmonics of the received vehicle signal

The fundamental frequency, which is linked to the engine speed as discussed in Section 2.2, is the strongest feature of a running piston engine. However, only a peak in the spectrum which can lie in the range of roughly 20Hz to 150Hz is just enough for an indication. Normally a driver is also changing the RPM. The derivative of the fundamental frequency will be the second feature. Features which appoint the coherence/constellation of the multiple (spectrum) peaks can also be used. Figure 5-2 gives an example of the selection of the first three harmonics. The above three vehicle properties will be present in the vehicle feature vector and thus, the experimental vehicle feature vector will contain these four values:

$$F_{vehicle\_i}[p] = \begin{bmatrix} f_{harm1}[p] \\ |f_{harm1}[p] - f_{harm1}[p-n]| \\ f_{harm2}[p]/f_{harm1}[p] \\ f_{harm3}[p]/f_{harm1}[p] \end{bmatrix} \quad (19)$$

The feature vector mostly contains frequency information and the derivative of the fundamental frequency with a positive integer n. When the last two feature values are not equal to two and three respectively, this can indicate that the found spectrum peaks may not be harmonics. The classification method can decide which features to use.

5.3 Human feature extraction

With the knowledge of Section 2.3 it seems very difficult to extract pedestrian features. Human feature extraction with different shoes, ground and gait allows only detection of walking, thus multiple footsteps. The variety of sounds created by the footstep force or friction of the pedestrian is too large for one-footstep recognition. Walking detection is similar to the following problem: how to recognize a clock, when only one clock tick is heard? The human feature extraction below has some overlap with tracking, but there is no other way for robust pedestrian feature extraction. The Human FE will use the time signals received from the Time SP.

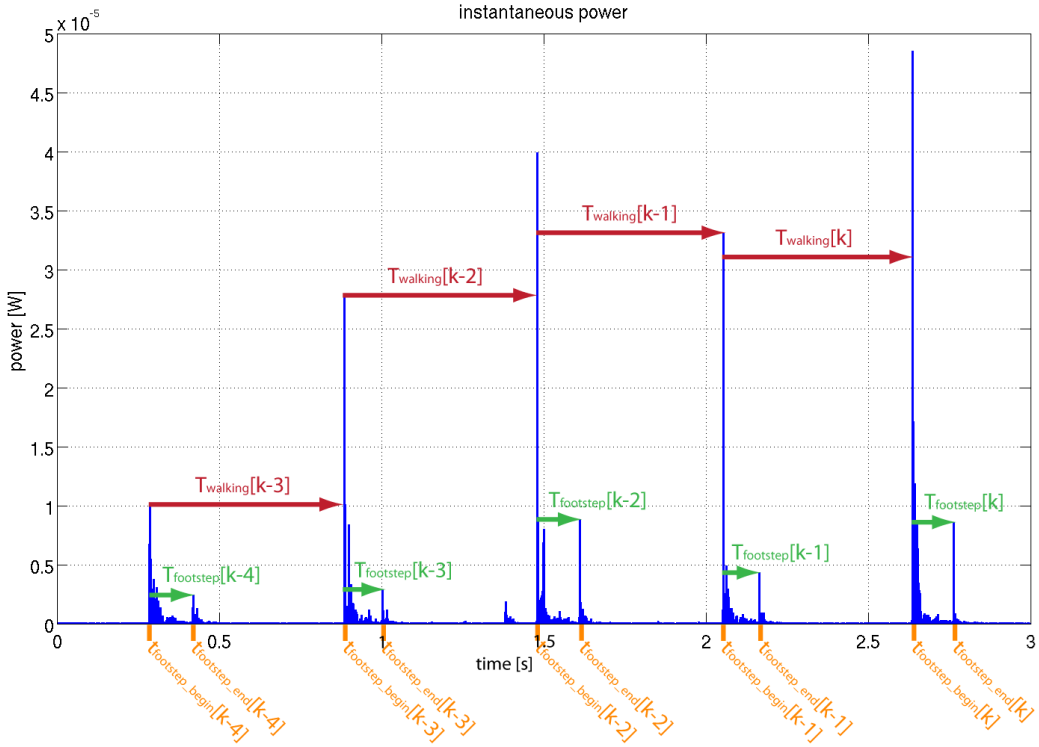


Figure 5-3: Example of time difference of a walking pedestrian

Figure 5-3 illustrates the features. The footstep and walking durations should have a certain duration and continuity. The begin time $t_{footstep\_begin}[k]$ and the end time $t_{footstep\_end}[k]$ can be used to estimate the footstep time:

$$T_{footstep}[k] = t_{footstep\_end}[k] - t_{footstep\_begin}[k] \quad (20)$$

The walking duration is given by:

$$T_{walking}[k] = t_{footstep\_begin}[k] - t_{footstep\_begin}[k-1] \quad (21)$$

For every last three footsteps a feature vector with six values will be constructed:

$$F_{human\_i}[p] = \begin{bmatrix} T_{walking}[k_{last}] \\ T_{walking}[k_{last}-1] \\ \frac{T_{walking}[k_{last}-1]}{T_{walking}[k_{last}]} \\ T_{footstep}[k_{last}] \\ T_{footstep}[k_{last}-1] \\ T_{footstep}[k_{last}-2] \end{bmatrix} \quad (22)$$

Where k_{last} is the last footstep number. The walking durations are the first two features, the continuity is the third feature. The first two features are very dependent on the pedestrian gait: for example, fast or slow walking pedestrian (e.g. running). The third feature is less dependent on the gait if the pedestrian is walking continuously. The last three features will be the three footstep times. However, the second peak of a footstep will not always be present, because this depend on the pedestrian gait. When a second footstep peak is not detected, the footstep length will be set to zero. However, when a second footstep peak is detected this can increase the certainty that a pedestrian is present.

It could be argued to use the walking pedestrian footstep locations for classification. Multiple footsteps positions should not be separated too much from each other. This is a respectful idea, but classification is then becoming very dependent on the localization performance. For the current implementation stage, it is preferred that the human classification is not dependent on the human localization performance.

5.4 Class estimator

The *class estimator* should decide what the target class is: gun, vehicle, human or unknown. The class estimator is outlined in Figure 5-4.

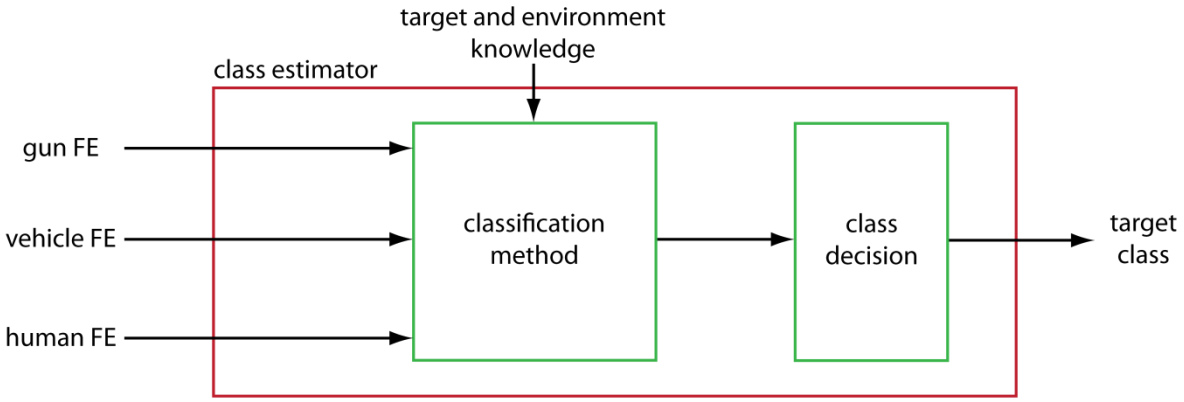


Figure 5-4: Class estimator

The class estimator will give the feature vectors to the classification method. Which classification method to choose or how to design such a classification method is considered beyond the scope of this thesis. However, it is certain that the parameters of the classification method depend on the targets and environment knowledge. Some classification methods are discussed in Appendix E. A classification method could provide four values, because there are four classes and these values should depend on the target recognition. The *class decision* will then receive the four values and will select the class with the best value (probably the highest).

6 Signal processor design

The signal processor has to process the raw recorded signal, and do it in such a way that the localizer and classifier can use the result. The signal processor has to cope with finite signals at the input and it should provide partitions ready to be analyzed at the output.

Three processes will be designed: *partition*, *power signal processor* and *time signal processor*. The recorded signal is first divided into multiple partitions by the *partition*. The experimental implementation of the partitioning is given in Section 7.1. The Time SP is designed to provide *time signals* which will have a high time accuracy. The goal of the Power SP is to provide *power signals* which represent the received power. The signal processor is outlined in Figure 6-1.

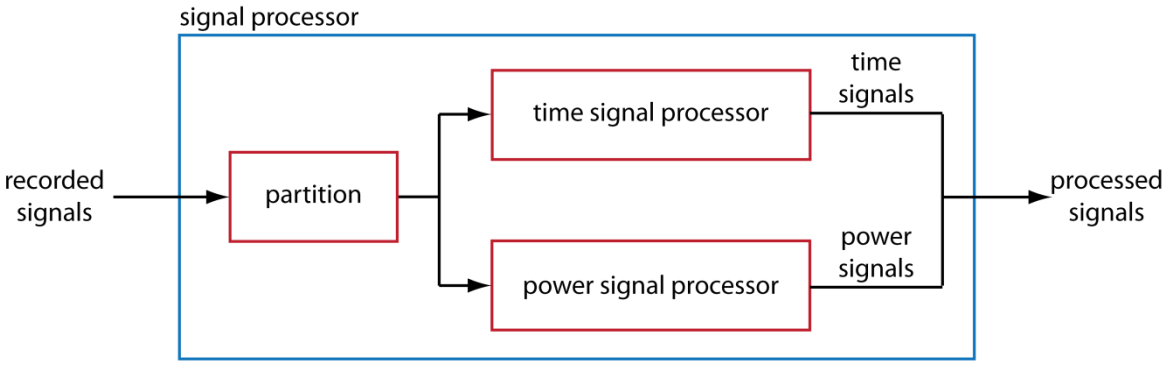


Figure 6-1: Signal processor overview

Wavelet analysis techniques can be used in an ASN. Although some papers [1], [13] design their ASN with wavelet techniques, it is still unclear why wavelets would perform better then Fourier. Wavelet analysis, which is discussed in Appendix C, is a relatively new field of study and much less used than the Fourier analysis. However, it has some advantages over the Fourier Transform, for instance the variable time and frequency resolution. This property seems to be very effective for certain applications, for example compression or filtering. However, De Groot Fourier Transform, which is discussed in Appendix D, provides with the FFT also variable time and frequency resolution. Wavelet analysis does not have extra potential for this project in comparison to Fourier Analysis: it is a method for very specific applications, but is has not the potential to extract extra general target characteristics. Maybe in the future wavelet techniques have the potential to be implemented in a ASN system, but it would be as an addition to the Fourier Transform. Therefore, it is decided that wavelet analysis is not very promising for this EASN.

At a certain node i, the microphone will measure a voltage which is proportional to the sound pressure with a certain sample time duration T_s and sample frequency F_s. The recorded signal $y_i[t]$, which is a finite discrete signal and a function of system time t, will be processed by the SP. The recorded signal $y_i[t]$ is first partitioned in multiple $y_{ip}[t]$ by the *partition*. The Time SP and Power SP will be designed and discussed in Section 6.1 and 6.2 respectively.

6.1 Time signal processor

The Time SP should provide a signal which has a very high time resolution, because the position accuracy depends on the time accuracy for the time-based localization. Although the time accuracy is important, the time offset is less important, because a time offset will not result in a different target location $p_t[t]$, but only in a different time of emission $T_{OE}[k]$. The Time SP will provide three outputs and they have in common that they provide partitions with the same length.

The first output $TSP_{ip1}[t]$ is the unedited version of $y_{ip}[t]$. The second output $TSP_{ip2}[t]$ is created by first calculating the instantaneous power of $y_{ip}[t]$ and next by filtering it with an edge mask:

$$\text{edge\_mask} = [-1 \quad -1 \quad \cdots \quad -1 \quad +1 \quad \cdots \quad +1 \quad +1] \tag{23}$$

Filtering is the convolution or cross-correlation of an inverted or non-inverted mask with the signal $y_{ip}[t]$. This selected mask will make the "real" edges in the instantaneous power signal better visible and will increase the robustness. This second output will be used for human feature extraction, but it can also be used for time-based localization if the third approach, which will be discussed below, does not work for certain target signals.

It would be better to use a filter which is based on the original signal to get a high time accuracy, but no such matched-filter can be constructed because nothing is known of the transmitted signal. Therefore an "on-the-run" mask is determined for every partition in one of the signals to filter all the signals. Thus, a kind of pseudo-matched-filter is created and then the third output $TSP_{ip3}[t]$ can be determined. How to choose the right mask is a challenging task. For instance how large the mask should be depends on the length of the target signal, but also which part of the signal should be selected is challenging.

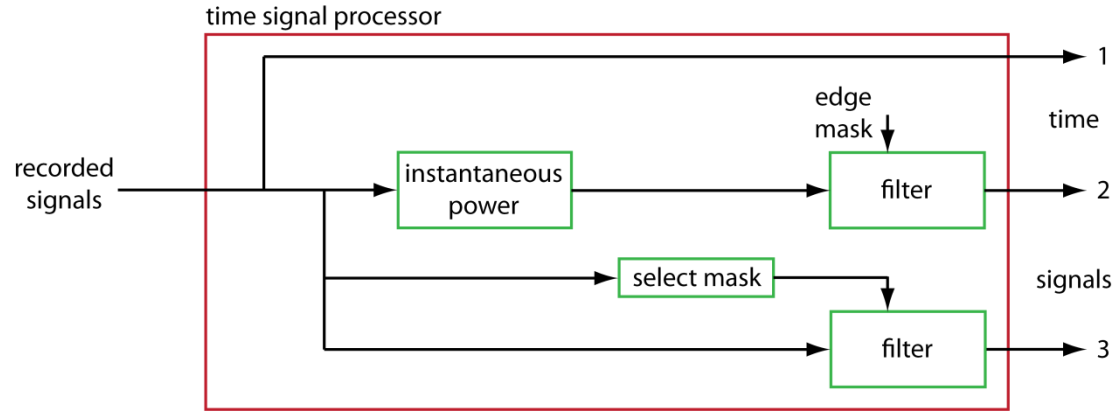


Figure 6-2: Time signal processing

Time SP with the three outputs is shown in Figure 6-2. In Section 8.1 appropriate solutions are suggested to select the correct masks.

6.2 Power signal processor

The Power SP will provide three outputs: original power spectrum, microphone gain compensated power spectrum and microphone gain and environment compensated power spectrum. How to process one of the partitioned signals $y_{ip}[t]$ to these spectra will be discussed in this section.

The first output $PSP_{ip1}[f]$ is the power spectrum and is calculated by windowing the signal $y_{ip}[t]$ to $y_{ipw}[t]$ (to avoid spectral leakage) and then calculating the mean square power spectrum:

$$PSP_{ip1}[f] = \left| \frac{FFT\{y_{ipw}[t]\}}{L_{ipw}} \right|^2 \quad (24)$$

Where L_{ipw} is the length of signal $y_{ipw}[t]$. The second output is the first output, but with microphone gain $G_i[f]$ compensation:

$$PSP_{ip2}[f] = \frac{PSP_{ip1}[f]}{G_i[f]} \quad (25)$$

It can be that although no targets are present in the covered area, still some power will be received. To compensate for this, the third output $PSP_{ip3}[f]$ will have environment compensation:

$$PSP_{ip3}[f] = |PSP_{ip2}[f] - ES_i[f]| \quad (26)$$

Where $ES_i[f]$ is the (estimated) environment spectrum at node i and is created in such a way that the length is equal to $PS_{ipw}[f]$. $ES_i[f]$ can be the power spectrum of a recorded signal at the calibration process. Crucial is that at the moment of recording the covered area should not contain any (defined) targets. $PSP_{ip3}[f]$ will show the power change relatively to the environment spectrum.

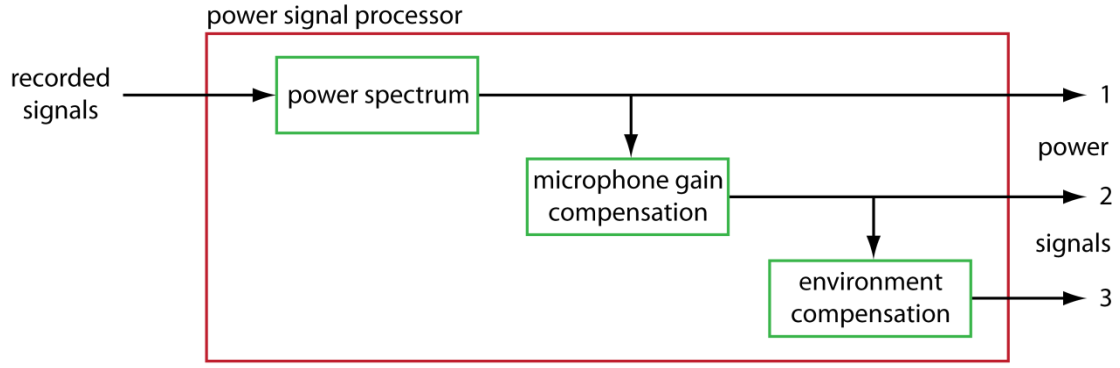


Figure 6-3: Power signal processing

Thus, the Power SP provides three outputs, as is illustrated in Figure 6-3. The above double-sided spectra can also be made single-sided. If all the microphone gains are equal and the $ES_{iw}[f]$ is zero, the power-based localization performance will not be effected by which output is chosen. But if there is a difference, the performance may be increased.

7 System implementation

The designed components will be combined into one experimental system. The experimental hardware and MATLAB will be used for the proof of concept implementation. Figure 7-1 shows the design overview after all the designed components are combined.

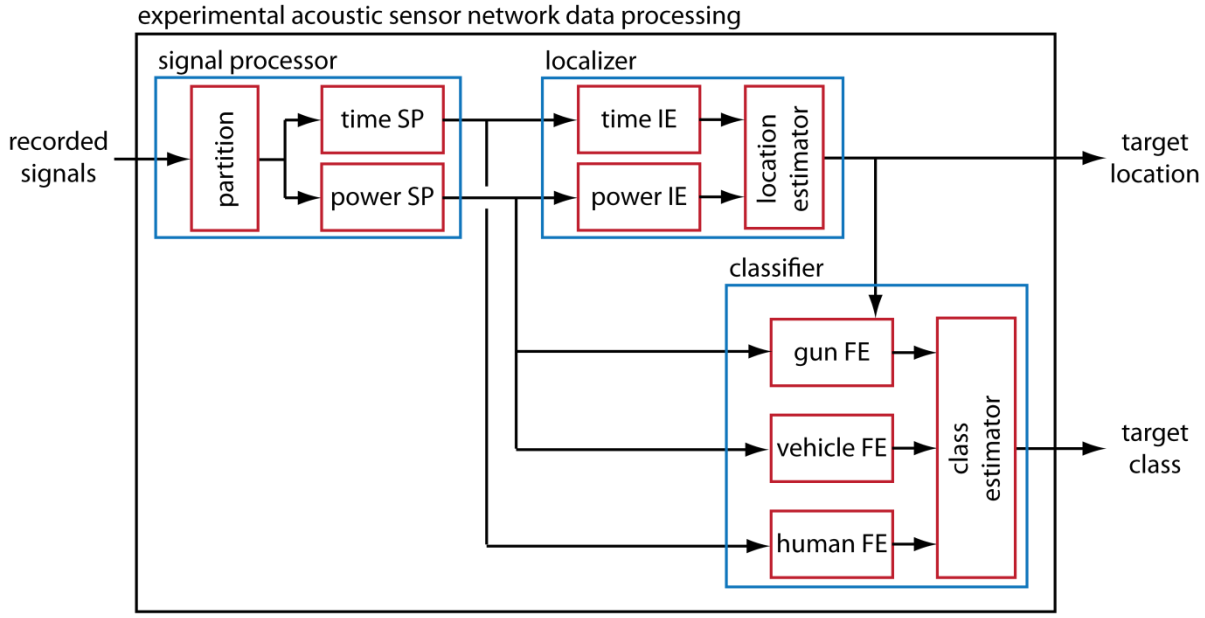


Figure 7-1: Final system overview

The system will be implemented as a single task system for experimental reasons. This means that the data processing only has one processor unit (one core) available. The single task process is discussed in Section 7.1. Section 7.2 further discusses the system choices for experimental purposes. Next, it is very interesting to discuss the system performance before it is evaluated with measured data in Chapter 8. Therefore, the theoretical localization and classification performance are discussed in Section 7.3 and 7.4 respectively.

