Sound Processing for Autonomous Driving

Vom Fachbereich
Elektrotechnik und Informationstechnik
der Technischen Universität Darmstadt
zur Erlangung des akademischen Grades
eines Doktor-Ingenieurs (Dr.-Ing.)
genehmigte Dissertation

von

M. Sc. Yury Furletov

geboren am 29. Juni 1993 in Moskau, Russische Föderation

Referent:	Prof. Dr.-Ing. Jürgen Adamy
Korreferent 1:	Prof. Dr.-Ing. Volker Willert
Korreferent 2:	Prof. Dr.-Ing. Sören Hohmann
Tag der Einreichung:	08. März 2022
Tag der mündlichen Prüfung:	22. Juni 2022

D17
Darmstadt 2022

Furletov Yury: Sound Processing for Autonomous Driving
Darmstadt, Technische Universität Darmstadt,
Year thesis published in TUprints 2022
URN: urn:nbn:de:tuda-tuprints-220908
Date of the viva voce 22.06.2022

Erklärungen laut Promotionsordnung

§ 8 Abs. 1 lit. c PromO

Ich versichere hiermit, dass die elektronische Version meiner Dissertation mit der schriftlichen Version übereinstimmt.

§ 8 Abs. 1 lit. d PromO

Ich versichere hiermit, dass zu einem vorherigen Zeitpunkt noch keine Promotion versucht wurde. In diesem Fall sind nähere Angaben über Zeitpunkt, Hochschule, Dissertationsthema und Ergebnis dieses Versuchs mitzuteilen.

§ 9 Abs. 1 PromO

Ich versichere hiermit, dass die vorliegende Dissertation selbstständig und nur unter Verwendung der angegebenen Quellen verfasst wurde.

§ 9 Abs. 2 PromO

Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.

Darmstadt, den 08. März 2022 Yury Furletov

Contents

Abbreviations

3D-CVF 3D Cross View Fusion

AD Autonomous Driving

ADAS Advanced Driver Assistance Systems

ANN Artificial Neural Network

ASR Auditory Scene Recognition

ASR Automatic Speech Recognition

BLSTM Biderectional LSTM

CLR Camera-LiDAR-RADAR

CPB Controlled-Pass-By

CR Camera-RADAR

DBF Digital BeamForming

DCNN Deep Convolutional Neural Network

DCT Discrete Cosine Transform

DFT Discrete Fourier Transform

DOA Direction of Arrival

DTW Dynamic Time Warping

EBM Empirical Binary Masks

ECU Electronic Control Unit

EmV Emergency Vehicle

FMCV Frequency-Modulated Continious Wave

FoV Field of View

GCC Generalized Cross Correlation

GMM Gaussian Mixture Models

HLF High-Level Fusion

HMM Hiden Markov Models

ILD Interaural Level Difference

ITD Interaural Time Difference

ITSP Inverse TSP

kNN k-Nearest Neighbor

LDA Linear Discriminant Analysis

LFMCV Linear Frequency-Modulated Continious Wave

LiDAR Light Detection and Ranging

LLF Low-Level Fusion

LMS Least Mean Squared

LSTM Long-Short-Term Memory

MAP Maximum A Posteriori

MDF Module Difference Function

MFCC Mel-Frequency Cepstral Cofficient

ML Maximum Likelihood

MLF Mid-Level Fusion

MLP Multi-Layer Perceptron

NBc Naive Bayesian classifiers

PBM Part-Based Models

PHAT Phase Transformation weighting function

PPCA Probalistic Principal Component Analysis

RADAR Radio Detection And Ranging

RANSAC Random Sampling Consensus

RMS Rated Maximum Sinusoidal

RNN Recurrent Neural Network

ROTH Roth autocorrelation weighting function

SDF Symbolic Dynamic Filtering

SIL Sound Intensity Level

SNIR Sound-to-Noise-plus-Interference Ratio

SNR Sound-to-Noise Ratio

SONAR Sound Navigation And Ranging

SPB Statistical-Pass-By

SPL Sound Pressure Level

SRP-PHAT Steered-Response Power with PHAT

SSL Sound Source Localization

SVM Support Vector Machine

TDOA Time Difference Of Arrival

TFFE Time Frequency Feature Extraction

ToF Time-of-Flight

TSP Time-Stretched Pulse

UBM Universal Background Model

Abstract

Nowadays, a variety of intelligent systems for autonomous driving have been developed, which have already shown a very high level of capability. One of the prerequisites for autonomous driving is an accurate and reliable representation of the environment around the vehicle. Current systems rely on cameras, RADAR, and LiDAR to capture the visual environment and to locate and track other traffic participants. Human drivers, in addition to vision, have hearing and use a lot of auditory information to understand the environment in addition to visual cues. In this thesis, we present the sound signal processing system for auditory based environment representation.

Sound propagation is less dependent on occlusion than all other types of sensors and in some situations is less sensitive to different types of weather conditions such as snow, ice, fog or rain. Various audio processing algorithms provide the detection and classification of different audio signals specific to certain types of vehicles, as well as localization.

First, the ambient sound is classified into fourteen major categories consisting of traffic objects and actions performed. Additionally, the classification of three specific types of emergency vehicles sirens is provided. Secondly, each object is localized using a combined localization algorithm based on time difference of arrival and amplitude. The system is evaluated on real data with a focus on reliable detection and accurate localization of emergency vehicles. On the third stage the possibility of visualizing the sound source on the image from the autonomous vehicle camera system is provided. For this purpose, a method for camera to microphones calibration has been developed.

The presented approaches and methods have great potential to increase the accuracy of environment perception and, consequently, to improve the reliability and safety of autonomous driving systems in general.

Kurzfassung

Heutzutage wurde eine Vielzahl von intelligenten Systemen für das autonome Fahren entwickelt, die bereits ein sehr hohes Maß an Leistungsfähigkeit gezeigt haben. Eine der Voraussetzungen für das autonome Fahren ist eine genaue und zuverlässige Darstellung der Umgebung des Fahrzeugs. Aktuelle Systeme stützen sich auf Kameras, RADAR und LiDAR, um die visuelle Umgebung zu erfassen und andere Verkehrsteilnehmer zu lokalisieren und zu verfolgen. Menschliche Fahrer haben neben dem Sehvermögen auch ein Gehör und nutzen neben den visuellen Hinweisen auch viele auditive Informationen, um die Umgebung zu verstehen. In dieser Arbeit wird ein System zur Verarbeitung von Schallsignalen für die auditiv basierte Umgebungsdarstellung vorgestellt.

Die Schallausbreitung ist weniger abhängig von Verdeckungen als alle anderen Arten von Sensoren und reagiert in manchen Situationen weniger empfindlich auf verschiedene Wetterbedingungen wie Schnee, Eis, Nebel oder Regen. Verschiedene Audioverarbeitungsalgorithmen ermöglichen die Erkennung und Klassifizierung verschiedener Audiosignale, die für bestimmte Fahrzeugtypen spezifisch sind, sowie die Lokalisierung.

Im ersten Schritt wird das Umgebungsgeräusch in vierzehn Hauptkategorien eingeteilt, die sich aus Verkehrsobjekten und durchgeführten Aktionen zusammensetzen. Zusätzlich wird die Klassifizierung von drei spezifischen Typen von Sirenen für Einsatzfahrzeuge vorgenommen. Im zweiten Schritt wird jedes Objekt mit Hilfe eines kombinierten Lokalisierungsalgorithmus lokalisiert, der auf der Amplitude und der Zeitdifferenz der Ankunft basiert. Das System wird anhand realer Daten evaluiert, wobei der Schwerpunkt auf der zuverlässigen Erkennung und genauen Lokalisierung von Einsatzfahrzeugen liegt. Im dritten Schritt wird die Möglichkeit geschaffen, die Schallquelle im Bild des autonomen Fahrzeugkamerasystems zu visualisieren. Zu diesem Zweck wurde eine Methode zur Kalibrierung von Kamera und Mikrofonen entwickelt.

Die vorgestellten Ansätze und Methoden verfügen über ein großes Potenzial, die Genauigkeit der Umgebungswahrnehmung zu erhöhen und damit die Zuverlässigkeit und Sicherheit von autonomen Fahrsystemen im Allgemeinen zu verbessern.

1 Introduction

1.1 Motivation

Autonomous Driving (AD) technology is becoming increasingly popular every year. In modern transport science, the introduction of such technologies is believed to have enormous potential in increasing efficiency, safety, and accessibility. The beginning of this industry is considered to be the DARPA Grand Challenge 2004, which was organized by the US Department of Defense. The aim was to motivate development and research in the field of fully autonomous vehicles. In 2004, the rules of the competition were quite simple:

- The car must not damage other vehicles, road surfaces or the environment;
- Any human-control is prohibited;
- There are no restrictions on the size of the car;
- The specific route is announced 2 hours before the start;
- The track may include asphalt sections, country roads, desert off-road;
- Obstacles can be ditches, berms, ruts, sand, standing water, rocks, and narrow tunnels among others.