7.1 Single task system

Because the experimental system will be implemented in MATLAB without any multitasking, the system also has to be outlined and with succeeding processing.

The time and power signal processors will provide a result for every partition. The input signals will not be infinite and thus no continued calculation can be performed by the SP. Therefore the SP blocks will process the received signals to the end and save the last part of the signal which could not be processed due the length of the recorded signal. This saved part will be used at the beginning when new recorded signals have arrived. Figure 7-2 illustrates this approach for a single signal.

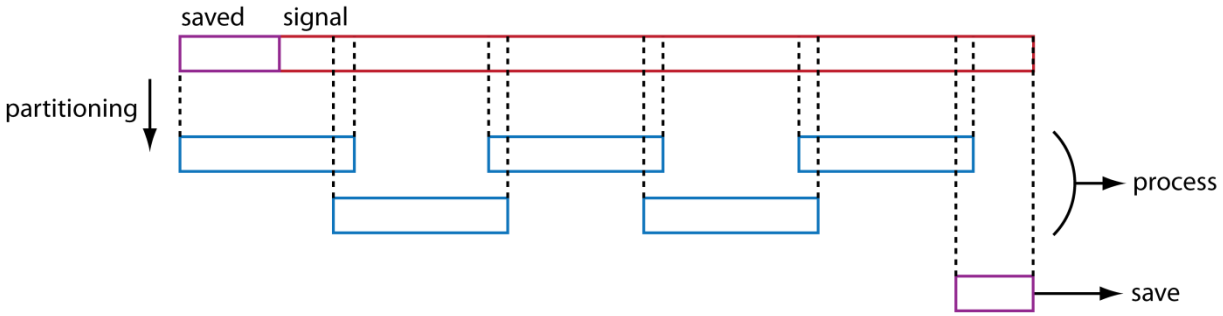


Figure 7-2: Partitioning example in the signal processor for a single signal

The partitions can have a certain length and a certain overlap. How large the partitions should be and with how much overlap depends on the target, the operational situation and the system demands. Large partitions can be chosen to benefit from signal integration, but then the target should not change his signal or position fast. An overlap in the partitioning can be chosen to cope with two problems. First, for windowing at the spectrum calculation. Second and more importantly, for different target distances which will result in different time of arrivals.

After the signal processing, the Time IE and Power IE can perform their tasks. The result of this process will be used by the location estimator. If no location is calculated, the Vehicle FE and the Human FE process are started and the output of the Gun FE is set to zero. Note that the Human FE and the Vehicle FE require memory to save respectively the previous footsteps and the previous fundamental frequencies. If a target location is found, the Gun FE process will also be started. If the Gun FE is not started due to no target location, is not very relevant, because normally a gunshot has so much energy that a target position should have been determined. After the features are extracted the class estimator will determine the target class. After the classification, the IE will process the next partition, which was delivered by the SP blocks.

When all the partitions from the SP are processed, the SP blocks will wait until new recorded signals are available and together with the saved recordings the process will start totally over.

7.2 Experimental choices

To implement a complete system which can acoustically localize and classify targets, requires certain choices to reduce the workload and to narrow the scope of this project. Although the decisions of the topics can be delayed after further research with modelling and simulations, eventually choices are needed for the EASN. The result of this first implementation will indicate if it is relevant to invest more time for ASN research. The first choice was that it had to be implemented in MATLAB which is a good application for testing the system. This section will further explain some other experimental choices.

partitioning
For outlining the system, first the partition time has to be chosen. This can be done freely by the designer, but with the knowledge of the targets a time partition of 0.5 second with an overlap ratio of 0.1 (i.e. 10%) is chosen. The overlap ratio indicates how much a partition overlaps with the previous partition.

environment compensation
The Power SP needs an environment spectrum estimation to compensate for acoustic noise background. For experimental purposes, the environment will be estimated using a 30 seconds recording from all nodes in the calibration process without any targets present in the area. For every node, the recording will be partitioned and the average spectrum for every node will be used as the environment spectrum.

location estimator
If no time information can be extracted but enough power is received, position estimation will be done with power-based localization. However, if time information is available (for example with a gunshot signal), time-based localization will be used. The decision if a target position is found depends on the squared error, which is explained in Section 4.5. If the error is too large, it is probably a false target position and therefore it is decided to ignore the position.

non-linear least square estimation
For solving the mathematical localization problem the iterative least square solver algorithm of MATLAB's Optimization Toolbox (*lsqnonlin* function) is used. This method can be used to find the parameters for the (local) minimum squared error.

feature extraction
Multiple node signals were needed for localization, but for classification a different approach is required. The EASN should provide one estimated class for every set of partitions. Therefore two solutions are suggested.
In Human FE the node which received most of the energies is selected to extract the time feature information. When the node selection switches many times, this can also result in a small error, because different microphones have a different distance to the target and thus different time of arrival. However, this error is very small, because a distance of ten meters correspondent to a travel time of approximately 0.03 seconds. A similar problem was already present because of the different footstep positions.

The Gun FE and the Vehicle FE will perform the feature extraction on every received signal, but will combine the results of the different nodes into one result using weights. The Gun FE will also aggregate the powers into less bins, in such a way that only six values will remain, to reduce the classification computation. Thus with a sampling frequency of 96kHz, frequency bands of 8kHz will be used.

classification method
The features can be used in an advanced classification method and some are discussed in Appendix E. For example, a Gaussian Mixture Model or a Neural Network can be used for classification. With a Gaussian Mixture Model it must be assumed that the features are Gaussian distributed. Neural Networks can also be used for classification, but they provide little insight. For a good insight of the features performance, a Classification Tree will be used.

peak finding
In certain processes it is required to find peaks in time or power signals. The used method is discussed in Appendix B.

weighting
In some design components weights are required. For experimental reasons, the weights are chosen by the implementer. This means that in most cases the weights are uniform or linear as a function of power and/or time. Three weighting example functions are given in Appendix B.

7.3 Theoretical localization performance

To provide a precise estimation of the theoretical localization performance is a very challenging task and probably therefore avoided in current research papers. For the current project stage it is too complex to provide a theoretical performance analysis, but this section will briefly discuss the localization performance dependency.

Two different localization methods were designed, both dependent on:

- Transmitted signal
- Propagation model
- Target distance
- Environment and ambient air
- Node position accuracy
- Hardware specifications
- Mathematical solving method

The transmitted signal is the most crucial, because the target location accuracy depends on the SNR measured by the microphone. The SNR is also dependent on the used hardware and operational situation. Due to the distance and the environment the signal can be modified, which can result in lower localization performance. The accurate knowledge of the node positions are crucial, because position estimation is relative to the node positions. The propagation model is also crucial, for instance wind is currently not included in the model, but the wind will affect the propagation time. The localization performance, and especially the power-based localization, is drastically decreased when the target is not emitting isotropically and/or the microphones do not receive omni-directional. To cope with this problem or to increase the performance, a more detailed model is required.

Time-based localization is less dependent on the signal power than power-based localization. For a good position accuracy a good time accuracy is needed. The time accuracy is inversely proportional with the transmitted bandwidth. In other words, a signal with a high bandwidth can be localized in time very accurate which can result in a very good location accuracy, because the time accuracy is linked with the distance accuracy with the sound velocity. Except the fact that the signal will be attenuated and altered, the time-based localization is less distance dependent than the power-based localization.

The power-based localization is more dependent on the used hardware, the environment and the target distance. It is required to calibrate the hardware for correct comparison at different microphones. Without considering the absorption, the geometric spreading will cause the signal power to decrease 6dB for every doubling of the target distance. Therefore, the power-based localization accuracy is inversely proportional with the squared target distance, and thus the accuracy is more distance dependent.

Furthermore, mostly the target has no desire to be localized and more probably the target has the desire to prevent localization. Because the target signal is not optimized for localization purposes, it will be a challenge to correctly localize targets.

7.4 Theoretical classification performance

An estimation of the theoretical maximum classification performance would be very interesting, but it is also very challenging to achieve. The Bayes error rate would be a candidate, but this requires much measurement work and solid class knowledge for a reliable value and is very challenging to compute. Therefore, the computation of the Bayes error rate is considered as a separate assignment topic to be calculated at a later project stage. This section will discuss the general classification dependency.

The classification performance depends on the following:

- Features
- Classification method

The classification performance depends on the features. If the features allow good discrimination, the classification method will provide a correct result. The method is considered outside the scope of this thesis, but the feature extraction is designed and the dependency will be briefly discussed.

The designed features extraction process depends on:

- Transmitted signal
- Target distance
- Environment and ambient air
- Hardware specifications
- Node position accuracy and localizer

The main dependency is on the SNR and thus depends on the first four points. For instance, the SNR of pedestrian footstep signals and the engine harmonics are crucial for the feature extraction. The SNR is the crucial factor for detecting peaks in the time or frequency domain. The gun feature extraction also depend on the node and target location accuracy, because the Gun FE uses the channel compensation based on the target distance.

8 System evaluation

The implemented system is tested with an experimental setup for a system performance indication. The top-view of the setup is given in Figure 8-1.

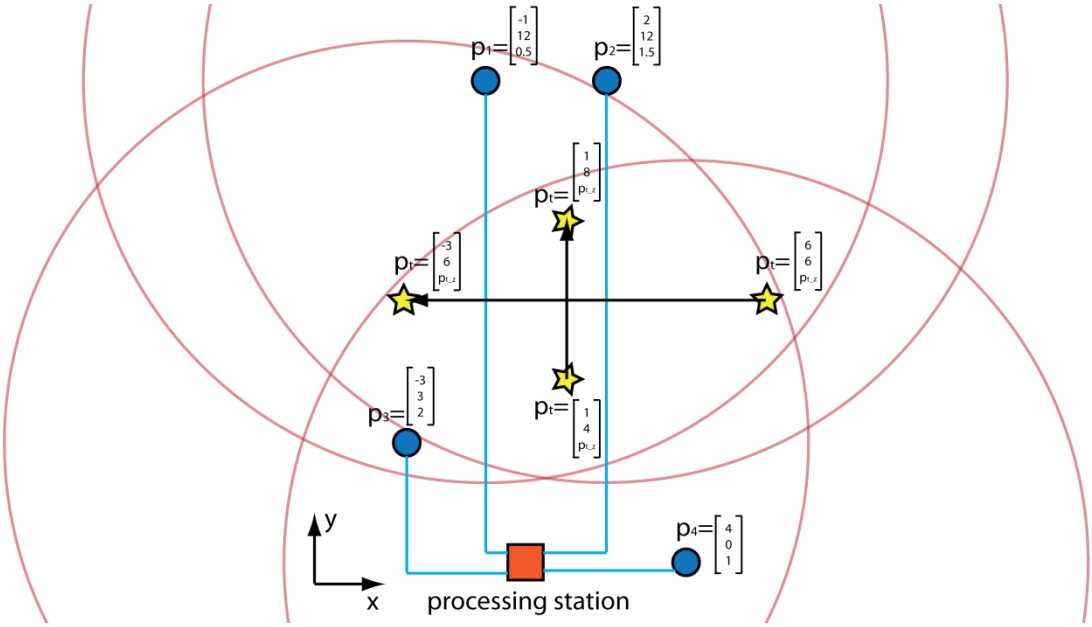


Figure 8-1: Performance measurement setup top-view

Four microphones are used in this experiment and two target trajectories are defined. Three coordinates are defined and the x, y and z coordinates are given in Figure 8-1. The path along (parallel) the y-axis is only used for the gun measurements. The red line indicates the assumed covered area of Section 1.7 of the microphone without considering the height.

Between the measurements, periodically a 30 second measurement is taken from the environment. These measurements without any targets are used to estimate the environmental noise spectrum.

The middle of the vehicle engine is placed along the path and for every new series of measurements the vehicle is moved 1 meter. Which means that 10 vehicle positions are defined. At every position three times a 3 second measurements is captured for both 1000RPM and 2000RPM. After the first six measurements, a 10 second measurements is captured of the vehicle when the driver is playing with the RPM. Thus in total 70 measurements are recorded. The height of the engine middle is approximately 0.65 meter.

On the ground the trajectory is drawn with chalk to assist the pedestrian. The footsteps are 0.8 meters separated and thus the trajectory allows 12 footsteps. Five pedestrians are asked to walk the trajectory four times, two in the positive and two in the negative direction along the x axis. Thus, in total 20 measurements of 12 footsteps. The footstep height is assumed to be 0 meter.

Finally, the toy-gun is mounted on a tripod to acquire a steady height of 1.35 meter. The toy-gun is placed on the two trajectories and is moved 1 meter after three shots. Thus, in total 42 toy-gunshots are fired.

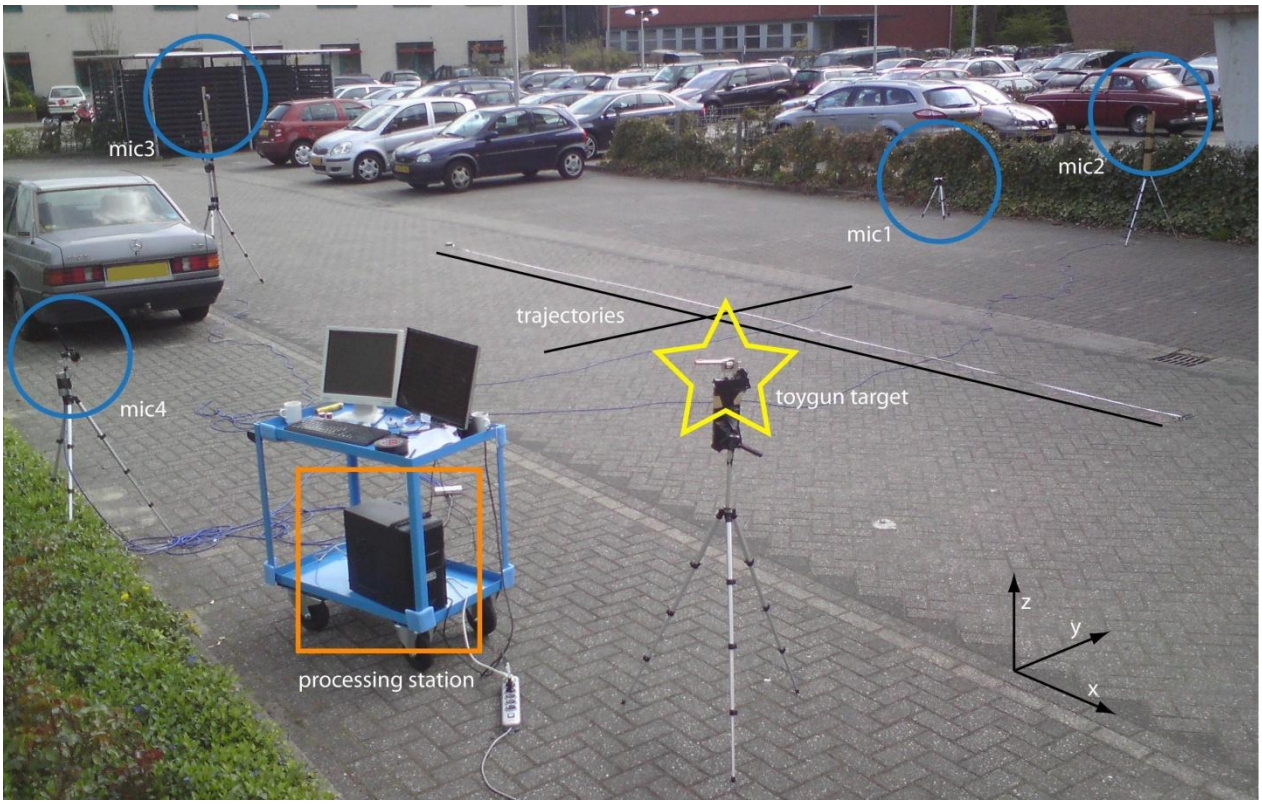


Figure 8-2: Performance measurement setup at Thales Delft parking area

In Figure 8-2 the real setup at the Thales Delft parking area is shown. Measurement date is 21 April 2010 and the temperature T_{air} is approximately +10°C and thus the approximately sound speed c_{air} is 337m/s. According to the weather report, the wind speed is approximately 5 m/s and the relative humidity is approximately 55%.

8.1 Data processing optimization

The experimental data processing, which is discussed in the previous chapters, is implemented in MATLAB. However, still some optimizations were needed to correctly localize and classify the targets. The optimizations, which are implemented, will be discussed in this section.

8.1.1 Localization optimization

Due to different microphone gains and multiple reflections some optimizations were needed for the Time SP, the Time IE and the Power SP for localization purposes.

The crucial point in time-based localization is the estimation of the correct time of arrival at different nodes of the Line-of-Sight emitted target signal. Reflection were present in the recorded data and the received Line-of-Sight signals did not always have the highest amplitude (peak). An algorithm is implemented which can cope with these problems and then can correctly and precise estimate the time of arrivals.

The selection of the mask for creating the third output of the Time SP, which is used for localization purposes, will first be discussed. In practice it became clear that the signal with the highest peak (local maximum) is not always the first received signal in time. In other words, the signal with the highest peak is not the signal which arrived the earliest. Most likely this is due to the different microphone gains and environment effects.

For the selection of the mask, it is desirable to select the earliest signal, but also a signal which has much power, thus a high amplitude. High power and early in time indicate that the target is nearby and that the signal has the least mutation. As mentioned above, the problem in practice is that the signal with the highest amplitude is not always the signal which arrived first. To cope with this problem, first the signal is searched for peaks. A peak search algorithm is described in Appendix B. After the peaks are found, the peaks receive a score:

$$W_{peak}(k) = \varphi_a \cdot W_{y\_a}\left[y_{peak}(k)\right] + \varphi_b \cdot W_{t\_b}\left[t_{peak}(k)\right] \qquad (27)$$

Where φ_a and φ_b are the parameters which decide how important the height of the peak is versus the time. The sum of φ_a and φ_b is one. $y_{peak}(k)$ is the amplitude of the k^{th} peak and $t_{peak}(k)$ is the time of the k^{th} peak. The function W_y gives the peak points for the amplitude and the function W_t gives the peak points for the time it was received. In Appendix B example weight functions are described.

After all the peaks are found and all the peaks have received a score, the peak with the highest score is selected. The signal around the selected peak is used as the mask for the third output of the time signal processing. Altogether, a mask is made which includes a peak which satisfies the two properties the most. The constructed mask is used to filter all the received signals.

The time information extraction has a similar approach. First the signals, which come from the time signal processor, are scanned for peaks. Also in this case, the highest peak is not always the correct peak, but also the earliest peak is not always the correct peak. It also seemed that

sometimes high peaks can be found far away in time from where the mask was created. To cope with all these problems, three properties are introduced:

$$W_{peak}(k) = \varphi_1 \cdot W_{y\_1}[y_{peak}(k)] + \varphi_2 \cdot W_{t\_2}[t_{peak}(k)] + \varphi_3 \cdot W_{t\_3}[t_{peak}(k)] \quad (28)$$

The summation of φ_1, φ_2 and φ_3 is one. The approach is similar as in the Time SP, and for every signal the best peak is selected. The time of the best peak is used as the T_{0Ai} for that particular partition.

To illustrate the approach, an example of time signal processing of four received signals is given in Figure 8-3. The upper plot shows the selection of the best mask from signal four, although it did not have the highest peak (signal two has the highest peak). The dotted green line with a slope illustrates the linear weighting of the peaks for their time property.

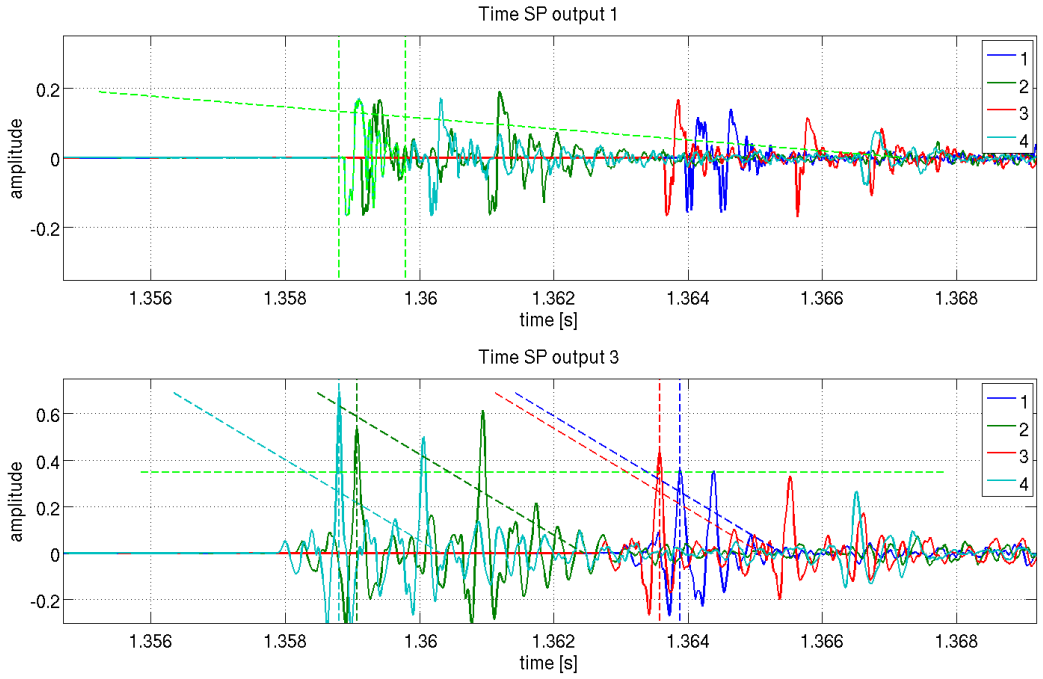


Figure 8-3: Time-based gun information extraction

The lower plot shows the selection of the correct time of arrivals. The horizontal dotted green line illustrates an uniform W_{t_3} weighting for the location around the mask. The other four dotted sloping lines illustrate a linear W_{t_2} weighting. The vertical dotty lines show the corrected selection of the time of arrivals.

Altogether, the above optimizations allowed a lower SNR - and still maintaining a correct location - than by just selecting the highest peak. With these optimizations is was possible to correctly localize certain footsteps instead of almost none.

Two things had to be solved in order to construct the power signal processor. First, a correct (relative) microphone gain had to be estimated and this is done as explained in Appendix A.

Second, the environment has to be estimated. This is done by measuring multiple times the environment without targets for 30 seconds. For every microphone separately, the environment recordings were partitioned and the average of all the spectra is used for the environment estimation. The environment signal did not contain much energy and therefore output 2 does not differ much from output 3. An example of the three outputs of the Power SP as explained in Section 6.2 are shown Figure 8-4 in power density.

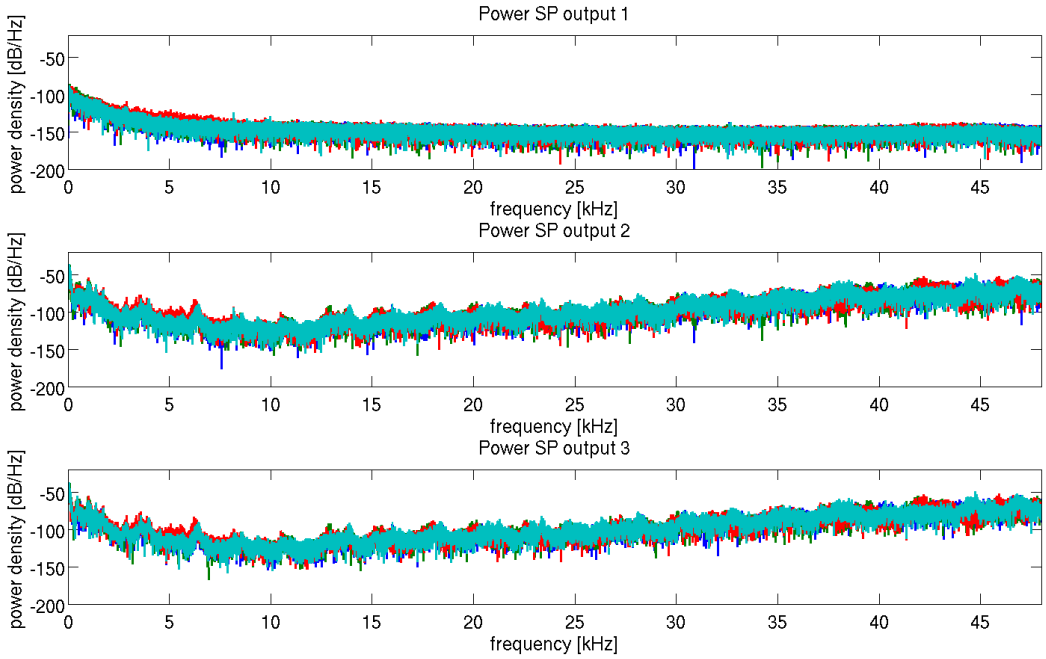


Figure 8-4: Power signal processing with 1000RPM running piston engine

The first Power SP output shows the standard mean squared power spectrum. The microphone gain decreases when the frequency increases, and the second output shows the compensation for this. The third output is, due to little environment noise, not much different from the second output. The idea of the Power SP is that when no target is present the third output would be flat, however mainly due to a not perfect microphone gain estimation, this could not be achieved. This situation indicates the need for further hardware study.

8.1.2 Classification optimization

In this section the optimization of some signal processing and feature extraction parts will be discussed.

For a correct Human FE, a suitable mask is needed to filter the instantaneous power to increase robustness. A sufficient long edge mask is selected and the result is given in Figure 8-5. For the correct footstep time extraction, the microphone with the highest received power is selected. Note: the plot also shows that the second Time SP output can also be used to extract time information for localization. Also note that the SNR of the second output is better and more robust in comparison to the first output. A new peak/footstep has to be at least 0.2 seconds separated in time from the previous footstep, otherwise it is most likely a reflection or the second sound of the footstep.

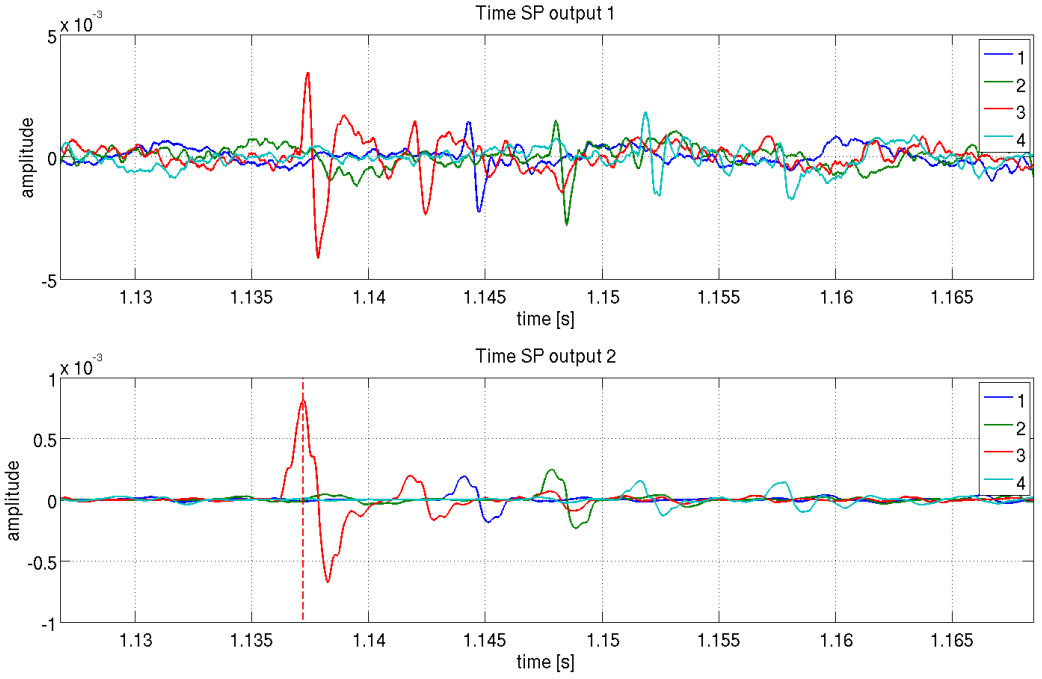


Figure 8-5: Time signal processing and human feature extraction

An illustration of the gun feature extraction is given in Figure 8-6. For the four microphones, four spectra with 12 points are plotted. The no target spectrum shows the spectrum when no target is present. When the toy-gunshot is present the power spectrum increases. The gun feature spectrum for every microphone is eventually the summation of the toy-gunshot spectrum and the channel compensation spectrum, which is explained in Section 5.1. The FE process is done with the four microphones and the last result can be the combination of the four result into one result by weighting.

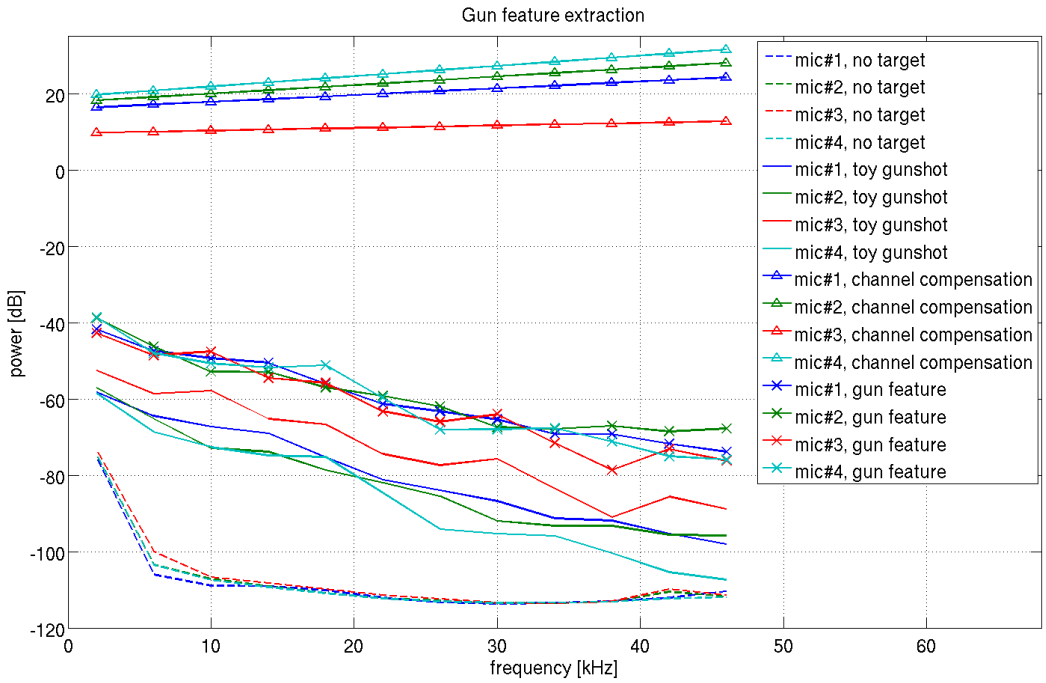


Figure 8-6: Gun feature extraction

The Vehicle FE was done by searching for the first three peaks in the power spectrum. Figure 8-7 shows a Vehicle FE example at four nodes with a 1000RPM running engine. As the example shows the SNR is not always good (barely 20dB), and with the measured data the SNR sometimes dropped further. Luckily, with the 2000RPM the SNR is higher.

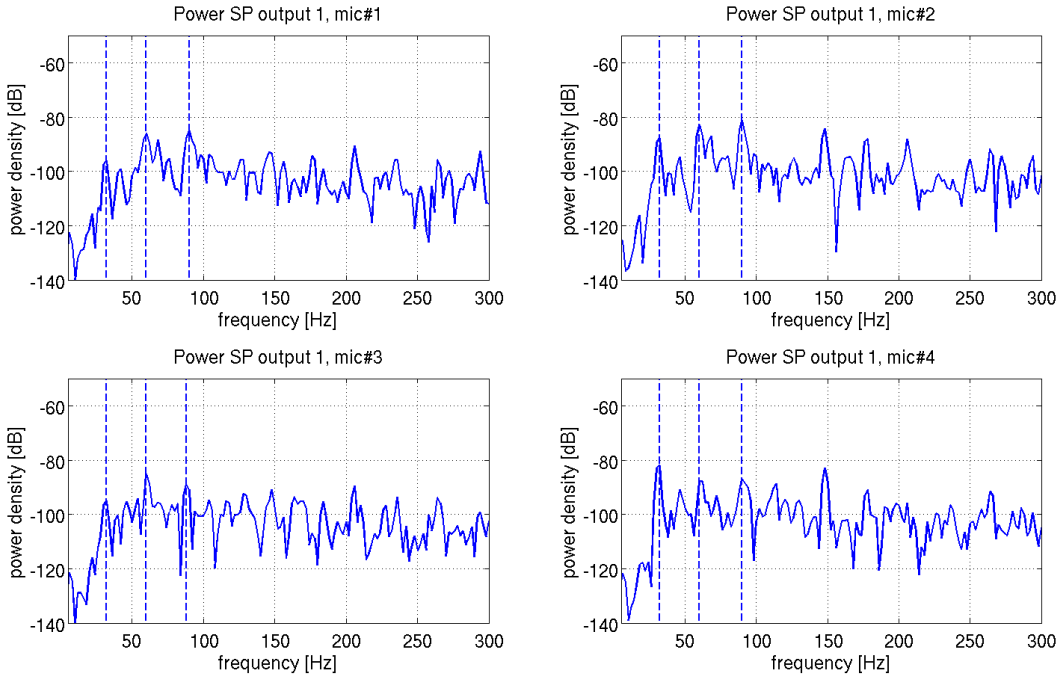


Figure 8-7: Power signal processing and vehicle feature extraction with 1000RPM vehicle

Feature extraction works in the EASN, however the FE has become a kind of detection. More clearly said, first a target has to be detected, before features can be extracted. For example with the vehicle features, first some peaks have to be found in the spectrum before the harmonics can be extracted. In other words, the FE process can only provide properties, if a certain class is detected.

Figure 8-8 shows three experimental detection trees, which are also known as binary classification trees. Every detection tree can decided if a certain target class is present. For every tree, two feature values are chosen. The black typed value can be enough, but the gray typed value can be added to increase the robustness. Adding the gray features will reduce the false alarm rate, but also increase the target miss rate.

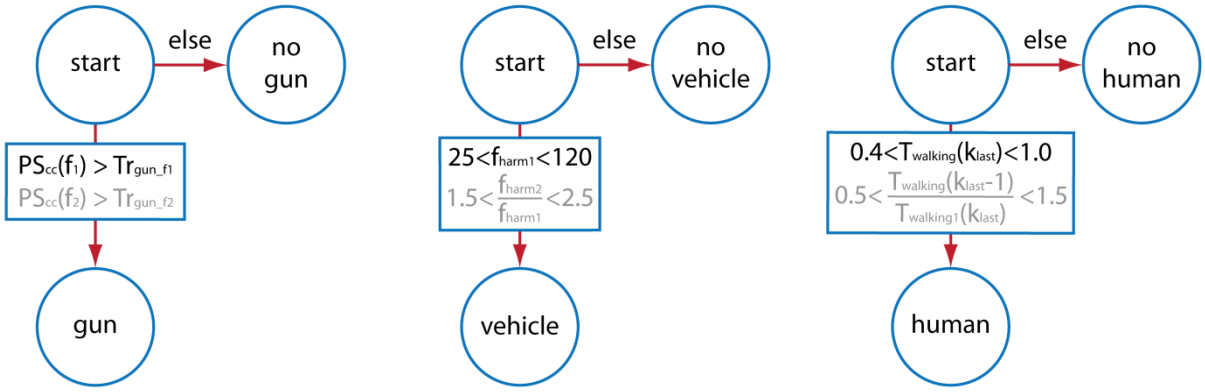


Figure 8-8: Experimental Detection Trees

The above values are experimental and chosen by the designer. Although class knowledge is beyond the project scope, this is needed for the experimental classification. The values are chosen by using the experimental recordings and the knowledge of the designer. The three detection trees will be discussed below.

The first detection tree decides if there is a gun with the gun features. A gun is detected if the channel compensated power is high enough. The threshold depend on which output of the Power SP is used for Gun FE.

The vehicle features are used for the decision if a vehicle is detected. The second vehicle feature, which is the ratio between the first and second harmonic, can be needed against gunshot sounds. With a gunshot a peak is sometimes found between 25Hz and 120Hz, but then the ratio between the first and second peak is not equal to two.

Also for human detection, the first feature is already sufficient. The second feature result in extra robustness, but also in a slightly (~15%) higher human miss rate. This is because, sometimes some footsteps are not detected and not both walking times are correct. In practice the second footstep tick is difficult to extract, due to the fact that the tick is not distinguishable from reflections and more importantly, it is not always present. Therefore the footstep times $T_{footstep}$ can best be ignored at the classification process for the used EASN.

8.2 Performance indication

Using the experimental implementation and the recorded data, an indication of the system performance can be given. First the localization performance will be discussed and second the classification performance.

8.2.1 Localization performance

With the measurements gathered from the experimental test, a performance indication can be given.

The mean error for N estimated positions will be defined as:

$$\mathrm{ME} = \frac{1}{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{N}}(\hat{\mathrm{p}}_{\mathrm{t}}(\mathrm{k}) - \mathrm{p}_{\mathrm{t}}(\mathrm{k})) \tag{29}$$

Where $\hat{\mathrm{p}}_{\mathrm{t}}(\mathrm{k})$ is the estimated target position and $\mathrm{p}_{\mathrm{t}}(\mathrm{k})$ the actual target position of the k^{th} signal. The Root Mean Square Error (RMSE) is given by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{N}}\|\hat{\mathrm{p}}_{\mathrm{t}}(\mathrm{k}) - \mathrm{p}_{\mathrm{t}}(\mathrm{k})\|^2} \tag{30}$$

The standard deviation of the estimated target positions is given by:

$$\mathrm{SD} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{N}}\left\|\hat{\mathrm{p}}_{\mathrm{t}}(\mathrm{k}) - \frac{1}{\mathrm{N}}\sum_{\mathrm{n}=1}^{\mathrm{N}}\hat{\mathrm{p}}_{\mathrm{t}}(\mathrm{n})\right\|^2} \tag{31}$$

In the experiment the $\mathrm{p}_{\mathrm{t}}(\mathrm{k})$ changes over k, therefore the average of $\hat{\mathrm{p}}_{\mathrm{t}}(\mathrm{n})$ is changed with different $\hat{\mathrm{p}}_{\mathrm{t}}(\mathrm{n})$ so the correct SD is calculated. The above three formulas can be used for all three dimensions.

First the time-based localization of the toy-gun was investigated to obtain a performance indication. The toy-gun was placed along the two trajectories which resulted in total 42 shots. All shots could also be properly localized with the third Time SP output, because the SNR was sufficiently high. The estimated toy-gun 2D locations are plotted in Figure 8-9. The toy-gunshot performance is given in Table 3.

The toy-gunshot localization measurements show two main things, which is also seen in the pedestrian localization measurements. Firstly, there is a offset in the system, and mostly the offset is present in the z dimension, but also in the x dimension. This system offset is probably due to an incorrect measurement setup. The parking area is not perfectly flat and maybe the height of the microphones is not correctly measured. Secondly, due to the different spatial microphone separations in the different dimensions, the SD is also different for the different dimensions. The y dimensions has the best microphone separation and the lowest SD, and the z dimension has the poorest microphone separation and the highest SD.

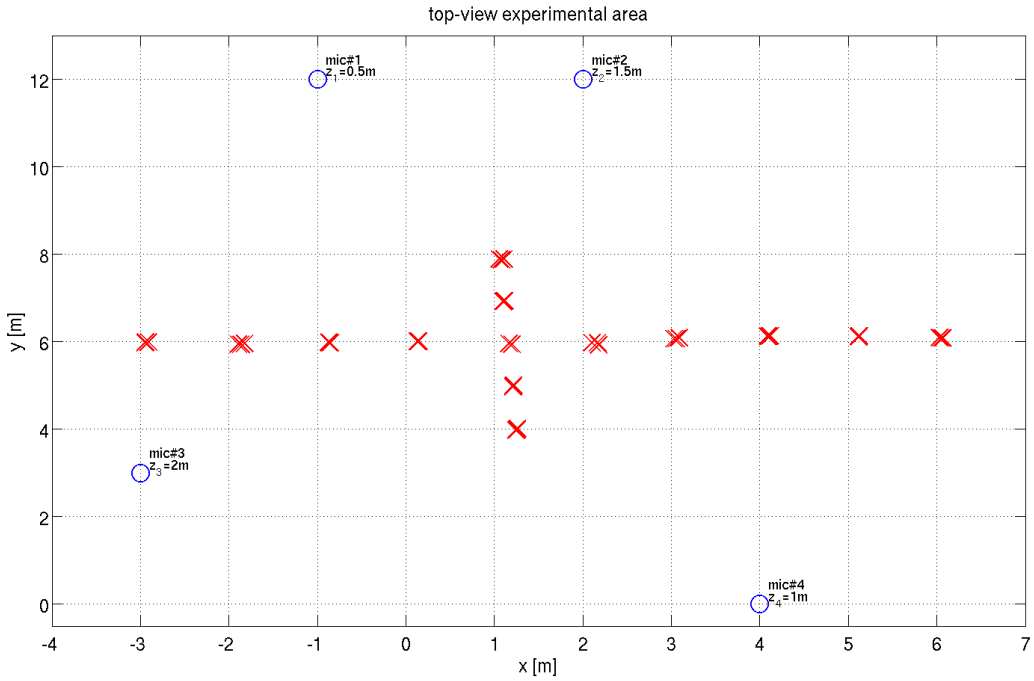


Figure 8-9: Top-view estimated toy-gun locations

Table 3: Muzzle blast localization performance

dimension	**ME [m]**	**SD [m]**	**RMSE [m]**
x	0.13	0.02	0.14
y	5e-4	0.01	0.07
z	0.22	0.03	0.24
x and y (2D)		0.02	0.16
x,y and z (3D)		0.04	0.29

The time-based localization of the pedestrian was a challenging task, because for a good target location a good signal was required at all four microphones. Although the persons were selected for their optimal shoes, only 46 of the 240 footsteps could be localized. One of the five persons had such good acoustical sound shoes that she was responsible for 29 good localizations.