The prize pool was 1 million US dollars and 15 teams participated. The teams were challenged with traversing a 230-km section between the cities of Barstow (California) and Primme (Nevada). That year, only eight cars out of 15 managed to leave the starting line, two cars managed to cover the distance of 11 km, and no one completed the entire route. The best result was 11.8 km, which was achieved by the Carnegie Melon University team with a Hummer H-1 car. At first glance, the results of this competition seem completely unacceptable, but after 17 years it is clear that this event generated a huge interest in AD and served as the beginning of a whole new

industry. For comparison, a year later in 2005, 195 projects participated in the qualifying competitions, of which 23 reached the final.

Currently, in several countries, AD technologies are being tested on public roads. Cars are allowed to drive autonomously, but with the obligatory presence of a human in the driver's seat. Furthermore, the use of various Advanced Driver Assistance Systems (ADAS) is allowed, such as adaptive cruise control, a lane-keeping system, or a collision-avoidance system. Thus, the automotive industry is moving toward automation at a fairly rapid pace. However, the widespread adoption of AD technology faces certain social challenges. The full implementation of AD technologies is impossible without public approval. In several countries that are pioneering AD, public debate about the technology has reached a very high level, which indicates the presence of several social problems.

The first and most pressing problem is responsibility. Public activists often pose the question of responsibility for the actions of autonomous vehicles. According to current legislation, all actions associated with a risk to life are somehow committed under personal responsibility. In the field of transport, full responsibility lies with the driver who drives the vehicle, who is criminally and financially responsible for the technical condition of the vehicle as well as for all damage caused by the vehicle while driving. Legally, the problem of responsibility has not been resolved and is not enshrined in any legislation. At the moment, all responsibility for the actions of an autonomous vehicle rests with the person in the driver's seat. Such a concept cannot be called fully autonomous because it implies the presence of an operator.

The second problem stems from the first, namely the proof of reliability. For the widespread introduction of technologies, it is necessary to prove their safety to society. There should be methodologies for validating and certifying autonomous vehicles, which are currently a broad and growing field of science.

The third problem is the lack of consensus in decision-making related to life-threatening decisions. The MIT Laboratory launched an online game that invites people to put themselves in the position of an autonomous vehicle and to make a decision in a desperate situation that in any case will entail the death of a person. For example, the user must to choose who should be run over by an autonomous vehicle, a man or a woman, a grandmother or a girl, and a pregnant woman or a woman with a baby. The game demonstrates that in a desperate situation, such a choice can be very difficult. Thus, situations of such a choice must be avoided.

After reviewing of all the aforementioned social problems related to AD

implementation, it became clear that they are all based on two technical problems with modern technologies – namely safety and reliability. This is why a huge amount of AD research is related to these two areas. This means that the introduction of AD technologies is entirely dependent on solutions to these technical problems in the near future. Reliability and safety must be increased significantly and proved in front of society.

Safe driving requires an autonomous vehicle with capabilities that are at least similar to those of a human driver, including visual, auditory, and haptic skills for robustly perceiving the environment and safely navigating the car. One way to improve robustness and reliability is to add more sensors that are based on complementary measurement principles to the system, which can fail under various environmental conditions but hold the same environmental information. Typically, these sensors vary in their sensory Field of View (FoV), range, accuracy, and spatial and temporal resolution. The underlying algorithms that extract the same information based on different types of sensors, such as stereo vision and LiDAR for extracting a three-dimensional (3D) point cloud of the environment, must consider all of these variations for proper sensor fusion [10, 68].

In this thesis, a sound processing system for AD is introduced. As far as it is agreed that AD systems must have at least the same performance and capabilities as human drivers, sound is one of the "must-have" types of data for environment representation. Due to air resistance and frictional forces, absolutely any action or movement of a person or mechanism is accompanied by sound, and the traffic environment is not an exception. Sound generated from other traffic participants or by the car itself contains much additional information about the traffic scene compared with visual cues. For example, combustion engines, bicycle or tram bells, Emergency Vehicle (EmV) sirens, or the surface that the car is driving on may produce specific sound patterns. Such sounds provide information that can be used by an autonomous vehicle to determine the correct driving strategy or plan a more optimal route. For example, the sound from the tires holds information for traction control [75]. Different sound processing algorithms are able to detect and classify various objects as well as to extract auditory cues that are specific to certain types of objects. Furthermore, such algorithms can localize the position of objects by extracting absolute loudness levels from a microphone or the Interaural Time Difference (ITD) and Interaural Level Difference (ILD) between several microphones [82]. Hence, sound localization and classification involve information that is not available to other sensors of an autonomous car. Thus, the accuracy and richness of the environmental representation can be improved when other

sensors are employed, while some perceptual capabilities are maintained for interacting with the environment if the other sensors are unable to work.

Sound propagation is less dependent on occlusion compared with all other types of sensory signals. Furthermore, in some situations, it is less sensitive to different types of weather conditions, such as snow, ice, fog, and rain, which one study mentioned as a call for AD [71]. Another advantage of sound is the large range of distances over which objects can be recognized, as long as the sound-to-noise ratio is high enough to detect and analyze specific sound sources. In accordance with road traffic regulations, several vehicles have priority on the road. To indicate exclusivity, sound and visual signals are used in the form of a siren and a flashing light. Consequently, current legislation implies that sound is a source of information on the road. Thus, sound processing is a necessary system for the full functioning of an autonomous vehicle.

For example, EmVs must be considered when driving but only rarely appear. According to Arnold Itkin's report titled "Statistics on Emergency Vehicle Accidents in the United States" (2018), an estimated 6500 accidents involving ambulances occur every year. Moreover, 35% of these cases result in the injury or fatality of at least one occupant of the vehicle. If an injury occurs, there are on average three unique injuries for every one accident. Approximately 60% of accidents with ambulances occur during emergency use, and those involved in medical emergencies have the highest risk of crashing. Fire truck accidents occur during emergency use in 70% of cases, and a rollover occurs in 66% of fatal fire truck accidents. The same dangerous conditions are applicable to police cars. Police officers have double the rate of motor vehicle crashes per million compared with the general public. Based on these statistics, it is obvious that EmVs quite often have accidents, which are associated with the highly risky driving style required in emergency situations.

In most cases, EmVs are driving outside of comprehensive traffic rules. Hence, it is necessary for autonomous vehicles to implement special algorithms to conveniently react to and interact with EmVs. Certain case-specific traffic rules exist that also vary across different countries. For example, an ambulance can drive in the opposite lane, cross the road at a red light, or overtake other cars where it is forbidden to do so. Furthermore, in a traffic jam, if drivers see an EmV behind them, they must let it pass or make a special lane in the middle of the road. All of these situations seldomly occur, but when they do, autonomous vehicles must react carefully and appropriately. When an EmV is detected, a sound-based approach has some advantages. All EmVs are equipped with a siren, which is audible

over large distances and even inside vehicles. Therefore, the siren of an oncoming EmV can be recognized and localized by sound sensors much earlier than by other sensors. Moreover, several types of siren sounds exist depending on the country to enable people to clearly differentiate between an ambulance, police car, and fire truck. According to the type of EmV, different interaction scenarios must be implemented. For example, a fire truck requires more space on the road, whereas a police car may order a vehicle to stop through an auditory or visual sign.

Currently, sound processing is a popular topic in robotics. This topic is especially relevant in the field of mobile robots. Based on acoustic data, it is possible to determine the type of surface, which is a crucial aspect for trajectory planning. Furthermore, some security robots use sound information in their work; for example, the breaking of silence can indicate the presence of strangers in a protected facility. Several studies in the field of robotics have demonstrated high efficiency in the use of sound for the indoor navigation of mobile robots. Good results have been reported for this method in small confined spaces where sound propagation is limited and there is little external noise. Due to the peculiarities of propagation, this method is applicable for cooperative robot communication as well as localization, especially in the case of indoor usage.

All of the aforementioned facts indicate the excellent potential and prospects of sound processing systems for AD and robotics in general.

1.2 Contributions

This thesis makes the following contributions:

- A comprehensive comparative study is conducted on the state of the art of current environment representation facilities of AD as well as current sound processing systems and applications for robotics.
- Sound localization techniques and methods are provided for AD applications.
- Extended sound classification capabilities are enabled for traffic environments, including object and motion classification as well as data sets for training.
- A camera-to-microphone calibration technique is provided for sound source visualization.
- A general sound processing system for AD is presented.

1.3 Dissertation Outline

The remainder of this dissertation is organized as follows.

Chapter 2 presents the state of the art of autonomous vehicle environment representation sensors and setups such as LiDAR, radar, camera-based systems, and ultrasonic sensors, as well as approaches for sensor data fusion. The chapter also provides an analysis of the weak points of current environment representation systems as well as an overview of current sound processing systems in automotive applications.

Chapter 3 presents sound processing algorithms for sound source localization and classification, including an overview of the direction of arrival and distance estimation as well as approaches for sound source classification.

Chapter 4 introduces sound processing system for AD which consists of sound based classification system with object and motion classification capabilities as well as dataset. Moreover, sound source localization system with angle and distance estimation functions is introduced.

Chapter 5 presents camera to microphone calibration method for audio-video data fusion.

2 State of the Art

In this chapter, an overview is provided of current environment representation systems for AD as well as existing sound processing algorithms. First, the sensors and setups currently used for environment representation in autonomous vehicles are presented in Section 2.1. Next, we outline actual sound processing techniques and their implementation methods in Section 2.2. Finally, the weak points of modern sensor setups are highlighted in Section 2.3 to explain the limitations, accuracy, and problems of existing solutions.

2.1 Sensors for Autonomous Driving

Modern autonomous vehicles have a very complex sensor setup. These sensors are combined to provide as high accuracy as possible in different road conditions. The most widely used sensor set is LiDAR, short range radar, long range radar, cameras and ultrasonic sensors (Figure 2.1). Such a combination provides adequate environment representation which could be compared with human drivers, or even better, but all of these sensors have limitations which can be caused by weather conditions, lightness or traffic.

2.1.1 Lidars

The first light beam-based range measurement approaches were developed in the 1930s to investigate the atmosphere's structure using searchlights. In 1938, light pulses were used to determine the height of clouds. In 1960, the laser was invented, and just a few years later, laser rangefinders began to be used in American tanks. In 1969, a laser rangefinder was used to measure the distance from the Earth to the Moon using a special target on Apollo 11. Today, Light Detection and Ranging (LiDAR) is used in various fields from archaeology to biology, but most actively in cartography, meteorology, aviation, and robotics, including in AD.

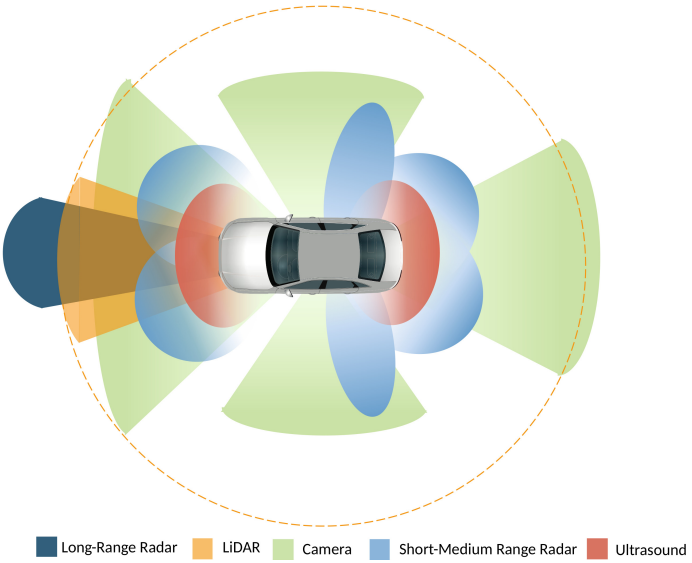

Figure 2.1: Environment representation sensor setup.

LiDAR is a laser sensor that can scan a 360° field around a car and collect high-definition data from the environment in the form of 3D point clouds. LiDAR demonstrated a very high potential for AD for the first time at the DARPA Grand Challenge in 2007. The top three projects were equipped with several LiDAR sensors (hereinafter "lidars"). Lidar perception and localization tools are used to detect, track, classify objects, evaluate posture, and predict intentions to represent the environment around an autonomous vehicle. This type of sensor has a large range of operation as well as high data quality. In addition to orientation in space, lidars are used to compile 3D maps of terrain that are used in some projects with autonomous vehicles.

Currently, most highly automated vehicles have one or more lidars despite their high cost and design disadvantages, such as moving parts and special placement schemes on the roof. A review of modern projects indicated that a trend exists in the industry of increasing the number of lidars in systems. Whereas autonomous cars previously had only one lidar located on the roof, there are now installations that include six to eight lidars located on the roof, bumpers, and external rear-view mirrors, as well as lidars located at an angle relative to the road plane, which reduce blind

spots on the sides of autonomous vehicles.

Principle of operation

The principle of operation of LiDAR is to emit a laser beam and capture its reflection. With data on the speed of propagation of the beam in space and the time spent on overcoming the distance, it is possible to calculate the distance to the object of interest. A laser diode generates a beam at a near-infrared wavelength, which propagates in the environment. After being reflected by an object or the environment, the signal is picked up by the receiver. The distance to the measured object is proportional to the difference between the transmitted and received signals and can be estimated. The energy difference between the emitted and reflected signals depends on the material and surface of the target object as well as on the state of the medium in between. LiDAR output data consists of 3D point clouds that are associated with the scanned medium and the intensity of the reflected signals.

Figure 2.2 presents the basic concept of the LiDAR scanning process.

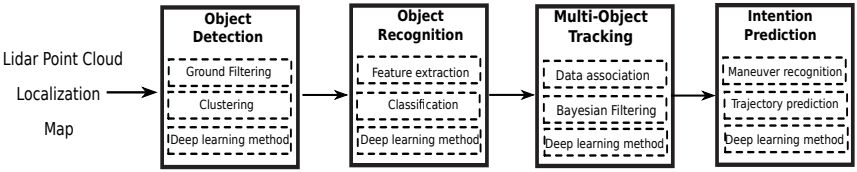

Figure 2.2: LiDAR scanning process.

The transmitter emits the wave and transmits it to the medium through the optics. After radiating and overcoming a certain distance in space, the laser's power weakens depending on the density of the medium; then, after being reflected by the surface of the target object, the laser is scattered. The receiving optics capture a part of a signal, which is converted into an electrical signal by a photodetector. After the emitted wave is reflected from the target surface at distance R, the photodetector receives the amount of power PR, which can be described by the LiDAR power equation as follows [79]:

$$P_R = E_p \frac{c\eta A_r}{2r^2} \beta T_r, \tag{2.1}$$

where E_p is the total energy of a transmitted pulse laser, c is the speed of

light, A_r is the area of receive aperture at range r, η is the overall system efficiency, β is the reflectance of the target's surface, which is decided by both surface properties and incident angle, and T_r is the transmission loss through the transmission medium.

From Equation (2.1), it follows that the power of the signal decreases following a pattern of inverse square proportionality relative to the distance; therefore, the magnitude becomes lower with increasing distance, which makes objects darker with increasing distance. Simultaneously, the power of the transmitted signal is limited by eye-safety regulations, so a simple increase in power is not the solution for operation range extension. Current research and development are working on improvements to optics, photodetectors, and signal processing algorithms.

The principle of the laser rangefinder distance measurement depends on the laser beam signal modulation type:

- Direct detection laser rangefinder,
- Coherent detection laser rangefinder.

A direct detection laser rangefinder uses a pulsed laser signal, allowing distance to be measured by Time-of-Flight (ToF). The time-of-flight method is based on the time delay between signal transmission and reception.

$$R = \frac{ct_d}{2n}, \tag{2.2}$$

where c is the speed of light, n is the index of refraction of medium, Δt the time difference between transmitted and received signals.

The main disadvantages of ToF lidars are limited potential in terms of increasing the operation range due to regulations on maximal laser power and the influence of weather conditions, such as strong sunlight, which interfere with the reflected signal. Moreover, ToF lidars are the most common type of LiDAR on the automotive market due to the simplicity of the signal processing algorithms and design [45].

A coherent detection laser rangefinder determines speed and distance by performing indirect Doppler effect-based measurements using a Frequency-Modulated Continious Wave (FMCV). The FMCV of coherent LiDAR signals consists of chirps stretched up and down in the frequency domain. The principle of operation is basically the same as FMCV radar but with the use of optical instead of electromagnetic signals. The wave is transmitted to the medium, reflected from the obstacle, and caught by the receiver. The received signal is the same signal that was transmitted but with a certain

delay in time if the object is not moving relative to the transmitter. In case the object is in motion relative to the transmitter, the received signal will (in addition to time delay) exhibit a shift in the frequency domain by a value of the Doppler frequency f_d:

$$f_d = \frac{2v_r}{\lambda}, \tag{2.3}$$

where v_r is the radial velocity of the target relative to the radar, λ is the wavelength of the laser

The relative velocity of the object in this case is

$$v_r = \frac{f_d \lambda}{2}. \tag{2.4}$$

Compared with ToF LiDAR, coherent or FMCV LiDAR is able to measure distance and velocity simultaneously. Furthermore, FMCV signal modulation is less dependent of interference. On the other hand, FMCV lidars have higher requirements for the quality of laser generators with longer coherent distances.