In Figure 8-10 a footstep is correctly processed by the Time SP and the Time IE. This example represents a good footstep signal, but mostly it is much worse (lower SNR) and localization is significantly more difficult than with the gunshots. The point is: correct time information extraction is very difficult with such a low SNR. Good and robust localization starts approximately when the SNR at each microphone is at least 15dB.

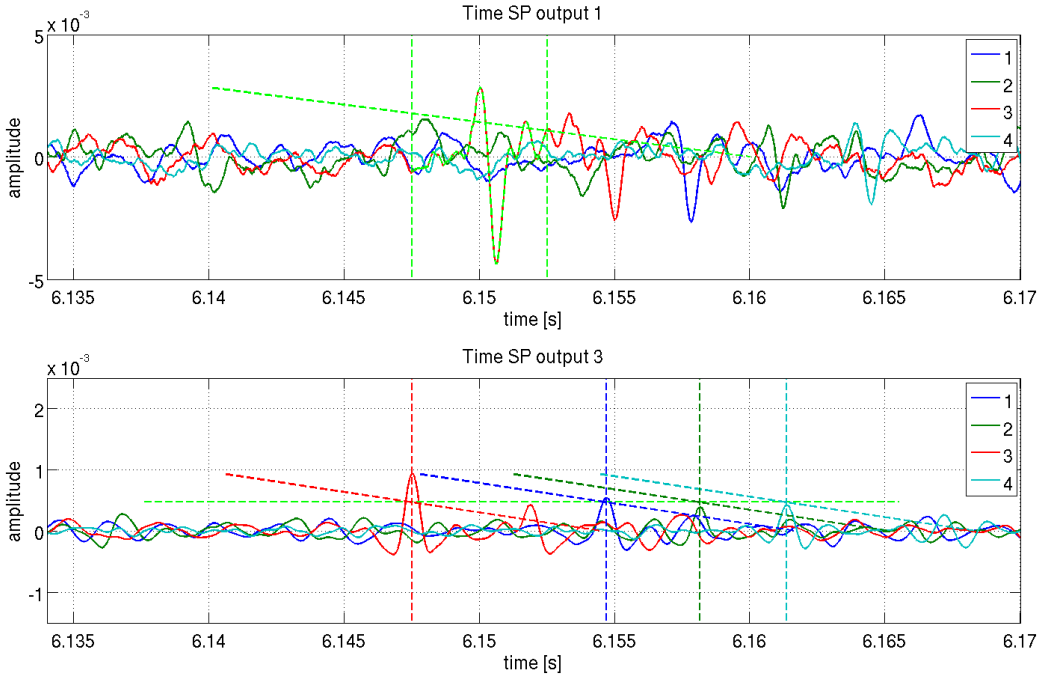


Figure 8-10: Time-based footstep information extraction

For footstep localization performance only two dimensions will be used. The dimension along the x-axis is calculated, but due to the different footsteps and persons, the estimation of the original position along the x-axis was too difficult for this experiment. This is because with the recordings it was unclear which footstep belongs to which footstep sound, due to the fact some footsteps were not present. Furthermore, the error along the x-axis is mostly due to the pedestrians, because the heel was not always placed on the correct location. Thus, the actual target position is 6m along the y-axis and 0m along the z-axis. The localization performance of the 46 footsteps is given in Table 4.

Table 4: Footstep localization performance

dimension	**ME [m]**	**SD [m]**	**RMSE [m]**
y	-0.06	0.27	0.28
z	0.31	0.70	0.76
y and z (2D)		0.75	0.81

The difference between the footstep and the muzzle blast performance is due to the different SNR and bandwidth, which is discussed in Chapter 7. The toy-gunshots are much louder and sharper in time. An examples of a pedestrian walking the trajectory is given in Figure 8-11.

After the time-based localization, the power-based localization performance was investigated. The power-based performance was significant worse than the time-based solution. For the power-based localization the vehicle measurements with 1000RPM and 2000RPM were used. Power-based localization was done with a 0.5 second recording (partition), thus with the recorded data 360 vehicle positions could be estimated.

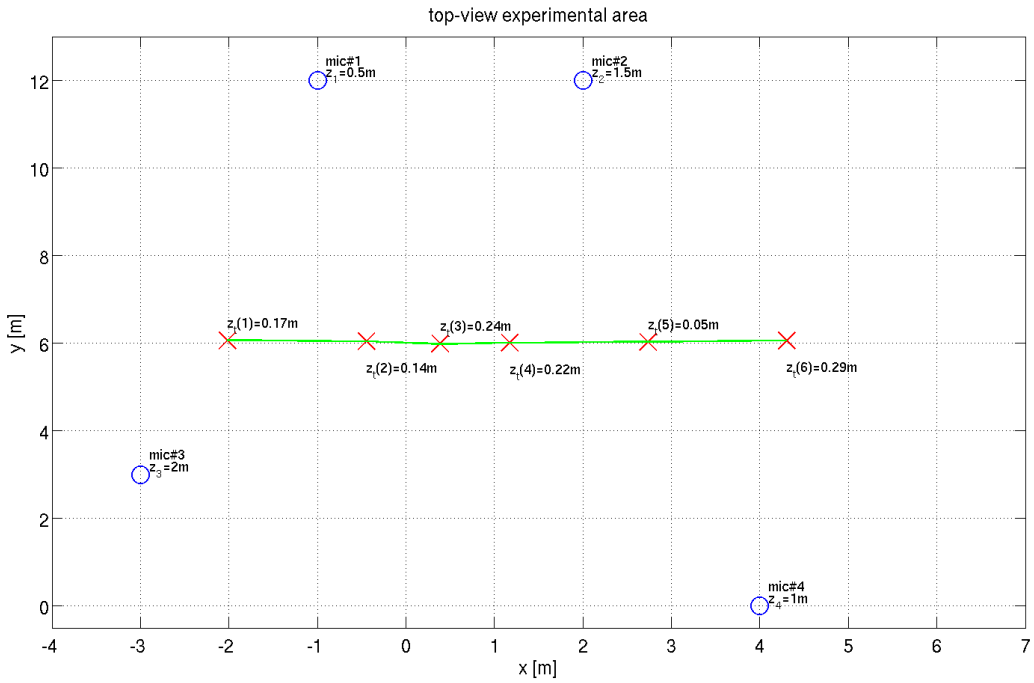


Figure 8-11: Top-view six estimated footstep locations from a pedestrian walking the trajectory ones

First the position estimation with the first input of the Power SP and without considering absorption was done. Thus, localization is done with the average power of the windowed partition. The performance is shown in Table 5.

Table 5: Vehicle localization performance without using the absorption model for 1000RPM and 2000RPM

dimension	**ME [m]**	**SD [m]**	**RMSE [m]**
x	2.08	1.35	2.75
y	1.45	0.83	1.76
z	2.52	1.28	2.89
x and y (2D)		1.58	3.27
x,y and z (3D)		2.04	4.36

The errors are calculated with assuming the middle of the vehicle engine as the emit point of the target. The SD performance along the y-axis is the best, and this is probably due to the vehicle dimensions and the spatial separation of the microphones along the y-axis. Along the x-axis the vehicle exhaust and the vehicle engine is most separated and therefore the performance along the x-axis is not optimal. The error along the z-axis is probably due to the least spatial separation of the microphones.

To improve the above performance is very challenging. The use of the second output of the Power SP, which includes the microphone gain, did not improve the performance. This is probably due to the incorrect microphone gain estimation. And the performance of the third output, which also includes the environment spectrum, was even worse.

To further test the power-based localization, the gun measurements were used. Without considering absorption and microphone gain compensation, the power-based gun localization performance is given in Table 6. With switching the gun for the vehicle, the performance increased, but this is mostly due to the fact that the gun dimensions are much less than the vehicle dimensions. The distance between the engine and the exhaust is huge (~3m). Figure 8-12 shows the top-view of the gun locations.

Table 6: Gun power-based localization performance without using absorption model

dimension	**ME [m]**	**SD [m]**	**RMSE [m]**
x	2.15	1.08	2.75
y	0.19	0.46	0.80
z	2.47	0.61	2.60
x and y (2D)		1.18	2.87
x,y and z (3D)		1.33	3.87

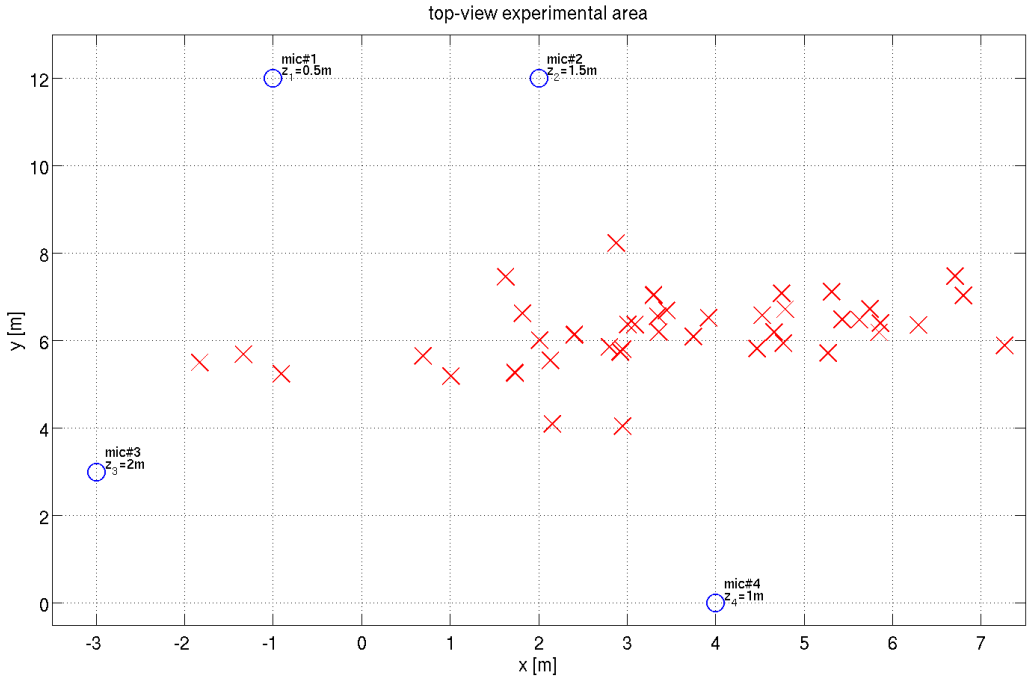


Figure 8-12: Top-view estimated toy-gun locations with power-based localization

The main conclusion of the power-based localization experiment is that better hardware is needed to improve the localization performance. Power-based localization requires extensive calibration and a good microphone gain has to be determined. Unfortunately, using the absorption model also did not improve the performance and perhaps a new absorption model is needed. Power-based localization with the provided hardware can only give a rough indication.

8.2.2 Classification performance

A classification performance indication can be provided with the recorded data. First, the performance of the vehicle and human feature extraction processes are discussed. Second, the performance of the features is investigated with a classification tree.

For the vehicle features, finding the harmonics in the spectrum is a challenge with a SNR lower than 20 dB. For 1000RPM the harmonics are mostly below this value, but for the 2000RPM measurements, the harmonics are mostly above this value. Another thing that strikes is that sometimes a certain harmonic signal power drops, and an example is given in Figure 8-13. The dotted line illustrates the selection of the peak. This example shows that the second harmonic at the fourth microphone is not detected, but sometimes the peaks drop even further in the noise. As the example plots show, at certain angles and distances of the vehicle some harmonics can be damped, but it also shows the potential for data-fusion techniques.

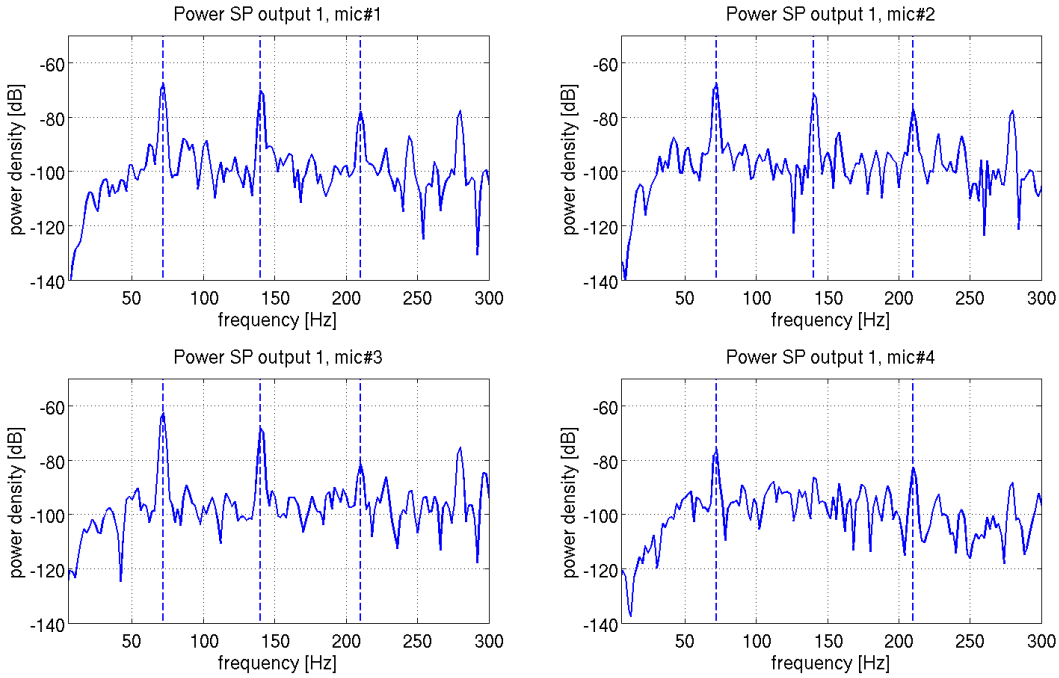


Figure 8-13: Power signal processing and vehicle feature extraction with 2000RPM vehicle

Where footstep localization required good signals at all the four microphones, Human FE only requires a good signal at one microphone. However, to determine the human features, three footsteps are required. With the recorded database, only 46 footsteps were localized, but 68 correct human feature vectors were created.

An example of Human FE is given in Figure 8-14. The example shows that a pedestrian footstep time is first extracted at the third node and next the footstep time is extracted at node two. This is explainable, because the pedestrian was walking in the positive x-direction and thus the target distance to node two decreased and thus the received power increased.

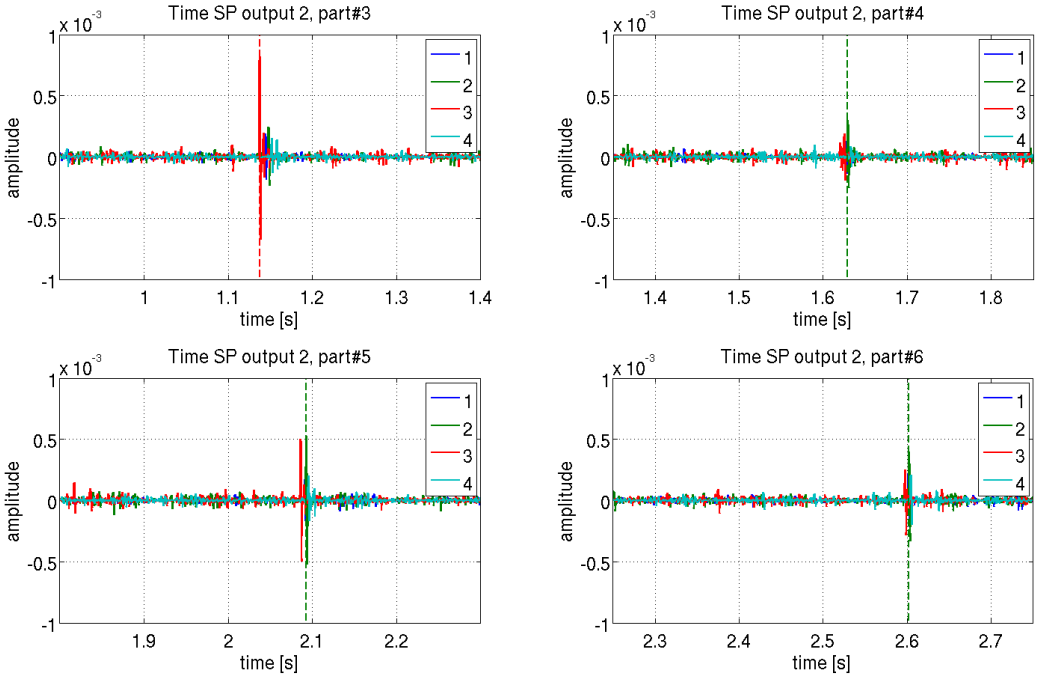


Figure 8-14: Time signal processing and human feature extraction for multiple partitions

When Human FE extraction is applied on a vehicle signal, there are no classification concerns. Although the vehicle signal is periodic, the extracted time at the Human FE is much smaller than with footsteps as Figure 8-15 shows.

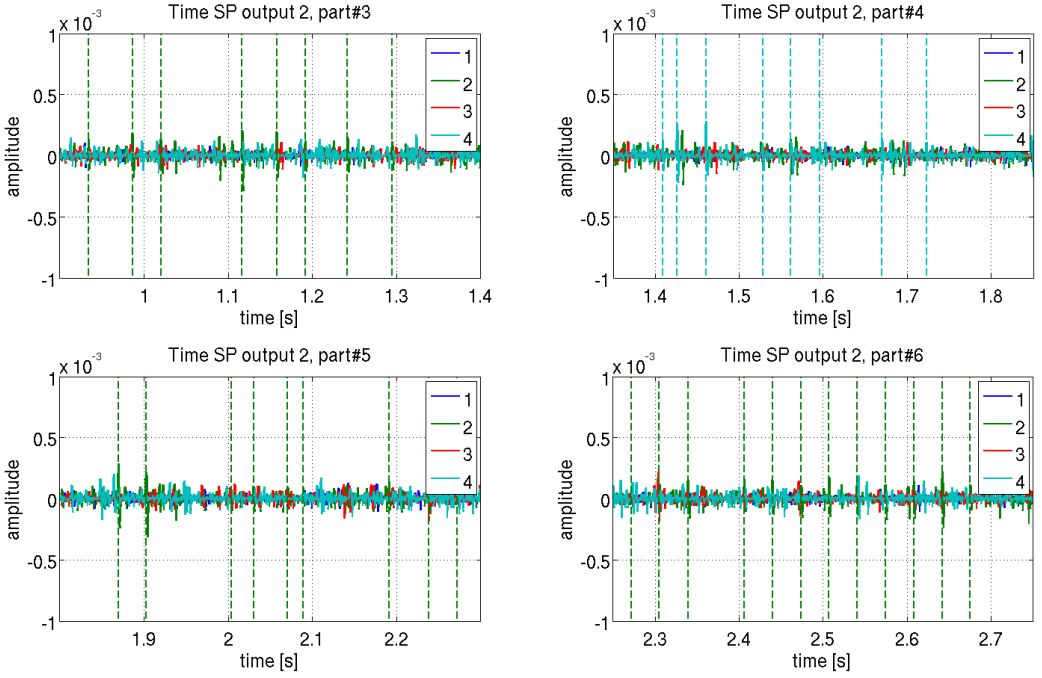


Figure 8-15: Time signal processing and human feature extraction with vehicle sound

As already discussed, the Gun FE depends on the localizer. If the threshold for the Time IE is very low to localize the footsteps, some footsteps are incorrectly localized. Therefore, if then the (incorrect) target distance is very high, channel compensation will be high and thus a footstep can be classified as gun. A solution is to increase the Time IE threshold and miss footstep locations.