Construction

LiDAR systems can in general be divided into two subsystems: the rangefinder and the scanning system.

The laser rangefinder contains the following:

- a laser transmitter for directing the modulated wave to the target,
- a receiver or photodetector for processing and conversions of the reflected photons into an electronic signal,
- a photoelectric converter of laser radiation, which includes an optical element for concentrating the transmitted laser and focusing the captured signal onto the receiver,
- signal processing module for final distance estimation.

The laser transmitter is essentially the laser generator. ToF lidars use a pulsed laser signal generated by a fiber laser or a pulsed laser diode. Wavelength modulation of the laser is performed in accordance with eye-safety requirements and cost. The most common lasers in automotive applications are near infrared (850–950 nm) and short-wave infrared (1550 nm) lasers.

The photodetector converts the reflected power into an electronic signal. The choice of photodetector type is related to the wavelength of the emitted laser. The most widely used photodetector types are described as follows [45]:

- *"PIN photodiode is formed by a p-i-n junction that creates a depletion region that is free of mobile charge carriers. By applying a reverse bias to a photodiode, absorbing a photon will generate a current flow in the reverse-biased photodiode.*
- *Avalanche photodiode (APD) use reverse voltage to multiply photocurrent through avalanche effect. The APD's ability to multiply signals reduces the effect of noise and achieves higher internal current gain (around 100) and SNR than the PIN photodiode.*
- *Single-photon avalanche diode (SPAD) is an APD designed to operate with a reverse-bias voltage above the breakdown voltage (Geiger-mode), which allows a detection of very few photons in very short time. SPAD can achieve a gain of* 10^6 *that is significantly higher than APD.;*
- *Silicon Photomultiplier (SiPM) is based on SPAD, while enable photon counting. The Geiger-mode in which a SPAD operates is a photon-trigger mode that a SPAD cannot distinguish the magnitude of received photo flux. To overcome this issue, SiPM integrates a dense array of 'microcells' (a pair of a SPAD and a quench resistor) working identically and independently."*

The scanning system scans laser beams in various directions along the vertical and horizontal planes, and is also known as a beam steering system. In current applications, beam steering techniques are grouped into mechanical spinning systems, which use rotating parts (mirrors and prisms) controlled by an electric drive, and solid state systems, which do not use moving parts to redirect the beam, as depicted in Figure 2.3.

- *"Mechanical Spinning is the most popular scanning solution for automotive LiDAR is the mechanical spinning system [33], which steers the laser beams through rotating assembly (e.g. mirror, prism, etc) controlled by a motor to create a large field of view (FoV). Conventionally, nodding-mirror system and polygonal-mirror system [57] are the main types applied (Figure 2.3 a).*

- *MEMS Micro-Scanning: MEMS (Micro-Electro-Mechanical Systems) technology allows the fabrication of miniature mechanical and electro-mechanical devices using silicon fabrication techniques. In essence, a MEMS mirror is a mirror embedded on a chip [85]. The MEMS mirror is rotated by balancing between two opposite forces: an electromagnetic force (Lorentz force) produced by the conductive coil around the mirror, and an elastic force from torsion bar, which serves as the axis of rotation.(Figure 2.3 b)*

- *Flash: Originally applied for spacecraft in autonomous landing and docking with satellites, 3D flash LiDARs [4] totally remove the rotating parts within scanning systems. Hence, they are truly solid-state. A flash LiDAR behaves as a camera. A single laser that is spread by an optical diffuser to illuminate the whole scene at once. Then, it uses a 2D array of photoiodes (similar to the CMOS/CCD for camera) to capture the laser returns, which are finally processed to form a 3D point clouds, as shown in Figure 2.3 c.*

- *OPA (Optical Phased Array): As a type of true solid-state LiDAR, optical phased array (OPA) LiDARs [53, 59] don't comprise moving components. Similar to the phased array Radar, an OPA is able to steer the laser beams through various types of phase modulators. (Figure 2.3 d)"*[45]

2.1.2 Radars

Radio Detection And Ranging (RADAR) or "radar" hereinafter refers to a distance measurement sensor based on the principle of electromagnetic radiation. This method of detection and ranging was developed over 100 years ago, with the first experiments on capturing reflected radio waves starting at the end of the 19th century. In the 1930s, the first radars were used to detect aircraft and ships for military defense purposes. Early radars could be installed only on land or on large ships due to the size of the equipment, but already at the beginning of the 1940s centimeter-range radars were installed on military airplanes. Since then, radar has been a widely used technology for detection and ranging in aviation, maritime, spacecraft, motor transport, and meteorological observation.

At present in the automotive industry, radar technologies are used not only for research in the field of autonomous driving but also as original

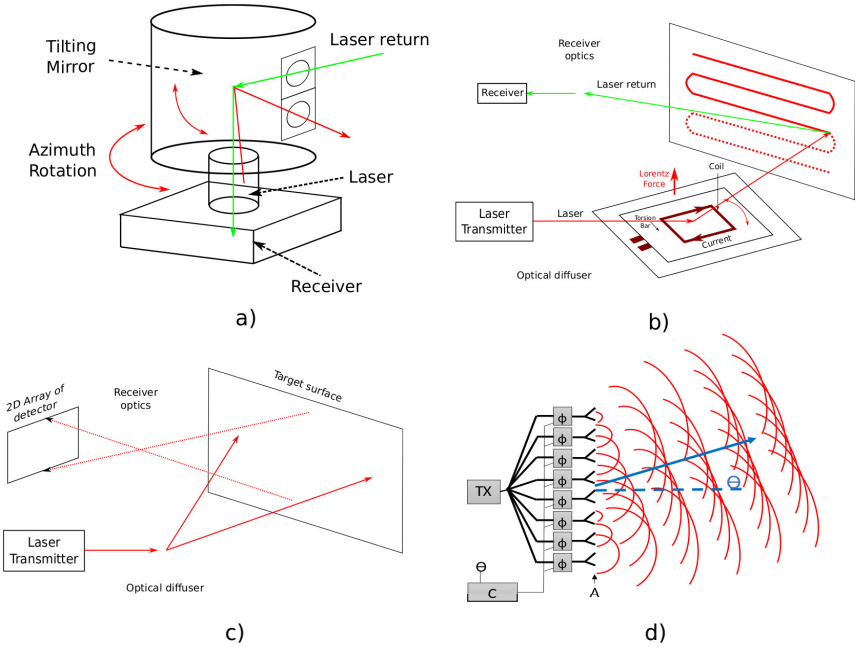

Figure 2.3: LiDAR systems categorized by scanning techniques [45].

equipment manufacturer (OEM) sensors for in-built driver assistance systems. Radars are used in blind spot monitoring as well as in collision avoidance, pedestrian avoidance, and other advanced driver assistance systems. The main function of these systems is to predict and warn the driver about a dangerous situation. Radars have a wide range of operating frequencies depending on the application, from high frequencies (3–30 MHz) for ultra-long-range radars to millimeter frequencies (40–300 GHz). Currently available commercial automotive radars operate at 24 GHz for short-range radars or at 77–81 GHz for long-range radars; however, there is a trend toward increasing this range in current research and development.

Short-range radars have a range of up to 80 m. Radars of this type, if installed on the sides of a car and directed in the opposite direction to that of driving, are used as part of blind spot monitoring systems as well as lane change assistance systems. When installed at the front of the vehicle, these radars are used in front cross-traffic alerts, such as when leaving a parking lot, as well as in side impact and cyclist warning systems.

Long-range radars provide a long range of operation up to 250 m, but they also have a narrow field in short ranges. Long-range radars are used in automatic cruise control systems as well as pedestrian collision warning and forward collision avoidance systems. As illustrated in Figure 2.4, long-range radar has a small width of operation in the near-field, whereas short-range radar has not been adapted for far-field applications; thus, these two types of radar are not interchangeable, which necessitates using both of them simultaneously for adequate environment representation.

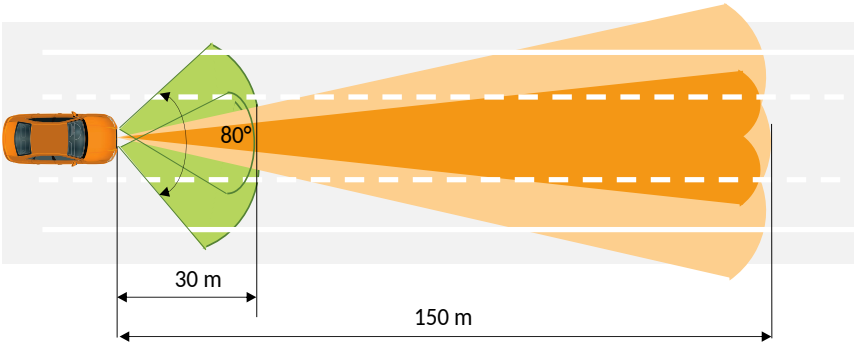

Figure 2.4: Short- and long-range RADARs FoV illustration.