All the three feature extraction processes benefit from the multiple microphones. The harmonics extraction is a little bit corrected by combining the features. Also the gun features benefit in the same way. The human features can be extracted from the best microphone, and also this works fine.

Classification can be seen as the detection of a certain class. The considered target classes are so much separated, classification is relative easy, because a kind of detection is already done in the FE process. To show the power of the provided features an Experimental Classification Tree (ECT) is used for the assumed operational situation and is shown in Figure 8-16. Eventually the ECT is the combination of the three detection trees of Figure 8-8. The class-detection order, thus which feature should be used first is a fundamental problem with classification trees. This ECT is chosen, because this mapping allows a good classification result with only three features values.

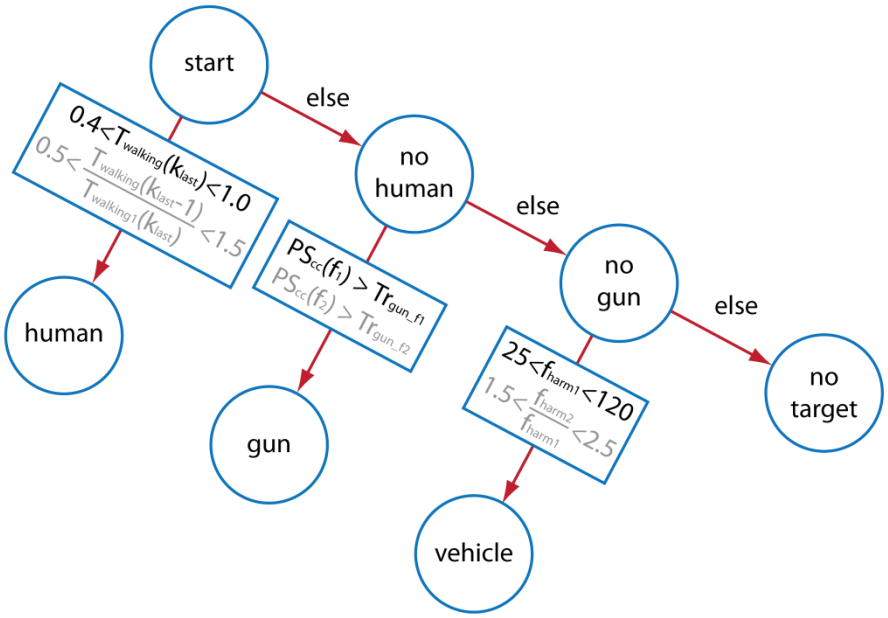


Figure 8-16: Experimental Classification Tree

The above ECT with only three features works perfectly well for the current operational situation. Four of the five pedestrians are correctly detected. With a correct threshold, all the gunshots are classified as gun. And all vehicle recordings are classified as vehicle. Thus, the miss rate and false alarm is - with this measured data - almost zero. When the thresholds of Time IE are decreased and footsteps are localized, the above discussed incorrect target distance arises, but also the power different of a footstep and a toy-gunshots is not always enough to distinguish. However, real gunshot muzzle blast (instead of toy-gunshots) produce signals with much more power, thus with real guns this will not be a problem.

9 Conclusion

To improve the situational awareness (via localization and classification) in urban environments, Thales Nederland has initiated a project to investigate the potential and feasibility of acoustic sensors. The operational problem is not merely to detect targets, but also to localize and classify them in a robust way. Current radar implementations do not provide enough performance to classify targets in complex urban environments while now acoustic sensors are seen as an extra source of information. Before this project started, the acoustic knowledge at Thales Delft was limited, no previous acoustic Thales projects were available and only few papers were downloadable concerning ASNs. This project has performed many tasks to bring a working ASN a little bit closer by providing new acoustic surveillance knowledge.

Three different kind of targets were investigated: guns (muzzle blast), vehicles (running piston engine) and humans (walking pedestrian). Furthermore, an operational situation was assumed to reduce the scope and to give the project a direction. The Experimental ASN would be deployed in an urban environment with air as transport medium. Only one target should be present, the target distance to the microphones was limited to 10 meters and a Line-of-Sight signal should always be received at each microphone.

Two project objectives were defined. Firstly, provide target information with propagation models, which allows target position estimation. Secondly, provide target features, which allows discrimination between the target classes. To achieve these objectives is a major challenge in complex urban environment. The challenge is that these goals should be achieved only with the recorded sound signals from a passive EASN deployed in an urban environment and that nothing is known about the target. Furthermore, the target has absolute no desire to be localized or classified. Experimental data processing had to be designed, implemented (proof of concept) and evaluated with measured data to obtain a performance indication.

To accomplish the project objectives, the project has followed an effective approach. The current knowledge of acoustics, localization, classification has been studied in an exhaustive literature survey to investigate the general possibilities. Then the acoustic sound signals of the selected targets have been measured and investigated to find out how localization and classification of these targets with sound can be achieved. Next with this knowledge, methods have been designed to extract the correct information for localization and to extract the required features for classification. After the methods were designed, a working experimental system had to be designed, implemented and tested to obtain a performance indication. With limited available previous research knowledge a proof of concept implementation with experimental hardware has been constructed to provide a performance indications and to show the potential for acoustic surveillance.

This project has clarified the difficulties, but also the opportunities for urban acoustic surveillance. Furthermore, this project makes it clear what the next steps should be to bring a working ASN closer. This chapter will discuss the project results more deeply in Section 9.1 and Section 9.2 will outline the recommendations for further research.

9.1 Project results

The localization method was to provide localization equations to a mathematical solver. These equations were built with the propagation time and propagation loss models and they allowed different kind of target class localization. With these models, a general, but robust design has been made to extract the correct time and power information. The time information is extracted from a pseudo-matched-filter signal. The power information could be extracted from three differently constructed power spectra. When the models and information were combined, a least square solver could estimate the target position.

Another goal of this project was to find features of the selected targets. The foundation was an acoustically analysis of the three selected targets. The muzzle blast, running piston engine and walking pedestrian had very different signal properties. It was discovered that the emitted powers, harmonics and footstep intervals could be used as features. These discriminative physical features, which are explainable and understandable, allows a simple classification method to distinguish between the classes. With the current hardware and targets, there were no special features found in ultrasound and in the current stage, Wavelet Analysis or De Groot Fourier Transform techniques, which is introduced in Appendix D, did not provide new kind of features.

To test the results and to provide a performance indication, an Experimental ASN system was implemented. The system design is done in a universal way which allows follow-up projects to use this design. Certain thresholds and parameters can still be chosen freely or determined by new designers. For example, the localizer is not matched to certain target classes and the human feature extraction is not matched to certain shoes. The experimental design is general which will allow very different types of signals.

The time-based localization needed some optimization, but then it performed good. It gives a very precise target position ($RMSE_{3D}$ of 0.29m, SD_{3D} of 0.04m) if the SNR is high enough (at least above 15dB at each microphone). Lower SNR is sometimes possible, but then the performance drops and the probability of false target positions will increase (and the error will probably not be Gaussian). Power-based localization works, but the current accuracy ($RMSE_{3D}$ of 3.87m, SD_{3D} of 1.33m) can probably be increased if the microphone gain is estimated correctly with more professional hardware. The current absorption model was insufficient and should be replaced with another model, alternatively power-localization can also be done in time-domain with the average power.
A noticeable thing is that the information and feature extraction components became detectors. First the peaks in time or frequency-domain should be detected before the feature values can be constructed. This allowed a simple classification tree to easily classify the very different targets. In other words, the chosen features allowed easy discrimination between the targets and the success rate with a classification tree was high (with the recorded data above 90%).

With the knowledge and experience gained in this project, the main conclusion is: passive acoustic surveillance is possible and it can provide extra situational awareness, but the system performance drastically depends on the target and the environment (e.g. background noise). In other words, system performance depends on SNR, and the SNR depends on the target class. To emphasize, the potential for a working acoustic surveillance system depends on the target class,

because this determines how much power is emitted. A short overview with the three investigated target classes is given in Table 7.

Table 7: Summarized conclusion for the current target classes, operational situation and experimental hardware

	Localization	**Classification** (/ detection)
Gun (muzzle blast)	**Possible**, SNR is high enough	**Potential**, depends on localization robustness
Vehicle (running piston engine)	**Potential**, but more hardware study is needed	**Potential**, but SNR has to increased for extra robustness
Human (walking pedestrian)	**Low potential**, the current SNR is too low	**Low potential**, for certain shoes, gait and/or ground

Human detection has the least potential. This is mainly due to the fact that for most pedestrians the SNR is too low. In the experiments of this project, pedestrians were selected for their shoes. And even with these pedestrians localization and detection was difficult. If the conditions are changed, for example with a better ground surface, the potential can be increased.

Vehicle localization and classification has higher potential. However, for extra robustness, the SNR has to be increased with further hardware study or environmental noise reduction. The first step to improve the power-based localization is to better estimate the microphone gains.

Gun localization has the best potential, because the SNR is high. It is also not surprising that current acoustic surveillance systems can already localize gunshots. The classification has also good potential, but the gun feature extraction depends on the localizer and thus the localizer has to be very robust. A false target position can result in a target miss or a false alarm, but gunshot localization is relatively easy.

The experimental data processing was limited by the used experimental hardware. For example, the microphone gain difference could not be estimated properly. Mainly the vehicle and human will benefit if the SNR is increased. The SNR can be increased with better hardware, but the environmental noise is difficult to reduce. The human localization would also benefit when the microphone density is increased. When the hardware performance and the microphone spatial density is increased the expected conclusion is given in Table 8. Localization is possible when the SNR is sufficient enough, but classification will always depend on the defined target classes.

Table 8: Estimated summarized conclusion for acoustic surveillance with improved (hardware) design

	Localization	**Classification** (/ detection)
Gun (muzzle blast)	**Possible**, SNR is high enough	**Potential**, depends on localization robustness
Vehicle (running piston engine)	**Possible**, new absorption model to improve	**Potential**, depends on other target classes
Human (walking pedestrian)	**Possible**, if SNR is high enough	**Potential**, depends on other target classes

9.2 Project recommendations

The project recommendations are based on the current system performance and the potential of the acoustic surveillance system. The research for ASN is huge and this project was only able to investigate a certain operational situation with certain targets, however this project shows the potential and possibilities of ASNs. First the main recommendations are outlined and next some recommendations are discussed to increase the capabilities of the current experimental system.

9.2.1 Main recommendations

For further research the recommendations are mainly related to the main issues that have not been covered in this project. In Chapter 1 some topics were excluded, but all those topics could have potential. The three main recommendations for further acoustic surveillance research are:

- Acoustic environment and target modelling
- Further hardware design and technology study
- Data-fusion to cope with many sensors

Modelling has the potential to give much insight into acoustic propagation in urban environments. With the results, the system parameters can be better determined and more can be concluded about SNR (e.g. background noise). The models can also be used to investigate multi-target detection, tracking, data-fusion and maybe multipath cancellation.
For a better indication of the performance limit, further study is needed in microphone and amplifier hardware. A crucial question is: can the system noise be further decreased?
The aim of Thales is to implement a system which contains many sensors. How to cope with all this information is crucial for the system design, but maybe also for the network energy consumption. Data-fusion can be used to cope with this problem, but also to further increase the SNR and to increase the feature extraction robustness.

9.2.2 Multi-target, tracking and extra classes

The basic idea behind multi-target detection is: the target has to be separated in time, space and/or features. If, for example, two targets have all the three things in common multi-target detection is impossible. However, sometimes tracking can cope with this problem in particular situations. Tracking can also be used to estimate the target velocity. The target velocity can then be used as a feature in the classification process.
Classification can become more difficult when the amount of classes increases. For example, the difference between a muzzle blast and a "standard" explosion is almost none. Classification has the precision and generality problem. Targets within the "same" class can sound very different, but targets from "different" classes can sound very similar. Maybe the investigation of targets in urban environments is too difficult. The acoustic surveillance of air units instead of ground units can also be interesting (less environmental effects). Altogether, adding new target classes can be very interesting, but the designer should aim to find physical and explainable features instead of providing dubious features with system learning which do not provide insight.
Other topics in acoustic research can be: theoretical system performance analysis for localization and/or classification. Investigate acoustic active, seismic or arrays sensors. The sky is the limit.

Glossary

Acoustic
Propagation of a wave due to pressure difference in a medium.

Acoustic Sensor Network
A sensor network where the nodes have acoustic sensors. In this thesis the sensors are measuring sound with microphones.

Classification
A target is classified when it is determined in which class the target belongs. Classification can be seen as detection of a certain class. A classifier decides to which class the target belongs.

Detection
A target is detected when it is determined that the target is present in the covered area. A detector classifies between target present and target absent.

Environment
Every sound and object except for the defined target sounds and the targets.

Experimental Acoustic Sensor Network
The ASN which is designed, implemented, evaluated and discussed in this thesis for experimental reasons.

Feature extraction
The construction of characteristic values for a certain signal, which will allow the classifier to discriminate among the classes.

Gun
A projectile firing weapon where mostly the projectile is propelled due to an explosion.

Human
A living creation of God, who has conscience, can make decisions and can have relationships.

Identification
A target is identified when the identity of the object is estimated. Identification is more precise than classification, but in principle it is the same.

Information extraction
The process of extracting values from a received signal for localization purposes.

Localization
A target is localized when the target position is estimated. A localizer estimates the target position.

Network
A group of distributed nodes which are connected to each other.

Node
Nodes, also known as mobile agents, can communicate with each other and are part of a network. For the EASN the nodes consist of one microphone.

Noise
Signals which are not produced by the considered target.

Partitioning
Division of the original signal into multiple parts with certain length and overlap.

Operational situation
The circumstances in which the system should work.

Processing station
The main computer where the network gathers and processes all the sensed data.

Recognition
A target is recognized, when the system has discovered sufficient distanguisable features of the target. Recognition comes after cognition.

Sensor
A device, which can measure physical facts, for example sound pressure.

Sensor Network
A network where the distributed nodes have sensors. The scenario generally consists of a large number of nodes that collaborate to gather information from multiple locations.

Situational Awareness
The perception of elements/things within a volume of time and space.

Sound
Acoustic signals through gas, for example air.

Supersonic
An object is going supersonic, when it is moving faster than the speed of sound.

System
A system is a construction, which can do a certain task with inputs and construct certain outputs. In this thesis the system is the EASN.

Target
An object, animal or person. This thesis defines three targets: guns, vehicles and humans. Although the word 'target' may suggest it need to be eliminated this is not necessary the case.

Tracking
A target is tracked when the tracker is focusing on the target. The tracker stores the previous target states and is able to estimate the next state.

Ultrasound
Also known as ultrasonic. Sound which cannot be heard by the human ear and/or an acoustic signal with frequencies above 20 kHz.

Vehicle
A mechanical object for carriage or transport.

Vibration
Acoustic signals through the ground or solid material.

Weighting
Data scaling according to a certain function.

Windowing
Weight the different signal samples with a window function, for example to make the signal periodic.

Wireless Multimedia Sensor Network
A sensor network where nodes are connected wireless. Furthermore, all the nodes can have different sensor types.

Abbreviations

Abbreviation	In full
ASN	Acoustic Sensor Network
EASN	Experimental Acoustic Sensor Network
ECT	Experimental Classification Tree
FE	Feature Extraction
FFT	Fast Fourier Transform
IE	Information extraction
ME	Mean Error
PSD	Power Spectral Density
RMSE	Root Mean Square Error
RPM	Revolutions Per Minute
SD	Standard Deviation
SN	Sensor Network
SNR	Signal to Noise Ratio
SP	Signal Processor
TDOA	Time Difference of Arrival
TOA	Time of Arrival
TOE	Time of Emission
WMSN	Wireless Multimedia Sensor Network

Symbols

Symbol	Definition	Unit
$A(d)$	Geometric spreading	1
$A_\alpha(d, \alpha_a)$	Atmospheric absorption	1
c_{air}	Sound velocity in air	m/s
d_i	Distance between the target and node i	m
$\hat{d}_i$	Estimated distance between the target and node i	m
$ES_i[f]$	Environment Spectrum for node i	W
$F_{gun}[p]$	Gun feature vector of the p^{th} partition	
$F_{human}[p]$	Human feature vector of the p^{th} partition	
F_s	Sample frequency	Hz
$F_{vehicle}[p]$	Vehicle feature vector of the p^{th} partition	
$G_i(f)$	Microphone gain at node i	1
$IP_{ip}[t]$	Instantaneous Power of $y_{ip}[t]$	W
L_{ip}	Length of $y_{ip}[t]$ in samples	1
$P_0(f,t)$	Received power at 1 meter	W
$P_{ei}(f,t)$	Power error at node i due to modelling and noise	W
p_i	Node i position	m^3
P_{ip}	Average power of $y_{ip}[t]$	W
$P_{Ri}(f,t)$	Received power at node i	W
$PS_{iw}[f]$	Power Spectrum of signal $y_{iw}[t]$	W
$PS_{iwa}[f_a]$	Aggregated Power Spectrum of signal $y_{iw}[t]$	W
$PSD_{iw}[f]$	Power Spectral Density of $y_{iw}[t]$	W/Hz
$PSP_{ipk}[f]$	Output k of Power Signal Processing for $y_{ip}[t]$	W

p_t	Target position	m^3
$\hat{p}_t$	Estimated target position	m^3
t	System time, all clocks are synchronized	s
T_{air}	Air temperature in Celsius	°C
T_{ei}	Time error at node i due to modelling and noise	s
T_{ip}	Length of $y_{ip}[t]$ in seconds	1
$T_{OAi}(k)$	Time of Arrival at node i	s
$T_{OE}(k)$	Time of Emission by the target	s
Tr_X	Threshold for process X	
T_s	Sample length in time	s
$TSP_{ipk}[t]$	Output k of Time Signal Processing for $y_{ip}[t]$	
T_{ti}	Travel time between target and node i	s
$W_X(x)$	Weight of x for process X	1
$y_i[t]$	Sampled signal from node i	V
$y_{ip}[t]$	Partition p of $y_i[t]$	V
$y_{ipw}[t]$	Windowed version of $y_{ip}[t]$	V
$\alpha_a(f)$	Air absorption coefficient	dB/m

References

[1] A. Ekimov, J.M. Sabatier, "Ultrasonic wave generation due to human footsteps on the ground", J. Acoust. Soc. Am., vol. 121, nr. 3, pp. 114-119, 2007.
[2] P. Atrey, N. Maddage, M. Kankanhalli, "Audio Based Event Detection for Multimedia Surveillance", Acoustics, Speech and Signal Processing, vol. 5, pp. 813-816, 2006.
[3] S. Ntalampiras, I. Potamitis, N. Fakotakis, "On acoustic surveillance of hazardous situations", IEEE International Conference on Acoustics, Speech and Signal Processing, icassp, pp.165-168, 2009.
[4] S. Shin, T. Hashimoto, S. Hatano, "Automatic Detection System for Cough Sounds as a Symptom of Abnormal Health Condition", IEEE transactions on IT in biomedicine, vol. 13, nr. 4, pp. 486-493, 2009.
[5] J. Whitaker and B. Benson, "Standard Handbook of Audio and Radio Engineering", P. Mc-Graw-Holl Professional, 2001.
[6] P.A. Tipler, G. Mosca, "Physics for Scientists and Engineers, Fifth Edition – Extended Version", P. USA: Freeman, 2004.
[7] A. Pierce, "Acoustics: An Introduction to its Physical Principles and Applications", New York: McGraw-Hill, 1989.
[8] R.C. Maher, "Acoustical Characterization of Gunshots". Signal Processing Applications for Public Security and Forensics SAFE '07, pp.1-5, issue: 11-13 April 2007.
[9] A. Averbuch, V. Zheludev, N. Rabin, "Wavelet based acoustic detection of moving vehicles". J. Multidimensional Systems and Signal Processing, vol. 20, nr. 1, pp. 55-80, 2009.
[10] M.E. Munch, "Bayesian Subspace methods for acoustic signature recognition of vehicles", P. Proc. of the 12th European Signal Processing Conf., Sept. 6-10, 2004.
[11] R.H. Mgaya, S. Zein-Sabatto, A. Shikhodaie, "Vehicle identifications using acoustic sensing", SoutheastCon, 2007. Proceedings. IEEE, pp. 555-560, Issue: 22-25 March 2007.
[12] M. Jing, C. Wang, L. Chen, "A real-time unusual voice detector based on nursing at home", Machine Learning and Cybernetics, vol. 4, pp. 2368-2373, Issue: 12-15 July 2009.
[13] A. Itai, H. Yasukawa, "Footstep Classification Using Wavelet Decomposition", International Symposium on Communications and Information Technologies, pp. 551-556, 2007.
[14] H. Wu, M. Siegel, P. Khosla, "Vehicle Sound Signature recognition by Frequency Vector Principal Component Analysis", Instrumentation and Measurement Technology Conference IMTC/98, vol. 1, issue: 18-21 May 2009.
[15] Y. Kim, D. Kom, S. Chung, "Target Classification in Sparse Sampling Acoustic Sensor Networks using DTWC Algorithm", The International Conference on Intelligent Pervasive Computing (IPC), pp. 236-241, 2007.
[16] A. Chancon-Rodrigues, P. Julian, "Evaluation of Gunshot Detection Algorithms", Micro-Nanoelectronics, Technology and Applications, pp.49-54, issue: 18-19 Sept. 2008.
[17] G. Valenzise, L. Gerosa, M. Tagliasacchi, "Scream and Gunshot Detection and Localization for Audio-Surveillance Systems", IEEE Conference on Advanced Video and Signal Based Surveillance (avss), pp.21-26, 2007.
[18] Y. Zhi-jun, Y. Guang-xin, W. Jian-ming, "Multi-model Rao-blackwellised particle filter for maneuvering target tracking in distributed acoustic sensor network". Acoustics, Speech and Signal Processing (ICASSP), vol. 3, issue 15-20 April 2007.
[19] Z. Yu, S. Dong, J. Wei, "Neural Network Aided Unscented Kalman Filter for Maneuvering Target Tracking in Distributed Acoustic Sensor Network". Proceedings of the International Conference on Computing: Theory and Applications (ICCTA), pp.245-249, 2007.
[20] K. Tumer, J, Ghosh, "Estimating the Bayes Error Rate through Classifier Combining", 13th International Conference on Pattern Recognition (ICPR), vol. 2, pp.695, 1996
[21] R. Rojas, "Neural Networks - A systematic Introduction", Springer-Verlag, Berlin, New-York, 1996.
[22] J. Lee, I. Oh, "Binary Classification Trees for Multi-class Classification Problems", Seventh International Conference on Document Analysis and Recognition (icdar), vol. 2, pp.770, 2003.
[23] L.B. Evans, H.E. Bass, L.C. Sutherland, "Atmospheric absorption of sound: theoretical predictions". J. Acoust. Soc. Am., vol. 72, pp. 1565-1575, 1972.
[24] L.B. Evans, H.E. Bass, "Tables of absorption and velocity of sound in still air at 68F". Report WR72-2, Wyle Laboratories, Huntville Ala, 1972.
[25] H.E. Bass, L.C. Sutherland, A.J. Zuckerwar, "Atmospheric absorption of sound: Further developments", J. Acoust. Soc. Am., vol. 97, issue 1, pp. 680-683, Jan. 1995.

Appendices

A Experimental hardware analysis

The equipment which is used for the Experimental ASN (EASN) is a professional four input audio card and four very small microphones with windjammers.

The sound card is the M-Audio Delta 44 - Professional 4-In/4-Out Audio Card. This is a external box connected to a PCI host adapter card, with a maximum sample frequency of 96 kHz with 24 bit and dynamic range of 99 dB. For the gunshot measurements the Creative Sound Blaster X-Fi Surround 5.1 soundcard is used. The Creative soundcard has similar specifications, but it allows more flexible use (usb connection to a laptop). However the Creative soundcard has only two record inputs. The M-Audio soundcard will be further discussed.

The FG-3329 microphones from Knowles electronics, which are omni-directional according to their specifications, are used for the experiment. Each microphones has its own amplifier, which is internally developed by Thales. It is known that the amplifiers amplify with a factor of 20 and that the internal noise is proportional to $1/f$.

The microphones need a windjammer for outside measurements. The cap and windjammer are from Raycote Lavaliere and the cap is made from foam and the jammer from fur, which is a artificial fibre fixed to a fabric mesh backing. Figure A-1 shows the microphone, amplifier, cap and windjammer.

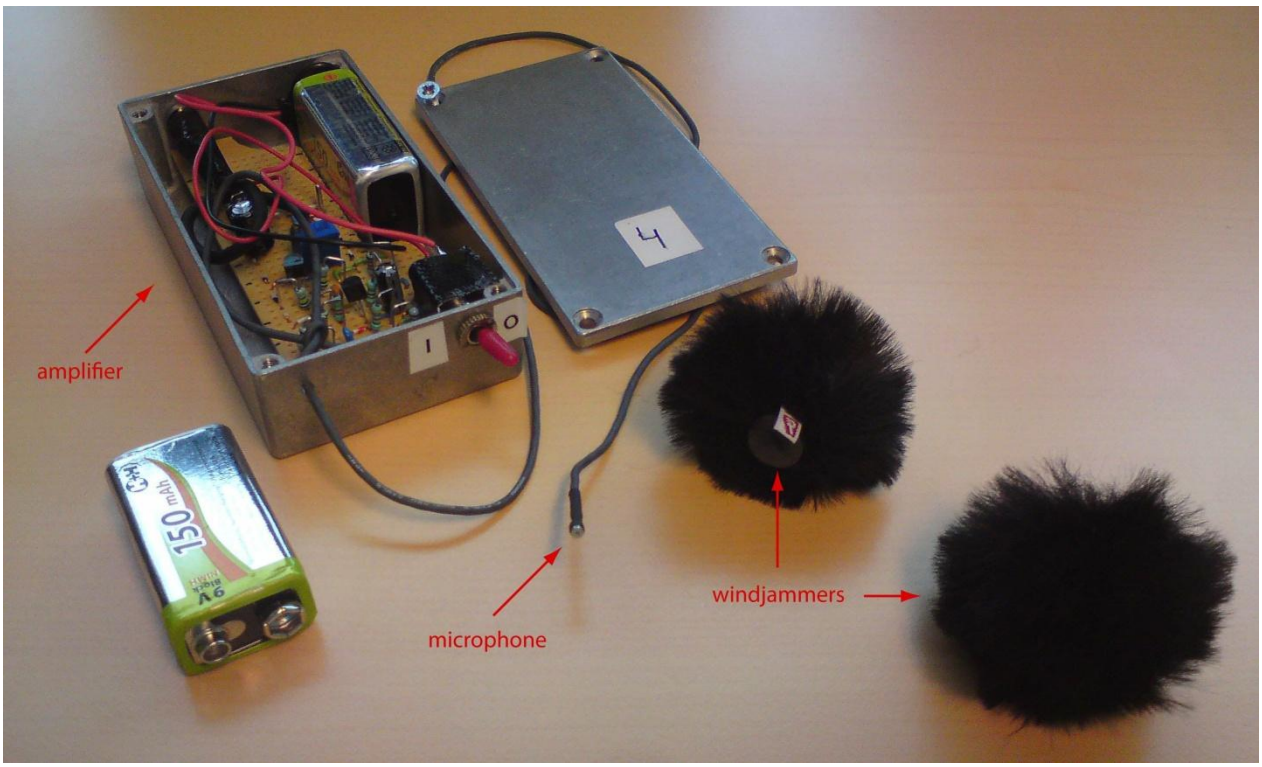


Figure A-1: FG-3329 microphone, windjammers and amplifier

The gunshots produce very strong signals and therefore the signal has to be attenuated. This is done with a yellow-ear-damper. A gap was made in the ear-damper and slid over the microphone. The standard and ear-damper-windjammer are shown in Figure A-2.

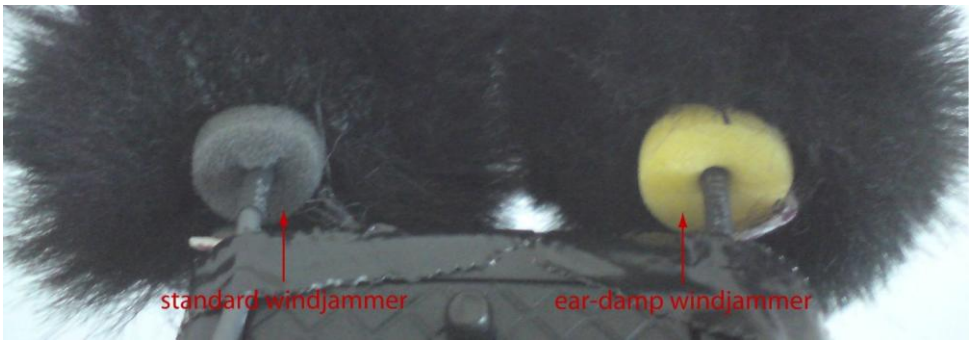


Figure A-2: The standard and ear-damp windjammer

Multiple measurements are done with the equipment and in the urban environment. Firstly, soundcard measurements without any microphones. Secondly, measurements with microphones. And thirdly, measurements in and outside and with and without windjammer. For the first two measurements the sample duration T_s was 10s and the sample frequency F_s was 96kHz.

Figure A-3 shows the PSD of the different soundcard inputs when the inputs are closed. As the plot shows, the different inputs have different internal distortions. The spectrum is not flat and the inputs have different peaks at different frequencies. The result of open inputs in comparison to closed inputs is not very different, but in theory an open port could better receive electromagnetic waves.

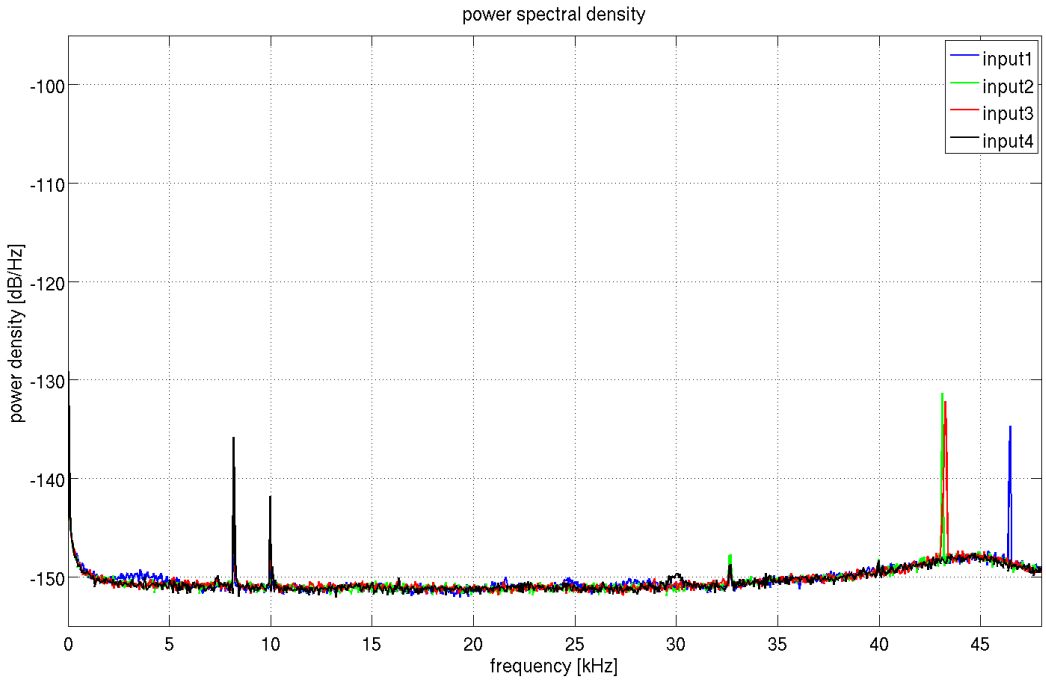


Figure A-3: PSD of the four closed soundcard-inputs

In Figure A-4 the four different microphones with amplifiers are measured on input1 inside the Thales Delft building (lab) after the air-conditioning has been shut down. However, still some signals were present and the most prominent was the signal in the frequency band of 25kHz till 35kHz. This ultrasound was not present outside, thus it can be concluded that the distortion was not from the equipment. The microphones differ in this situation at most 4dB and mainly in the 1kHz till 13kHz (important) frequency band.

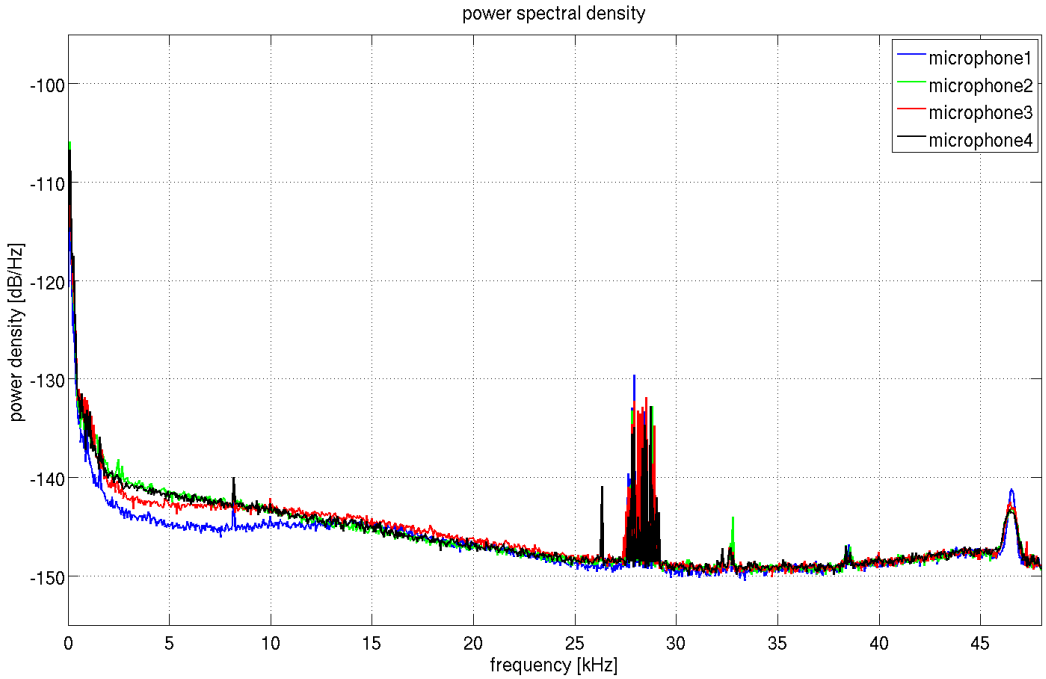


Figure A-4: PSD of the four microphones inside at Thales Delft without air-conditioning

Another microphone measurement was done to obtain an indication of the microphone gain as a function of frequency. All the four microphones where placed together and a speaker which emitted a pseudo white noise signal was placed at a distance of 30 centimetres. The sample duration T_s was 30 seconds and the sample frequency F_s was 96kHz. This experiment does not show the exact gain of the microphones, because a speaker, cables and a distance is included, but it shows the difference between the microphone gains. The PSD of the received signals of the four microphones is shown in Figure A-5. Although the batteries were al freshly charged, the spectra show that a gain difference of 3dB and even 6dB is present at certain frequencies between the microphones.

Measurement date is 29 March 2010, little windy outside (~5 m/s) and at approximately hundred meters cars were moving. Figure A-6 shows multiple PSDs of indoor measurement at Thales Delft and outdoor measurements on the Thales Delft parking area. The sample duration T_s was 3 seconds and the sample frequency F_s was 96kHz. One of the things the plot shows, is the difference between in and outside and with and without a windjammer. Although the windjammer reduces the wind noise, it also results in an attenuation of a wide-band signal (hitting hands together). In other words, the windjammer results in a reduction of the wind noise at the cost of signal power. Although loss of signal power is not desirable, wind noise reduction

in urban environments is needed. Therefore, in further outdoor measurements the windjammer will be used.

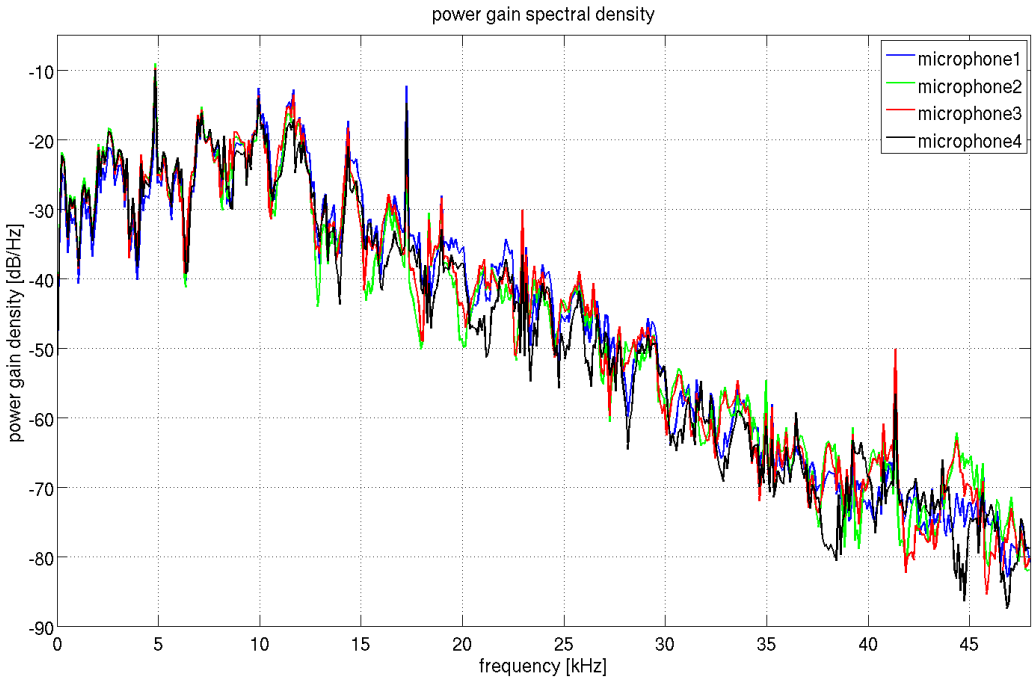


Figure A-5: PSD of the four microphones with white noise source

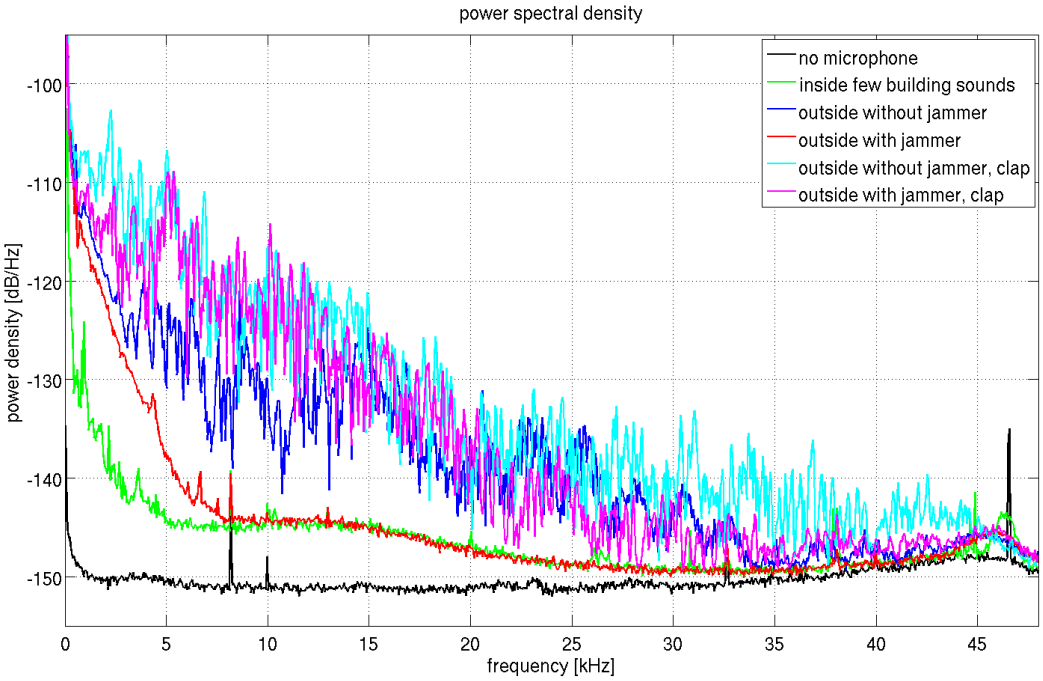


Figure A-6: PSD of different measurement situations

B Signal Processing

B.1 Power calculations

To obtain the *instantaneous power* from signal $y_{ip}[t]$ the following has to be done:

$$IP_{ip}[t] = \frac{{y_{ip}}^2[t]}{R} \tag{32}$$

Where R is the electrical resistance and if it is assumed to be one, then the *average power* is:

$$P_{ip} = \frac{1}{L_{ip}} \sum_t {y_{ip}}^2[t] \tag{33}$$

Where P_{ip} is the mean square of signal $y_{ip}[t]$ and L_{ip} is the length of signal $y_{ip}[t]$.