Construction

Similar to LiDAR, the key functions of radar are the detection, localization, and tracking of objects of interest in the FoV around the vehicle. In the case of automotive transport, objects of interest are other road users, such as cars, bicycles, and pedestrians. In ongoing research, radars also have the ability to recognize and classify objects. A radar system consists of three parts: the transmitter, the receiver, and the signal processing system, as well as an antenna as an external interface.

The transmitter consists of a signal generator, a converter, and a power amplifier. The transmitter is essentially a signal emitter, the design and power of which mainly determine the detection range of the radar; thus, the higher the power of the emitted signal, the greater the range of the radar. Most radars have the ability to adjust their power, which makes their design more complex. The maximum range of a radar is proportional to the fourth root of the transmitted power; therefore, to increase the

detection range, the power of the emitted signal must be increased 16 times.

The receiver receives a signal reflected from an object in the FoV. Due to noise and interference, the receiver must maximize the signal-to-noise ratio by suppressing and rejecting interference. The main components of the receiver are a low noise amplifier and a buck converter.

The signal processing system applies various algorithms to process the received signals in order to extract data for the determination of the location, as well as for the identification, classification, and tracking of objects. Noise and interference necessitate advanced signal processing algorithms. Furthermore, the capabilities of complex algorithms suffer from hardware limitations as automotive applications have strict requirements for the size and shape of the installation [22].

An antenna is provided for receiving and transmitting radar signals. Depending on the number of antennas, radars are divided into monostatic, bistatic, and multistatic varieties. Monostatic radar uses a single antenna for transmitting as well as receiving signals. In bistatic radar, the transmitter and receiver are separated; that is, two antennas are used, one of which transmits a signal and one receives. A multistatic radar is essentially a combined system consisting of several monostatic or bistatic radars with a common coverage area.

Several types of antennas are used in automotive radars: planar, waveguide, reflector, and lens. Planar antennas are in high demand due to their low cost and simple design. Modern radars use multiple antennas combined into antenna arrays. An antenna array is a set of several antennas, the data from which are processed together in some combination, which improves the accuracy compared with a separate antenna [43, 77]. The main advantages of using an antenna array are increased signal power, increased directivity, decreased side-lobe power, increased Sound-to-Noise Ratio (SNR), maximized Sound-to-Noise-plus-Interference Ratio (SNIR), and increased antenna gain [60]. When designing antennas and antenna arrays, it is necessary to reduce losses, size, and cost as much as possible. Several configurations of arrays exist, the most common of which are linear, flat, rectangular, and circular.

Principle of operation

This section explains the principle of operation of radar using the example of monostatic radar. Monostatic radar consists of a transmitter, a receiver, and one antenna that performs both functions at once. The transmitter

emits a signal with power P_t and wavelength λ that is transmitted by the antenna in the form of an electromagnetic wave. Antenna gain G is the maximum radiation intensity of an antenna compared with an isotope radiator. Furthermore, the waves propagate in space until they reach an obstacle at a distance R from the emitter. After that, parts of the radiated waves are reflected from the obstacle and return, where the antenna converts part of them into a received signal with power P_r. The fraction of energy reflected toward the receiver depends on the target's radar cross-section σ. The radar range equation is used to calculate the effective range of the radar [63]:

$$P_R = \frac{P_t G^2 \sigma \lambda_0^2}{(4\pi)^3 R^4 L_G}, \tag{2.5}$$

where L_G is the total loss, which consist of system loss, non uniform radar cross-section loss, and atmospheric propagation loss.

The design of the radar depends on the choice of the transmitted signal and its modulation; thus, this parameter must be selected and considered at the design stage of the radar. Modern radars used in automotive applications use Linear Frequency-Modulated Continious Wave (LFMCV) with slow chirp or fast chirp. The difference between them lies in the modulation time. The slow chirp modulation time is usually in the range of a few milliseconds to 100 milliseconds, and it is currently widely used in automotive radars. Fast chirp FMCW modulation was first applied in automotive radars in 2007 [83]. The main idea is to reduce the chirp modulation time to the microsecond range, thus enabling the usual chirp to be replaced by a sequence of shorter chirps.

The radar transmits a linear signal stretched in the frequency range B and in time T_s. Part of the transmitted energy is reflected from the obstacle at a distance R from the transmitter. The radar receiver receives the reflected wave of the same signal but with some delay depending on the distance to the obstacle. If the object does not move relative to the radar, then the signal is the same, but with a delay in time domain t_d. The time delay between the transmitted and received chirp is essentially the time of travel to the obstacle and back as follows [39]:

$$t_d = \frac{2R}{c}. \tag{2.6}$$

If the object is in motion relative to the radar, then the received signal, in addition to the time delay, has a shift in the frequency field by the value of the Doppler frequency f_d which is equal to

$$f_d = \frac{2v_r}{\lambda}, \tag{2.7}$$

where v_r is the radial velocity of the target relative to the radar, λ is the wavelength of the transmitted signal.

The transmitted and received signals are down-converted to provide the beat frequency f_b. In the case of zero relative speed, the frequency is

$$f_{b_0} = \frac{B}{T_S} * \frac{2R}{c}. \tag{2.8}$$

Furthermore, if the radar and the object are in motion relative to each other beat frequency is a difference between up $f_{b_{up}}$ and down fb_{down} chirps

$$f_{b_{up}} = f_{b_0} - f_d, \tag{2.9}$$

$$f_{b_{down}} = f_{b_0} + f_d. \tag{2.10}$$

The distance to the object of interest R can be estimated as:

$$R = \frac{cT_S}{4B} * (f_{b_{down}} + f_{b_{up}}). \tag{2.11}$$

The relative radial velocity v_r of the object can be estimated as:

$$v_r = \frac{\lambda}{4} * (f_{b_{down}} - f_{b_{up}}). \tag{2.12}$$

In addition to information about speed and distance, radar estimates the direction to the object of interest. The quality of angular information depends on the type of antenna installed and the design of the radar sensor; thus, the concept of the antenna system is one of the main distinguishing factors of different types of radars. In [30], the authors considered the most widely used antenna systems of modern 77 Hz automobile radars. Radar characteristics such as FoV and achievable angular separation are determined by the performance characteristics of the antenna installation, such as gain, input matching, bandwidth, and the number of channels. Therefore, the antenna system determines the angular properties of the radar.

Quasi-optical beamforming is a concept that uses quasi-optical elements, such as a dielectric lens, to shape the antenna pattern and FoV. Typically, this concept is used in long-range radars where a narrow FoV but a long

operating distance are required. This concept allows a large antenna aperture to be obtained at a low manufacturing cost since the quasi-optical elements are easy to manufacture. When using several antenna feeders, each of which is connected to a separate receiver channel, it is possible to receive several multidirectional beams that are processed simultaneously. This concept is implemented in fixed-beam long-range radars, such as the Bosch LRR3 with an operation range up to 250 m.

Mechanical scanning is a concept where directionality is controlled mechanically using various mechanisms to influence the antenna. This method requires the implantation of additional elements into the design, which increases design complexity and cost as well as reduces reliability. Furthermore, mechanical control is slower due to the actuation time. However, mechanical scanning provides flexibility in the range and width of operation. This concept was realized in the Conti ARS300 radar sensor, which provides a width of 56° for ranges under 60 m and 17° for ranges up to 200 m.

Analog scanning is a concept that uses active elements, such as phase shifters and variable gain amplifiers, to electronically control the radiation pattern directly at the external millimeter wave interface. This concept allows one to quickly change the width and direction of the beam, which enables working in several modes. The Spektra radar from Metawave was built with this concept and, together with advanced digital signal processing techniques, provides 5D imaging capabilities.

Digital BeamForming (DBF) is a concept where the beam is steered in a digital frequency band. The whole structure remains fixed and does not require additional elements. The flexibility of the radiation pattern is achieved using digital signal processing algorithms, which require multiple antenna channels. DBF is widely used in short-range radars where a wide field is required as well as in long rage radars, such as the Denso DNMWR004.

Radars have already gained much popularity, and their use in cars will grow at a rapid pace in the near future. At present, countries such as the United States, Japan, and EU Member States have already developed and implemented roadmaps to improve road safety through automation and driver assistance systems, which make the presence of radar-based systems necessary.

2.1.3 Image Processing

Computer vision systems are now highly common in the field of AD systems, driver assistance, and robotics. Development in this area began in the 1970s, when computers reached a certain level of power that allowed them to process large amounts of data, such as images. The main idea behind computer vision is basically the same as human vision. This is every image is analyzed in order to detect, localize and recognize various objects in the environment.