To transform $y_{ip}[t]$ to the frequency domain the Fast Fourier Transform (FFT) can be used. Before the FFT can be applied on a signal, the signal has to be weighted by a window that is not rectangular to avoid spectral leakage. Which window to choose is a difficult decision and depends on the target, but normally the Hanning window is used. To construct the *Power Spectrum* (PS) of a received signal the following is done:

$$PS_{ipw}[f_{double}] = \left| \frac{FFT\{y_{ipw}[t]\}}{L_{ipw}} \right|^2 \tag{34}$$

Where $y_{ipw}[t]$ is the windowed version of $y_{ip}[t]$. The above PS has a double-side spectrum. For a single-side spectrum $PS_{ipw}[f]$ the first half of $PS_{ipw}[f_{double}]$ is taken and multiplied by two. The result of the FFT is divided by L_{ipw} so the amplitudes of $PS_{ipw}[f]$ are not depend on the length of $y_{ipw}[t]$. The $PS_{ipw}[f]$ shows how the power is distributed over the frequencies, and thus:

$$P_{ipw} = \sum_f PS_{ipw}[f] \tag{35}$$

When the $y_{ipw}[t]$ signal is very long, the length of $PS_{ipw}[f]$ will also be long. To reduce the length and frequency points of $PS_{ipw}[f]$ the amplitudes can be aggregated. The *power aggregation* of $PS_{ipw}[f]$ is the power summation of multiple bins and place the result in a wider new bin.

The *Power Spectral Density* (PSD) of the signal $y_{ipw}[t]$ is given by:

$$PSD_{ipw}[f] = \frac{PS_{ipw}[f]}{F_{bin}\{PS_{ipw}[f]\}} \tag{36}$$

Where $F_{bin}\{PS_{ipw}[f]\}$ is the bin width of $PS_{ipw}[f]$ in Hertz. The amplitudes of the PSD are power aggregation independent, because more aggregation means a higher $F_{bin}\{PS_{ipw}[f]\}$ value. The PSD is usually plotted in logarithmic scale is:

$$PSD_{ipw}[f]\ \ (dB) = 10\log_{10}(PSD_{ipw}[f]) \tag{37}$$

The *Short-term Power Spectral Density* (SPSD) is created by calculating the PSD for multiple partitions.

B.2 Data weighting

Weighting of information, which is not always the same as signal windowing, is needed in the design process. Data weighting will not be discussed deeply, but a general approach is discussed to illustrate the process. Many functions can be constructed to weight, but three basic functions, uniform, linear and exponential, are suggested to weight real positive data x:

$$w_{uni}[x_k] = \begin{cases} 1 & \text{if}\{x_{start} \le x_k \le x_{stop}\} \\ 1 & \text{if}\{x_{stop} \le x_k \le x_{start}\} \\ 0 & \text{else} \end{cases} \tag{38}$$

$$w_{lin}[x_k] = \begin{cases} 1 - [x_{scale}\,(x_k - x_{start})/(x_{stop} - x_{start})] & \text{if}\{x_{start} \le x_k \le x_{stop}\} \\ 1 - [x_{scale}\,(x_{start} - x_k)/(x_{start} - x_{stop})] & \text{if}\{x_{stop} \le x_k \le x_{start}\} \\ 0 & \text{else} \end{cases} \tag{39}$$

$$w_{exp}[x_k] = \begin{cases} \exp[-x_{scale}\,(x_k - x_{start})/(x_{stop} - x_{start})] & \text{if}\{x_{start} \le x_k \le x_{stop}\} \\ \exp[-x_{scale}\,(x_{start} - x_k)/(x_{start} - x_{stop})] & \text{if}\{x_{stop} \le x_k \le x_{start}\} \\ 0 & \text{else} \end{cases} \tag{40}$$

Where x_k is the k^{th} component of data x. The constant variables x_{start}, x_{stop} and x_{scale} can be chosen by the designer. x_{start} can be smaller or larger then x_{stop}, but the weight of data x_k will be zero if x_k has no value between x_{start} and x_{stop}. After the weights are constructed, the designer can choose to be normalize the weights as follow:

$$W_{max}[w_k] = w_k/\max_n\{w_n\} \tag{41}$$

$$W_{sum}[w_k] = w_k/\sum_n w_n \tag{42}$$

B.3 Peak finding

In the design process, multiple times an algorithm is needed to find peaks in a signal. To find peaks is to find maxima in local areas. Thus if peaks has to be found in a signal $y_i[t]$, the signal is first partitioned into $y_{ip}[t]$ with 0.5 overlap ratio and the partition size determines the local area size. For every partition a maximum is recognized as a peak, if the maximum is in the middle, which size is 0.5 of the partition, and the maximum is higher than a certain threshold. The threshold can for example be fixed or a function of the mean and standard deviation of $y_i[t]$. An example of finding peaks is shown in Figure A-7.

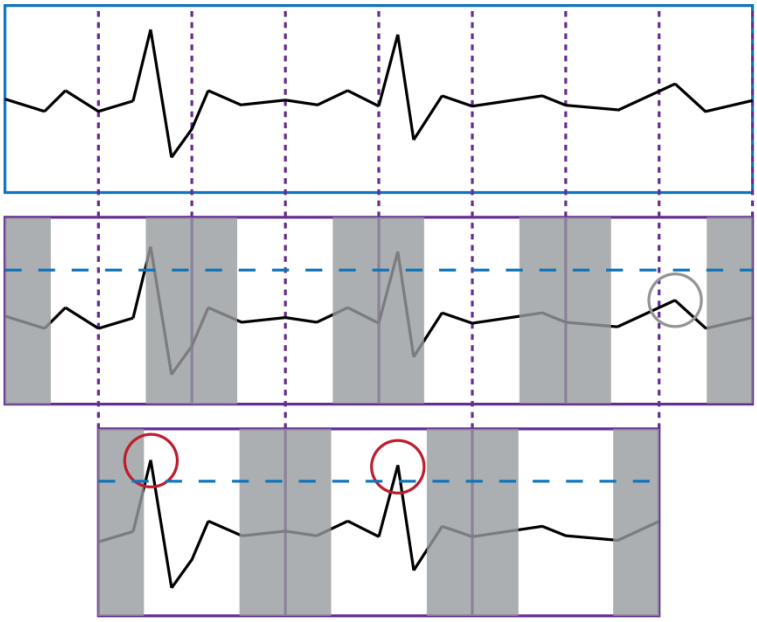


Figure A-7: Peak searching and finding

C Wavelet Analysis

Instead of using the harmonics, wavelets[1] can be used to analysis the signal. The Wavelet Analysis has an infinite set of possible (wavelet) basis functions. A wavelet is a waveform of effectively limited duration and the mean is zero. The Daubechies family wavelets are well known in the wavelet research and are shown in Figure A-8. Sine waves, which are used in Fourier Analysis, do not have a limited duration and are smooth and predictable. However, wavelets are irregular and asymmetric. The wavelet basis functions are scale varying, which means that the wavelet basis functions are self-similar: scaled in time to maintain the same number of oscillations and scaled in amplitude to maintain energy.

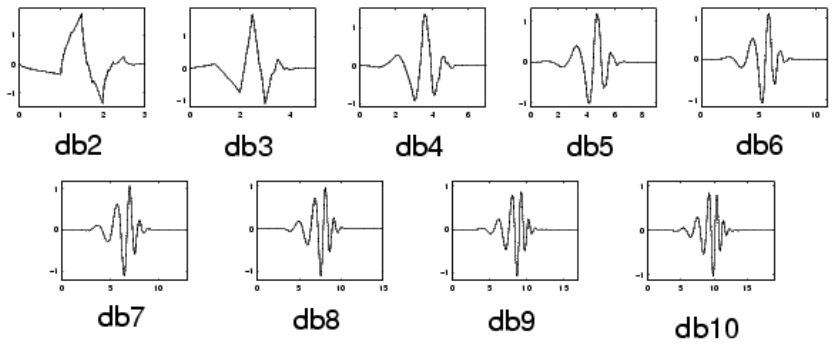


Figure A-8: Daubechies family wavelets, source: MATLAB Wavelet Toolbox documentation

The most interesting dissimilarity between Wavelet Analysis and Fourier Analysis is the time-frequency resolution[2]. With the STFT a spectrogram is created. With the Wavelet Analysis a similar picture can be made, but the frequency is replaced with the scale. An illustration is shown in Figure A-9. The Short-term Fourier Transform (STFT) has a fixed window size, but Wavelet Analysis allows the use of variable window sizes. For low frequencies long time windows can be used and for high frequencies short time windows can be used. Instead of transforming to time-frequency in STFT, Wavelet Analysis is transforming the signal to time-scale. The scale is connected to frequency in the following way: low scale wavelets, contain mainly high frequencies and vice versa.

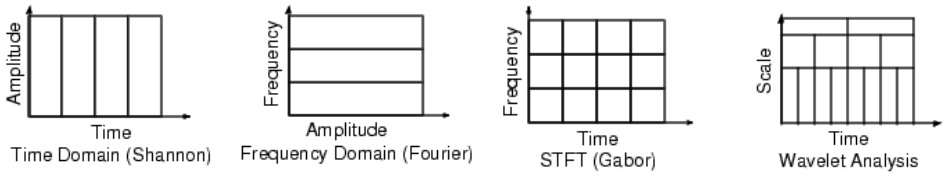


Figure A-9: Wavelet Analysis, source: MATLAB Wavelet Toolbox documentation

The Wavelet Transform (WT) derives amplitude for every scale and time of the basis function. On the one hand you would like to have some very short basic functions, to describe the fast signal changes. On the other hand you would like very long basis function. A way to achieve this is to have short high-frequency basis functions and long low-frequency ones. This is exact what

[1] A good introduction to Wavelets is provided in the MATLAB Wavelet Toolbox
[2] L.A. Barford, R.S. Fazzio, D.R. Smith, "An Introduction to Wavelets". Instrument and Photonics Laboratory, 1992.

the wavelet transform provides. A short comparison between the wavelet and Fourier transform is showed in Table 9.

Table 9: Differences between Fourier Analysis and Wavelet Analysis

	Fourier Analysis	Wavelet Analysis
Amplitude of each time is transformed to	Amplitude and phase for each frequency	Amplitude for each scale and time
Localization in frequency/scale	Yes	Yes
Localization in time	Limited (with STFT)	Yes

Because the WT is so powerful in localization in time it is a good method to analyze non-stationary signals. For every scale there is another frequency and time resolution. It is like the WT bridges the gap between time-domain and frequency-domain representations of a signal[3]. In contrast, the STFT is limited in both time and frequency resolution by the fixed width of its window.

C.1 Continues Wavelet Transform (CWT)

After the wavelet scales are chosen, each scaled wavelet is shifted over the time signal and the correlation coefficient is continued (for every shift/sample step) determined. Thus, for every scale wavelet coefficients are calculated. Multiplying each coefficient by the appropriately scaled and shifted wavelet describes the original signal. Two steps are shown of correlation coefficient calculation in Figure A-10.

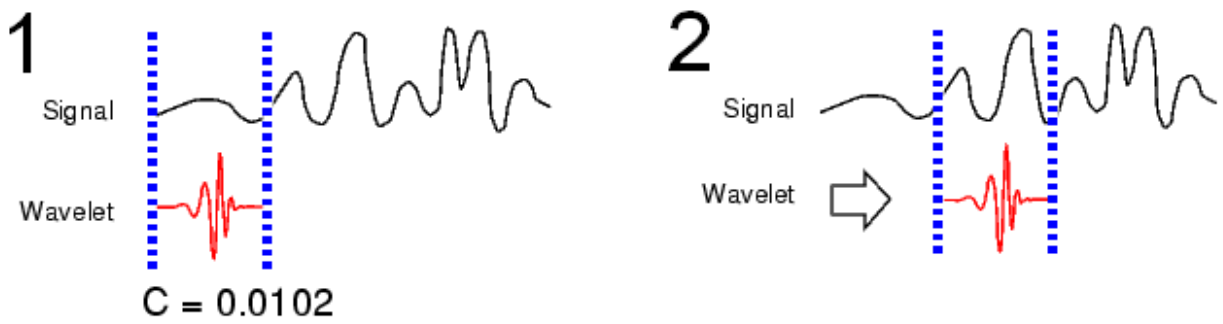


Figure A-10: Continues Wavelet Transform, source: MATLAB Wavelet Toolbox documentation

This process is done for every scale, however the amount of coefficient for every scaled wavelet is the same and thus the same time resolution, because the wavelets were continues shifted over the signal.

[3] A. Graps, "An Introduction to Wavelets". IEEE Computational Science and Engineering, vol. 2, nr. 2, 1995.

C.2 Discrete Wavelet Transform (DWT)

Instead of calculating the coefficient at every possible scale, a subset of scales and positions can be chosen. It turns out, rather remarkably, that if the scales and positions are chosen as a power of two, the analysis will be much more efficient and just as accurate[4]. Filters can be used to implement DWT as shown in the first part of Figure A-11.

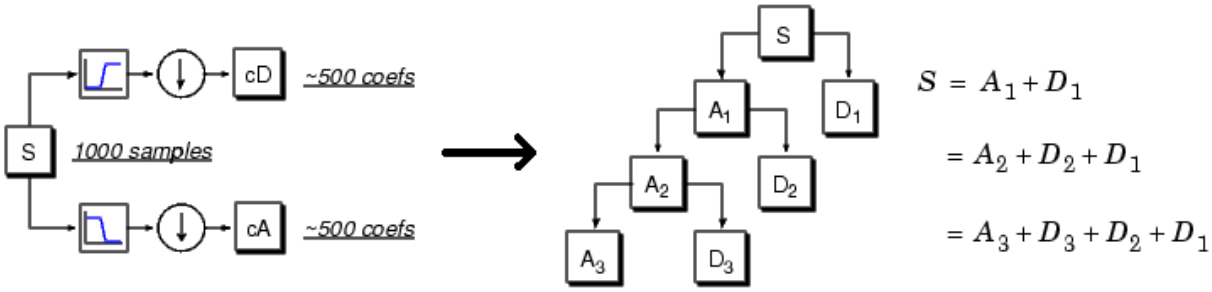


Figure A-11: Discrete Wavelet Transform, source: MATLAB Wavelet Toolbox documentation

The signal is split with filters and then down sampled, which result in an approximation and a detail part. The length of the original signal is the same as the sum of the length of the approximation and the detail. After one operations, again the same procedure can be applied on the approximation. This is called multiple-level decomposition, as shown in the second part of Figure A-11.

C.3 Wavelet Packet Transform (WPT)

WPT is the same as DWT, but in WPT the details are also decomposed. An illustration is given in Figure A-12.

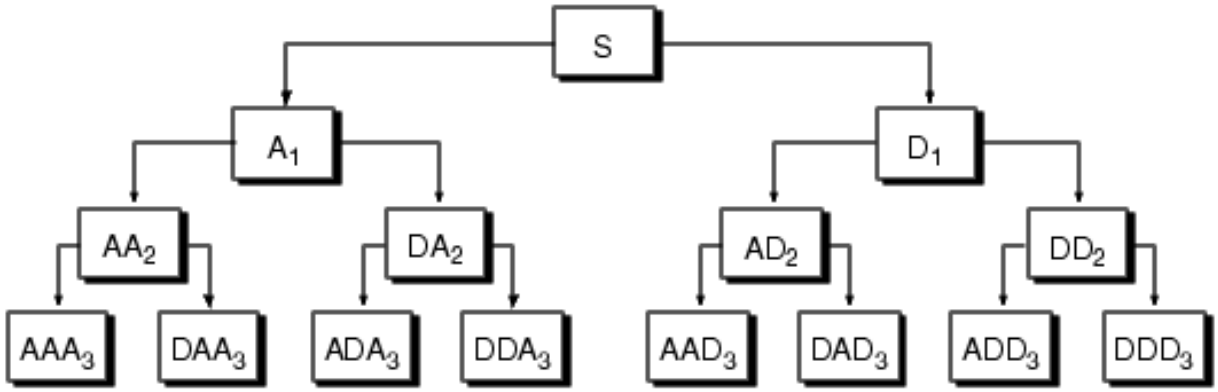


Figure A-12: Wavelet Packet Transform, source: MATLAB Wavelet Toolbox documentation

For instance, WPT allows the signal S to be represented as A1 + AD2 + ADD3 + DDD3. Algorithms exist for both wavelet packet decomposition and optimal decomposition selection.

[4] Discrete Wavelet Transform - The MathWorks, http://www.mathworks.com/access/helpdesk/help/toolbox/wavelet/ch01_i11.html, consulted: 28 January 2010.

D De Groot Fourier Transform

The *De Groot Fourier Transform* (DGFT) is a simple method which introduces variable time and frequency resolution in de spectrogram analysis. Wavelet Analysis is proclaiming to have variable time and frequency resolution, which seems to be an advantage for certain applications. However, DGFT is providing this same advantage in the Fourier Analysis. This paragraph is an introduction to the DGFT.

The STFT is easy understandable and uses the harmonics, but it has a fixed time and frequency resolution depending on the chosen window size. When the window size is decreased, the time resolution increases, but the frequency resolution decreases. Ideally, a method is required to give low frequencies components a wide window, but high frequencies a narrow window. DGFT is providing low frequencies large windows and high frequencies small windows.

DGFT is a method with two variables: $groot$ and $power$. The name 'De Groot' is Dutch and can be translated to 'the great'. The name 'De Groot' is due to the fact that the DGFT is using a large set (equal to $groot$) of standard STFTs and also because it is the last name of the author.

The two values $groot$ and $power$ can be chosen freely, depending on the application and goals of the analyzer. In Figure A-13 an example construction of the DGFT is illustrated. As it is shown, DGFT is based on multiple STFT, which result in that higher frequencies get a higher time resolution. STFT can still be constructed with a certain window length, window weights and window overlap.

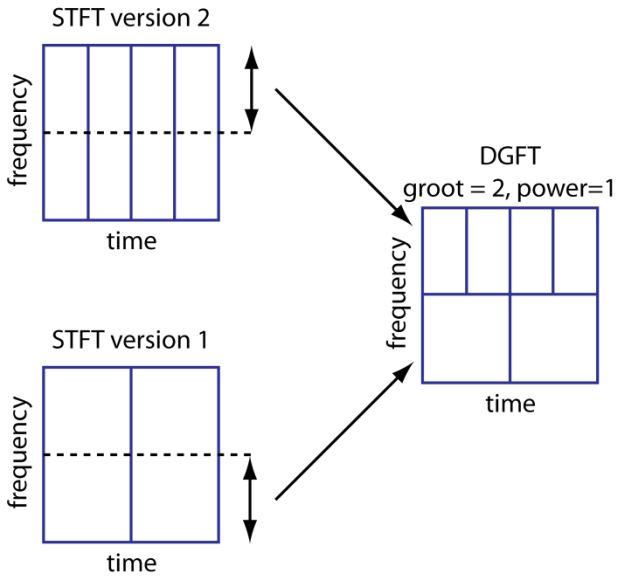


Figure A-13: Construction example of the DGFT

The idea is thus: construct multiple STFT with different window sizes and cut the right frequency part of the different results. The selection of the right frequencies can also be seen as filtering.

The part length is defined as:

$$L_{part}(n) = \frac{power^{n-1}}{\sum_{k=1}^{groot} power^{k-1}} \qquad n = 1,2,\ldots,groot \qquad (43)$$

Where n is an integer from one to $groot$. Therefore the top frequency of a certain part is given by:

$$F_{top}(n) = \frac{F_s}{2}\sum_{k=1}^{n} L_{part}(k) \qquad n = 1,2,\ldots,groot \qquad (44)$$

When the different STFT are combined, the low frequency of part n is:

$$F_{low}(n) = \begin{cases} 0 & if \quad n = 1 \\ F_{top}(n-1) & else \end{cases} \qquad n = 1,2,\ldots,groot \qquad (45)$$

The time part is given by:

$$T_{part}(n) = \begin{cases} \frac{T_{part}(1)}{n} & if \quad power = 1 \\ \frac{T_{part}(1)}{power^{n-1}} & else \end{cases} \qquad n = 2,3,\ldots,groot \qquad (46)$$

Where $T_{part}(1)$ is chosen by the analyser. The overlap ratio remains the same over all the STFT, thus the time overlap is given by:

$$T_{overlap}(n) = T_{part}(n) \cdot R_{overlap} \qquad n = 1,2,\ldots,groot \qquad (47)$$

Couple of examples without overlap are given in Figure A-14.