The first implementations of computer vision were in the fields of military equipment and heavy industry. The use of computer vision enables visual quality control of manufactured products on conveyor belts, such as defect checks. Furthermore, cameras are used to measure an object's position and orientation during automated production processes, such as robot arm operations.

The military field of application has long been the largest. Computer vision is installed in various aircraft in the form of soldier and enemy equipment detection systems for missile guidance.

Relatively new areas of computer vision application are various autonomous underwater, ground (robots and cars), and air vehicles with a wide range of autonomy, ranging from support for the driver or pilot, up to fully autonomous systems. In AD applications, computer vision systems are used for navigation and environment representation to obtain information about the cars' location and to create a map of the environment. Several car manufacturers have demonstrated autonomous vehicle control systems based on computer vision as well as in combination with other sensors and systems.

Today, computer vision systems used in the automotive industry perform the tasks of detecting, recognizing, and classifying objects around the car as well as determining the location of objects of interest relative to the camera. Cameras are integrated into standard driver assistance systems, and they are used for recognizing road signs and markings as part of various ADAS.

Principle of operation

The vision system of an autonomous vehicle can be divided in two main functions: localization and classification.

In the context of computer vision, localization refers to the estimation of the moving camera's pose change between consecutive frames. This process is also known as visual odometry. This motion consists of the rotation

matrix R_V and subsequent translation T_V. The trajectory of the moving camera can be reconstructed by concatenating the pose changes (R_V, T_V) over time.

The starting point of a vision system is image formation. This process can be explained with an ideal pinhole camera, which is a simple camera without a lens; thus, it has no optical blur, distortion, or defocus. The camera has a 3D coordinate system, where the X-axis faces to the right, the Y -axis faces downwards, and the Z-axis points to the front. An object with 3D coordinates is relative to the camera's coordinate system and given by the projective transformation:

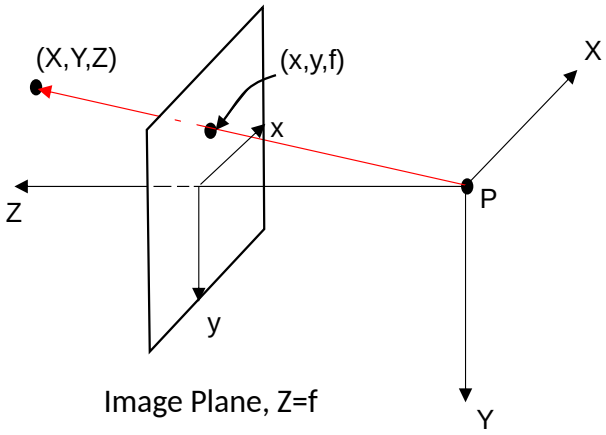

Figure 2.5: Image formation process.

$$\begin{pmatrix} x \\ y \\ f \end{pmatrix} = \frac{f}{Z} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}. \tag{2.13}$$

The distance between the image plane and the projective point P is called the focal length f. The result of the image formation process is a digital image $I(t)$ at time t.

The second step is feature extraction. Feature extraction is a dimensionality reduction process that decreases the number of variables to specific groups, after which further image processing occurs. Feature extraction highlights certain characteristics of an object in an image from the original

raw frame. It is assumed that the features must be informative and characteristic only for certain type of object, which allows further classification of the image or part of the image based on the received attributes. Thus, a large amount of raw data is reduced to a small set, but one that contains only crucial data.

The third step is classification, which refers to the recognition of patterns in the image. For classification tasks, convolutional neural networks are used in most cases. Convolutional neural networks can classify an entire image, but it is necessary to distinguish several features in one image for computer vision tasks. To solve this problem, the sliding window method was proposed. A sliding window smaller than the picture scans it, thereby dividing the picture into many pictures and classifying each one separately, forming a map of class probabilities. Thus, it is possible to recognize several images in one image, as depicted in Figure 2.6. In AD, image classification is performed to recognize road markings and signs, traffic lights, pedestrians, and other traffic participants.

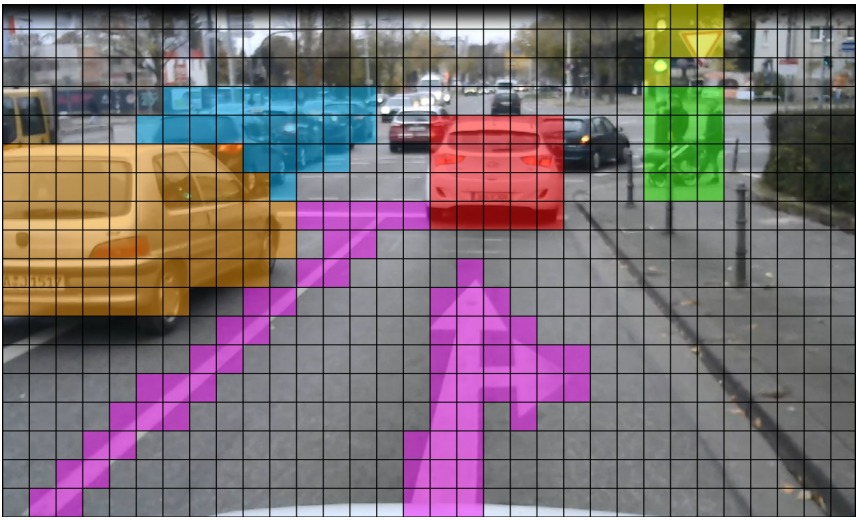

Figure 2.6: Example of image classification.

Construction

Computer vision systems in automotive applications are mainly based on a two-camera system. However, in the field of AD, solutions already exist that offer a 360° view around the vehicle.

In terms of the integrated computer vision systems used in driver assistance and collision avoidance systems, the most widely used solutions are from Bosch and Continental.

The construction of a camera system is illustrated in Figure 2.7.The dual camera system is combined in a common housing and located inside the car. The housing is rigidly attached to the car's windscreen near the rear-view mirror and within the windscreen wiper operating zone, thus avoiding failure due to dirt, snow, and rain.

Figure 2.7: Computer vision sensor of Tesla Model 3.

In a range of autonomous driving projects, cameras are placed on the roof of the vehicle, together with other sensors traditionally placed on the roof, such as LiDAR and GPS antennas. This design certainly has several advantages, such as a larger FoV and the ability to distinguish objects at a greater distance due to the higher position of the cameras. However, this approach suffers in harsh weather and road conditions as the mounted external camera are exposed to frost, moisture, dust, and dirt.

2.1.4 Ultrasonic Sensors

Ultrasound sensors, also known as Sound Navigation And Ranging (SONAR) is acoustic-based rangefinder sensor. They were developed at the end of the 1910s for the detection and ranging of underwater objects. The basic principle of operation is the same as radar, but sound is used as the signal. These sensors are widely used in marine transport for measuring the distance to the bottom or other underwater obstacles. On submarines, sonar is used as the main sensor for navigation underwater.

In automotive applications, ultrasonic sensors are very widely used for parking assistance. The "ultrasonic Back Sonar" was introduced by Toyota in 1982 and patented in 1988. Today, most vehicles are equipped with parking sensors based on sonar technology. In most cases, from two to four sensors are installed in the back bumper of the vehicle and provide information to the driver about the distance to the obstacle. Simultaneously, a vehicle could be equipped with both rear and front parking sensors, which make parking in narrow spaces safer and more convenient.

Principle of operation

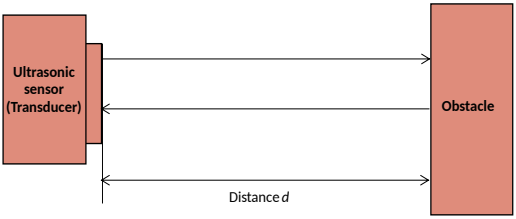

Figure 2.8: Principle of ultrasonic sensor operation.

Ultrasonic ranging is based on the ToF method, which is applied to radar and LiDAR as well. Sensors transmit the ultrasonic pulse to the medium, usually at a frequency of 40 kHz. After the signal reaches the obstacle, a reflected echo signal propagates back to the sensor and is caught by a receiver. For one sensor, the distance to the obstacle R is measured by the time delay t_d between the signal being transmitted and received as follows [70]:

$$R = \frac{t_d c_s}{2}, \tag{2.14}$$

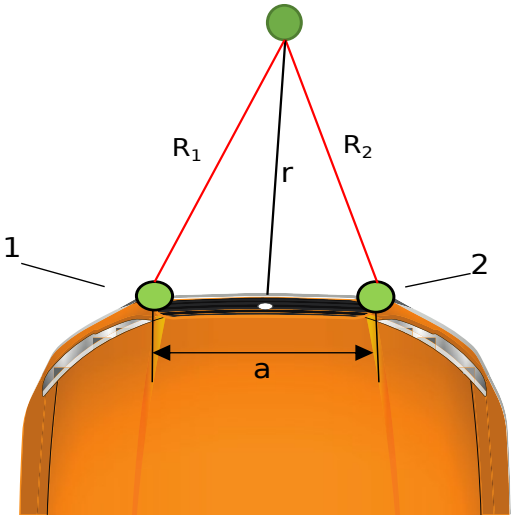

Figure 2.9: Ultrasonic sensor localization method.

where c_s is the speed of sound in air (approximately 340 m/s).