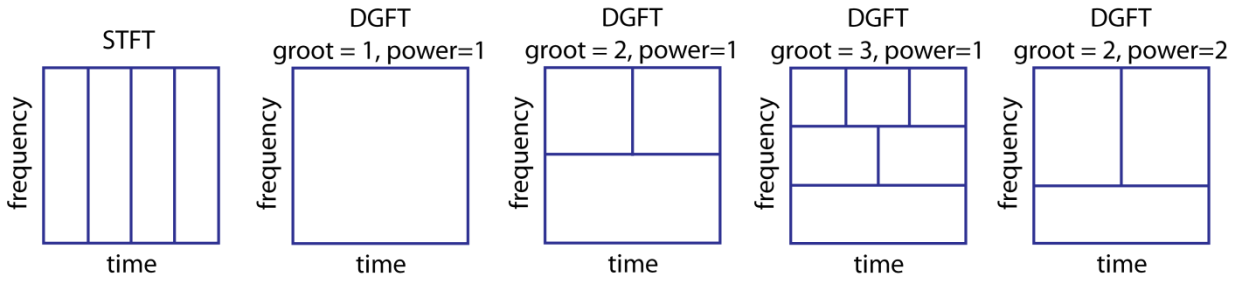


Figure A-14: DGFT examples

In Figure A-15 an example signal is shown in time-domain and with a spectrogram (STFT). Sample frequency is 96kHz of a duration of 0.2 second. Figure A-16 shows the same signal, but then analyzed with the DGFT.

Thus, DGFT brings a variable time and frequency resolutions to the Fourier Analysis and DGFT can be seen as a flexible STFT. DGFT is a strong method, because the different frequency components get the window size which they "deserve".

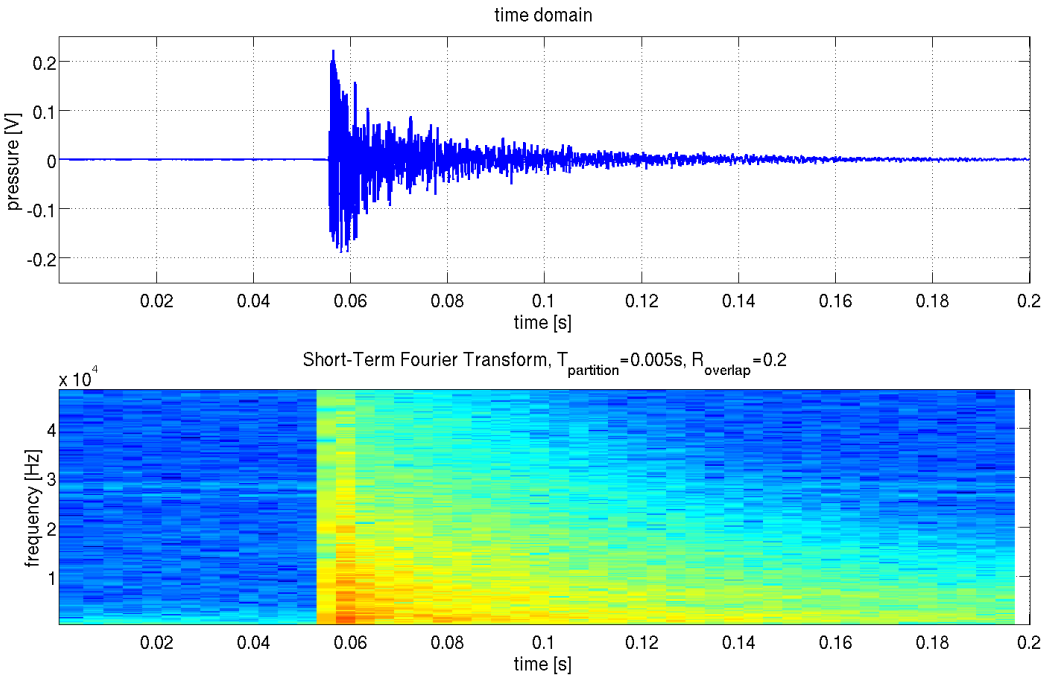


Figure A-15: Acoustic toy-gunshot signal in time domain and spectrogram

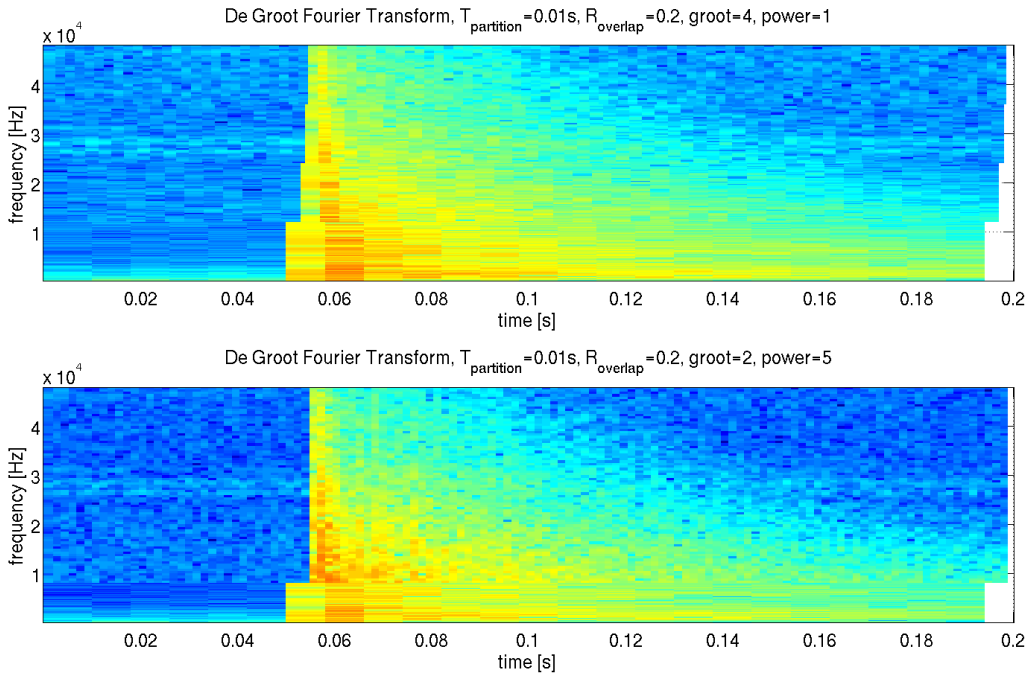


Figure A-16: Acoustic toy-gunshot signal analysed with DGFT

Although the author of the DGFT did not use it for his master graduation project, it could maybe useful for certain specific applications, for example compression techniques.

E Classification

E.1 Minimum Distance (MD)

The MD classifier, also known as nearest neighbor, is searching for the training feature, which is at minimum distance from the input feature. The feature distance function can be defined in many ways, but is mostly defined as the root of the sum of the squared distances:

$$d(F_1, F_2) = \sqrt{\sum_i^N (F_1(\mathrm{i}) - F_2(\mathrm{i}))^2} \tag{48}$$

Where N is the length of the feature vectors F_1 and F_2. There are other functions available, but when a distance function is chosen, the MD classification problem for features F_x is defined as:

$$C = \arg\left[\min_j \left(d(F_x, F_j)\right)\right] \tag{49}$$

Multiple techniques are possible to find the minimum distance, but they are all very exhausting.

E.2 Classification Tree (CT)

The CT is a well-known method[5]. In most general terms, a CT has multiple if-then logical (split) conditions, together they form a tree. The goals is to determine a set of if-then logical (split) conditions that permit accurate class prediction (classification).

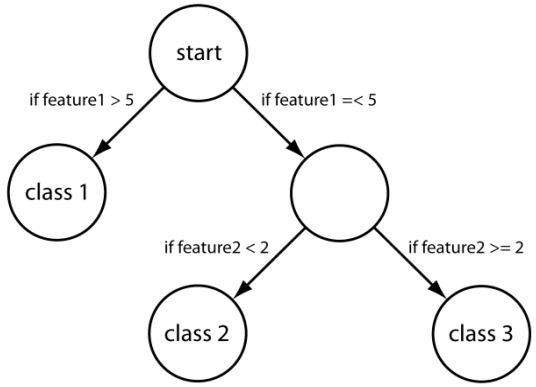


Figure A-17: Classification Tree

An CT example is given in Figure A-17. The CT is very clear in its result estimation. The interpretation of results summarized in a tree is very simple. This simplicity is useful not only for rapid classification, but it also explains why certain observations are classified in a particular manner.

[5] L. Breiman, J.H. Friedman, R.A. Olshen, "Classification and Regression Trees", Chapman & Hall, Inc., New York, 1993.

E.3 k-Nearest Neighbor (k-NN)

The k-Nearest Neighbor (k-NN) is a simple classification method[6]. The method searches for k nearest training features in respect to the input feature. These k training features all have a class assigned. The most present class will be the result of the k-NN classifier. An example of a 4-NN classifier is shown in Figure A-18.

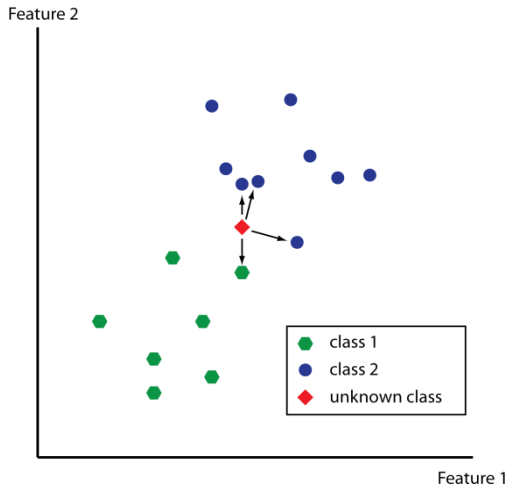


Figure A-18: Example of a k-NN (k=4) classifier

An 1-NN is thus actually the same as the MD classifier. The k is usually chosen to be odd to avoid ties. The main drawback in k-NN is the majority voting: when many training features of a certain class are present, this class will mostly win. To avoid this problem the same number of training features has to be present from every class.

If MD is computationally intensive, then k-NN is even more computationally intensive. Especially when the size of the training set grows. There are algorithms proposed to reduce this problem[7].

[6] K-nearest neighbor, http://www.scholarpedia.org/article/K-nearest_neighbor, consulted: 5 February 2010.

[7] V. Garcia, E. Debreuve, "Fast k Nearest Neighbor Search using GPU". Universite de Nice-Sophia, France.

E.4 Neural Network (NN)

Neural Network[8] theory is inspired by the biological nervous systems. NNs can be used for recognizing patterns and are composed of multiple artificial neurons operating in parallel. A neuron model is shown in Figure A-19.

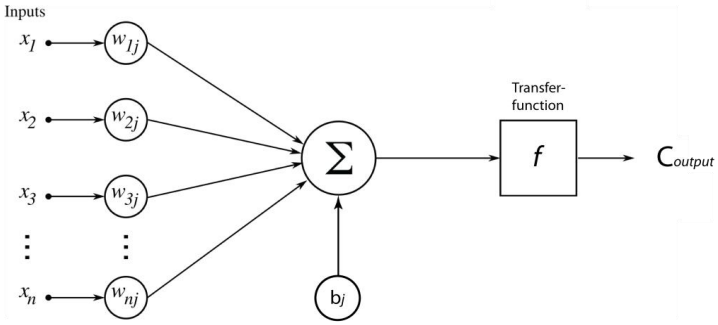


Figure A-19: Neuron model

The neuron sums all the weighted inputs, add a bias and gives it to a transfer function, which gives the output. The formula:

$$C = f\{b + \sum_i w_i x_i\} \tag{50}$$

The bias can also been seen as a fixed weighted input. The model can have many variants. For example multiple transfer function can be chosen or weights can be negative. Examples of transfer functions are shown in Figure A-20.

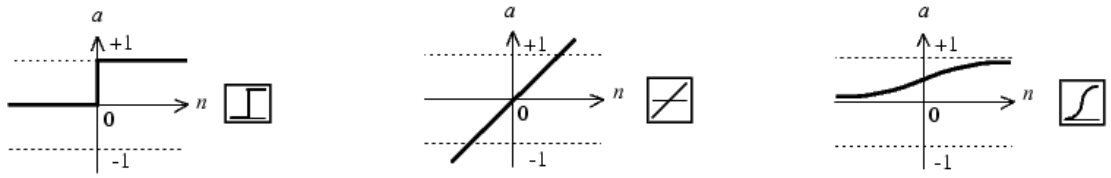


Figure A-20: Three transfer functions: Hard-Limit, Linear and Log-Sigmoid

In fact, one neuron can be a classifier: it transforms an input to an output. Combining multiple neurons results in a network of neurons: an artificial NN. Normally a NN is divided in three layers: input layer, hidden layer(s) and output layer. Figure A-21 shows the layers of a NN.

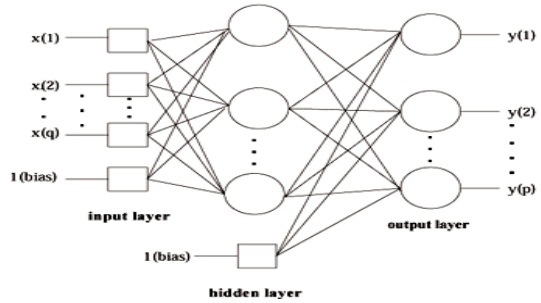


Figure A-21: Input, hidden and output layers

[8] A good introduction to Neural Networks is provided in the MATLAB Neural Network Toolbox

This NN can form a classifier, because it transforms an input to an output. The output of NN can be calculated in a static and dynamic way. When the output is static calculated, the input vector is not changing and treated as fixed. When the NN contains delays, the input will be a sequence of input vectors that occur in a certain time order and then the output is dynamic calculated.

The main problem in NNs is how to choose the weights and the transfer functions. When the parameters of the NN are derived, this is called learning or training. NNs are mostly trained supervised in such a way that a particular input leads to a specific output. NNs can also be used for clustering in an unsupervised learning process.

NNs can perform complex functions in various situation. There is much research in weight estimation in NNs, and multiple methods have different names. However in the end it comes to: train a NN in such a way that it converts an input to a correct output.

E.5 Gaussian Mixture Model (GMM)

Another approach for classification is modelling the features with the GMM[9]. With GMMs every feature in the feature vector is assumed Gaussian distributed[10]. In Figure A-22 an example is shown of two classes in a two dimensional feature space modelled with GMM. For the clusters the mean and the variance is obtained.

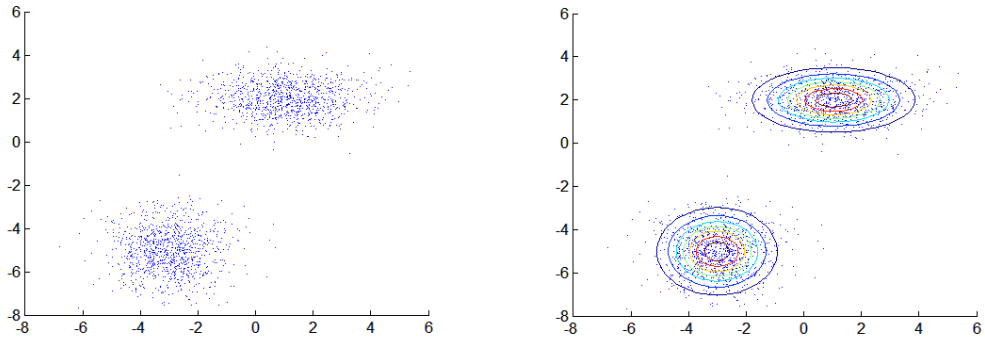


Figure A-22: Modelling with Gaussian Mixture Model, source: MATLAB Statistics Toolbox documentation

GMM can also be used to discover classes in data by forming clusters in an unsupervised way. An objective function is defined to assess the quality of clustering and optimize the objective function.

GMM classification works in the following way: for a certain class the feature vector is modelled with the feature training set. After the feature vector of a certain class is modelled, the probability can be calculated that a new feature vector is from the same class. Thus, if for every class the feature vector is modelled, the likelihood that a new feature vector is from a certain class can be calculated for every class.

[9] A short introduction to Gaussian Mixture Models is provided in the MATLAB Statistics Toolbox
[10] Also Mixture Models exists, which do not assume a Gaussian distribution

F Detection

A detector decides if a target is present or not. A detector is eventually a classifier which classify between present of absent. Frequently, but not necessary, the result of a detector is one or zero, present or not. This is called a binary classifier, a classifier with two classes. Also a probability value can be given.

When discussing *detectors*, it has to be clear what the detector detects. This is required to talk about detectors. On the other hand, when talking about detectors, mostly a non-noise detector is meant. However it remains unclear, because mostly noise is also not defined.

An *abnormal detector* detects if there is something unusual. It could work in the following way: it first learns about the environment, and when it detects a suddenly change it detects an abnormal event. In fact it is a normal versus abnormal classifier.

An *target detector* detects if the defined target is present or not. A target detector is a target versus other classifier, where the result can be target detected or no target detected.

Multiple detectors can be combined and placed together to form a larger system. Splitting the detection problem in multiple detectors/classifiers can improves the (accuracy) performance of the system. For example, the signal is first scanned with an abnormal detector. When an abnormal event is detected the vehicle, person or gunshot detector is initialized. The combination of detectors can also be seen as a tree classifier.

G Recording filenames of the plots

Table 10: Link between MATLAB plots and recording filenames

Figure	**Filename**
Figure 2-2 & Figure 2-3	17-May-2010_21-39-03.mat
Figure 2-4 & Figure 2-6	17-May-2010_21-45-41.mat
Figure 2-5 & Figure 2-7	17-May-2010_23-06-22.mat
Figure 2-9 & Figure 2-10 & Figure 2-13	19-May-2010_14-14-47.mat
Figure 2-11 & Figure 2-14	19-May-2010_14-19-40.mat
Figure 2-12 & Figure 2-15	19-May-2010_14-25-23.mat
Figure 2-17	30-Mar-2010_15-10-00.mat 30-Mar-2010_15-12-22.mat 30-Mar-2010_15-20-14.mat 30-Mar-2010_15-21-54.mat
Figure 2-18	30-Mar-2010_15-10-00.mat 30-Mar-2010_15-12-29.mat 30-Mar-2010_15-14-06.mat 30-Mar-2010_15-15-31.mat 30-Mar-2010_15-16-58.mat 30-Mar-2010_15-18-30.mat
Figure 2-19	30-Mar-2010_15-10-00.mat 30-Mar-2010_15-12-40.mat 30-Mar-2010_15-14-19.mat 30-Mar-2010_15-15-51.mat 30-Mar-2010_15-17-18.mat 30-Mar-2010_15-18-47.mat
Figure 2-20	30-Mar-2010_15-10-00.mat 30-Mar-2010_15-12-29.mat 30-Mar-2010_15-15-31.mat 30-Mar-2010_15-18-30.mat
Figure 2-21	30-Mar-2010_15-10-00.mat 30-Mar-2010_15-12-40.mat 30-Mar-2010_15-15-51.mat 30-Mar-2010_15-18-47.mat
Figure 2-22	30-Mar-2010_15-14-41.mat
Figure 2-24	30-Mar-2010_15-25-14.mat
Figure 2-25	30-Mar-2010_15-25-45.mat
Figure 2-26	30-Mar-2010_15-27-03.mat
Figure 2-27	30-Mar-2010_15-28-54.mat
Figure 2-28	30-Mar-2010_15-29-26.mat
Figure 2-29	11-Mar-2010_15-43-26.mat
Figure A-3	07-Apr-2010_12-10-45.mat 07-Apr-2010_12-11-12.mat 07-Apr-2010_12-14-47.mat 07-Apr-2010_12-22-11.mat
Figure A-4	30-Apr-2010_18-11-41.mat

	30-Apr-2010_18-12-05.mat 30-Apr-2010_18-12-29.mat 30-Apr-2010_18-13-04.mat
Figure A-5	MicrophoneGain.mat
Figure A-6	29-Mar-2010_11-53-24.mat 29-Mar-2010_10-59-18.mat 29-Mar-2010_11-06-27.mat 29-Mar-2010_11-08-10.mat 29-Mar-2010_11-08-55.mat

H Experimental MATLAB code structure

The global MATLAB code structure is outlined in Table 11.

Table 11: Experimental MATLAB structure

```
Load file_name

EASNDataProcessing.m

      %signal processor
      partitioning.m

      TimeSP.m
            instantaneouspower.m
            peaks.m
            weigthing.m

      PowerSP.m
            powerspectrum.m

      for all partitions %process result of the signal processor

            %localizer
            TimeIE.m
                  peaks.m
                  weighting.m

            PowerIE.m
                  aggregation.m
                  weighting.m

            LocationEstimator.m
                  weighting.m
                  time_localizer.m
                        lsqnonlin.m
                  power_localizer.m
                        lsqnonlin.m

            %classifier
            VehicleFE.m
                  peaks.m

            HumanFE.m
                  peaks.m

            if targetposition is relevant
                  GunFE.m
            end

            ClassEstimator.m
      end
```