Calculation of the exact position of the object is performed with at least two sensors. As mentioned before, automotive applications use from two to four sensors. The localization of the object is calculated using the trilateration method, as presented in Figure 2.9.

$$R = \sqrt{R_1^2 \frac{(r^2 - R_2^2 + R_1^2)^2}{4r^2}}, \tag{2.15}$$

where r is the distance between corresponding sensors, R_1 is the distance from sensor 1 to the object, R_2 is the distance from sensor 2 to the object.

Construction

The sensor consists of a housing with a connector, ultrasonic converter, and circuit board with electronic circuitry for the transmission, reception, and evaluation of signals. The ultrasonic part consists of an aluminum part with a piezo element inside connected to a printed circuit board. A digital pulse from the Electronic Control Unit (ECU) is received by the sensor, after which the membrane is excited by an electronic circuit in the form of rectangular pulses at a resonant frequency. As a result of vibration,

ultrasonic radiation is generated. During signal generation (approx. 700 ms), the reception of the reflected signal is impossible. This limits the minimum measured distance to approximately 15–20 cm. After returning to the resting state, the membrane begins to vibrate. The piezoceramic plate converts these vibrations into an analog electrical signal, which is then amplified and converted into a digital signal by the sensor electronics. Sensors with a digital interface calculate the distance and transmit it to the ECU [1].

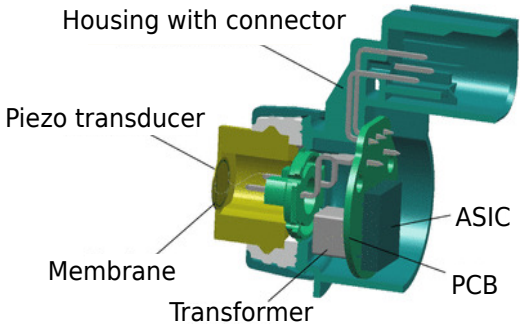

Figure 2.10: Ultrasonic sensor construction [1].

Ultrasonic sensors are already widely used in the automotive industry as part of parking assistance as well as collision avoidance systems. The main disadvantages of these sensors is their very short range and low operation speed, which make them efficient only at low speeds. Most manufacturers declare a range of 0.15 to 6 meters, which is suitable mostly for parking. As the signal generation time is approximately 0.7 seconds, it is impossible to use these sensors at speeds higher then 10 km/h. Based on the aforementioned characteristics, these sensors are efficient only for very short-field and low-speed tasks, such as automated parking.

2.1.5 Sensor Fusion

Modern AD projects use sensor fusion to enhance the performance of autonomous vehicles in general. Different sensors have advantages and disadvantages, such as their range of operation, resolution, and weather condition requirements. The purpose of sensor fusion is essentially to combine different types of sensors in a combination that will provide the

optimal total performance, through combining all of the advantages of the sensors and compensating for their weak points and disadvantages.

The combinations of Camera-RADAR (CR) as well as Camera-LiDAR-RADAR (CLR) are the most commonly used for environment representation in current projects.

The CR combination offers high-resolution images from cameras in addition to distance and velocity information of objects around the vehicle. For example, Tesla's automated driving system employs the CR sensor combination in addition to ultrasonic sensors [81].

The CLR sensor combination can be used in a larger range of operation with adequate resolution. The LiDAR point clouds and depth map information provide precise environment representation and improve the safety and reliability of the autonomous system in general. The CLR sensor combination for environment perception has been implemented in AD projects by companies such as Waymo and Navya.

In multi sensor data fusion three primary approaches exists for combining the sensory data from various sensors: High-Level Fusion (HLF), Low-Level Fusion (LLF) and Mid-Level Fusion (MLF) [6].

The principle of the HLF approach requires separate operation of the sensors; thus, fusion occurs only at the level of the output data of each sensor. For example, in [72] the authors proposed the use of a nonlinear Kalman filter for the independent and subsequent fusion of radar signals and LiDAR point cloud data. Due to its relatively low complexity compared with the LLF and MLF approaches, HLF is quite widely used. However, HLF provides poor quality and less precision of classification.

With the LLF approach, data fusion occurs at the lowest level of abstraction (i.e., raw data). In this case, all of the information is saved and can potentially improve the detection accuracy. The LLF approach is realized in 3D Cross View Fusion (3D-CVF) [86], which is a two-stage method. In the second stage, the LLF approach is used to fuse the combined camera and LiDAR feature maps received from the first stage. This method has been evaluated on KITTI and nuScenes datasets with outstanding object detection results. Nevertheless, LLF is associated with several implementation problems, such as the requirement for high precision of sensor calibration for accurate environment perception fusion. The sensors must also balance the ego-motion and be calibrated in time [6].

The MLF approach is essentially the middle level between HLF and LLF and is also known as feature-level fusion. The basic principle is to recognize and classify combined multi-sensor objects based on the fusion of features extracted from data from different sensors, such as color information from

cameras or radar and LiDAR location characteristics. For example, in [46] the authors presented an approach to object detection in a dynamic background environment with limited communication possibilities using a feature fusion scheme. A set of infrared sensors pointing in different directions is used to extract low-dimensional features, using a Symbolic Dynamic Filtering (SDF) algorithm. However, limited possibilities for environment perception and the loss of contextual information make it impracticable for MLF to match the SAE's Level 4 and 5 requirements.

Sensor fusion is a consensual method for increasing the safety, accuracy, and reliability of AD. Because all of the sensors have limitations, sensor fusion makes the system more universal and suitable for various operating conditions. However, the fusion of sensor data increases the costs and complexity of the system [84].

2.2 Sound Processing in Automotive Applications

Current AD research and projects use a variety of sensors to navigate and sense the environment. The previous section covered most of these (i.e., LiDAR, radar, cameras, and ultrasonic sensors). With the exception of Waymo, sound as a source of information is not used in the navigation and perception of an autonomous vehicle. Thus, it can be said that hearing is not part of the set of necessary senses for an autonomous vehicle. However, several applications and studies have used hearing for specific tasks where sound is the dominant characteristic. Several studies have suggested using sound to detect and identify abnormal sounds, such as sirens, alarms, and horns as well as for sound-based vehicle classification. Moreover, studies have demonstrated the possibility of using sound information to determine the road surface, since the contact between the tire with the road produces a sound that changes depending on the type of terrain. Furthermore, modern highways have special road markings, the crossing of which produces sound and vibration to attract the driver's attention. Some studies have considered the possibility of applying sound analysis when driving off-road and in other hard-to-reach places, where the effectiveness of other sensing and localization sensors is greatly reduced. Another problem that can be solved with the use of sound is self-diagnosis. Most car breakdowns are first detected by the driver through various abnormal noises, such as knocking, creaking, or grinding, which mean that the analysis of the ego-noise of the

car itself can help to identify various types of breakdown [50]. This section provides an overview of various sound processing techniques in automotive applications.

2.2.1 Emergency Vehicle Detection

In accordance with local regulations, which differ from country to country, all EmVs are equipped with light and sound warning devices to alert other traffic participants, such as a flashing beacon and a siren. The siren is a powerful generator of sound signals with certain frequency-time characteristics in accordance with various standards. In Germany, the parameters of the siren are determined by the DIN 14610 standard. A siren sound should consist of repetitions of low and high tones with a time period of (3 ± 0.5) seconds. The pause between two signals should not exceed 0.8 seconds; the tone frequencies should be in the range of 360–630 Hz; and the ratio between alternating tones should be 1.333. In rural areas, the frequency of sound signals can be reduced compared with the urban environment.

Automatic siren recognition has been reviewed in some studies. In [54], the authors presented a system based on a modified pitch detection method, which was realized in two parts. The first part involves the pitched–unpitched classification of each portion of the sound signal with Module Difference Function (MDF) and the estimation of the pitch frequency with the peak searching algorithm. In this part, a signal representing a pitch that changes over time $Pitch(t)$ is obtained.

In the second part, $Pitch(t)$ is analyzed to detect the presence or absence of a siren. This technique allows the avoidance of the influence of additional harmonics, which occur during the generation of an audio signal, on the recognition quality.

In [7] the authors proposed using the standard method for recognizing acoustic images. This method is based on the Mel-Frequency Cepstral Cofficient (MFCC) and Artificial Neural Network (ANN) which are well-known in the field of Automatic Speech Recognition (ASR). The ANN output was averaged over 400 ms windows. The dataset consisted of Italian EmV sirens and road noise. Although the method achieved an accuracy of 99% for SNR of 0 dB, the authors did not clearly indicate whether they considered the Doppler effect.

The detection of unknown alarms was also investigated in [15]. A comparative analysis of the standard ASR approach and sinusoidal modeling was conducted. The database consisted of various alarms and background

noises from the Internet. All tests were performed at an SNR of 0 dB. The authors reported that neither system worked well.

Fazenda et al. [16] suggested using a siren recognition and localization system to alert the driver and passengers through speakers inside the car. For the analysis, a cross-shaped microphone array was used, which recorded the sound signals of the environment. Furthermore, based on the received data, the siren signal was detected using the adaptive predictor noise canceler, which is based on the Least Mean Squared (LMS) algorithm, thus separating the siren signal from the mixed ambient noise. In the next stage, the Direction of Arrival (DOA) of the sound source was determined by the time delay estimation method based on the Generalized Cross Correlation (GCC) algorithm. The authors also used the sound intensity probe method to determine the DOA, but found it to be less accurate. Subsequently, the driver was notified of an incoming EmV through four speakers installed in the car. The DOA was indicated by volume; that is, the speaker in the same direction played the siren sound louder than the speakers in other directions, as illustrated in Figure 2.11.

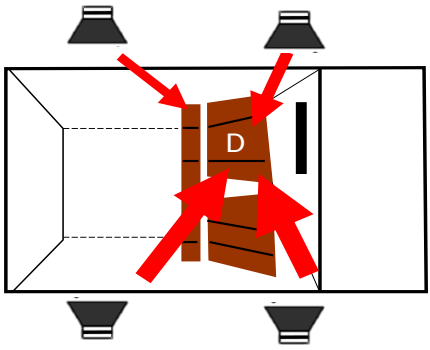

Figure 2.11: Driver notification principle.

The recognition of sirens using machine learning was first considered in [69]. The study was conducted on the high-low siren type of German EmVs, which conforms to the DIN 14610 standard. The authors proposed using a modified Part-Based Models (PBM) applied to the Mel spectrum of the siren signal. After performing the experiments in various conditions, the authors compared the results with the HMM-MFCC model. The results revealed that compared with Hiden Markov Models (HMM), the PBM approach provided a higher degree of modeling flexibility, particularly in a

noisy environment. Nevertheless, when the SNR value dropped below -5 dB, the performance was critically reduced.

In [55], the categorization of ambient sounds into five classes was proposed: siren, train crossing bell, screeching tire, car horn, and glass breaking sounds. The Gaussian Mixture Models (GMM) was used to classify sound events in conjunction with the MFCC and derivatives such as delta and delta coefficients. The proposed principle was to combine the GMM-supervector system and the GMM-based Universal Background Model (UBM) for use in voice-based speaker detection. The audio signals were represented as a supervector, which was obtained by combining average GMM vectors of audio segments and adapting them to the UBM using the Maximum A Posteriori (MAP) approach [12]. A Probalistic Principal Component Analysis (PPCA) model and a Linear Discriminant Analysis (LDA) projection were then applied to reduce the dimensionality. Then, based on the cosine distance, the final classification was obtained. Validation of the approach was conducted on a set of audio data collected from the Internet. The results demonstrated high robustness of the structure based on the GMM supervector with a discriminative internal classifier.

In [52], the authors proposed improving classification accuracy using preliminary noise removal. This method consisted of two parts. First, anomalies were detected in the audio signal through single-class Gaussian processes [38]. If an anomaly was present, then the event was classified using the k-nearest neighbor k-Nearest Neighbor (kNN) structure. The concept of Empirical Binary Masks (EBM) was applied before classification. EBM, similar to ideal binary masks [80], aims to remove unwanted masking signals from a noisy mixture. K-means segmentation was applied to the gammatonogram [49] of a noisy signal for EBM generation. Categorization was performed on five classes, such as sirens, horns, and the sound of pedestrian traffic lights. The evaluation was performed by comparing the performance of the k-NN structure when working in conjunction with EBM, the original noisy gammatonograms, and the MFFC. This approach was highly effective, but the accuracy was greatly reduced at low SNR values. The authors assumed that the signal of interest was the signal with the highest power. This is true in most cases since all warning signals are powerful and loud; however, this assumption is not applicable to low SNR values.

The same authors in [51] suggested using deep learning to recognize sirens. Audio signal classification was performed through the simultaneous segmentation of the gammatonogram using a multi-task deep learning architecture. Next, cross-correlated segmented gammatonograms were

transmitted to a Deep Convolutional Neural Network (DCNN) to estimate the direction of sound arrival. This method exhibited high efficiency, including at low SNR values from -40 to 10 dB. It can be particularly useful for solving some problems with AD, where the ability to work in noisy conditions is crucial for practical applications on public roads [50].

2.2.2 Vehicle Detection and Classification

In the current state of the art, vehicle classification is applied for traffic monitoring purposes and consists of car/truck classification.

In the study [27], the authors proposed combining audio and video training approaches based on Naive Bayesian classifiers (NBc), he primary training of which was conducted on labeled images. Next, based on the confidence level of audio and visual models in combination, labels for new samples were repeatedly generated. Evaluation of the proposed approach was conducted on audio-video data recorded on a bridge over a road by both a camera and a microphone simultaneously, which demonstrated high efficiency. An approach with a similar functionality was presented in [9], with the difference that the audio classifier acted as an autonomous supervisor, which was used to support continuous training of the visual online classifier. The efficiency assessment process was based on real data collected from multi-lane highways and exhibited high efficiency.

The algorithms presented above use combined audio and video methods for classification. There are also detection techniques based only on sound. In [34],an approach was presented that considered the creation of sound maps based on an audio signal from two microphones installed at the side of the road. This approach allowed the determination of the speed based on audio data. A sound map was created through the determination of the time delay between two microphones through a generalized cross-correlation. One of the approaches was based on Dynamic Time Warping (DTW) [35], which did not allow sequential or simultaneous detection since it does not have the ability to distinguish overlaps on the sound map. In a newer approach, the same authors proposed an algorithm that removes points from the sound card that correspond to the identified vehicle, thereby reducing interference for the following vehicles. The relationship between the sound map and the vehicle was achieved using the Random Sampling Consensus (RANSAC) algorithm [17]. This approach provides high detection accuracy, including when working with sequential vehicles.

2.2.3 Road-Tyre Interaction and Terrain Classification

Several studies have considered the dependence of the sound produced by the car when driving and the surface on which it moves. In [24] the authors confirmed that the noise level depends on the surface, humidity, speed, and type of vehicle. When water was present on the road surface, a significant increase in noise levels occurred, and some differences were also noted between the sounds produced by cars and those produced by trucks when driving. This was proven in [18] when the authors conducted a comparative analysis of the noise produced by passenger and cargo cars and confirmed the presence of some differences. In most cases, the reviewed studies have employed the Controlled-Pass-By (CPB) and Statistical-Pass-By (SPB) methods in noise measurements in accordance with ISO 11819-1 [36]. However, both methods are designed to measure road noise around highways and assume specific infrastructure.

The variation of self-noise generated by the interaction of tires with road markings has been investigated in [20, 21] on seven different types of road markings as examples. The following were chosen for comparison: irregular scattered dots, irregular dense structure, irregular longitudinal structure, regular wide drops, regular dense dots, regular narrow drops, and irregular perforated plate structure, with stone mastic asphalt as a reference. Experiments revealed a considerable rise in sound pressure levels at the lower end of the frequency spectrum (i.e., 800–1000 Hz), which was caused by the interaction of the tire with the road markings. Thus, it can be concluded that an autonomous vehicle has the ability to track the intersection of road markings based on acoustic data.

In [41], the authors presented an approach for the automatic classification of road status. They distinguished wet, dry, and snow-compacted surfaces using data from microphones located on the roadside through a multimodal structure based on audio-visual information and neural networks. Furthermore, systems for the onboard determination of road surface humidity based on machine learning were proposed in [2]. The systems can be used to enable autonomous vehicles to adopt an adequate driving strategy depending on current road conditions. In [3], the Support Vector Machine (SVM) classifier was employed for road status determination. Sound was recorded using microphones located on both sides of the rear wheels of the vehicle, thus reducing the influence of the ego-noise of the engine. In [2], the microphone was located behind the rear wheel and data was collected under various driving conditions, speeds, and road

surface qualities as an extension of the set presented in [3]. The authors used Recurrent Neural Network (RNN) in the form of Long-Short-Term Memory (LSTM) and Biderectional LSTM (BLSTM) [32], and obtained impressive accuracy.

In addition, several approaches exist for acoustic-based road surface classification. The first was proposed by Odedra in [56]. The first working concept of such a system was presented in [47], where SVM classifiers were used to classify several types of surface, such as grass, pavement, gravel road, and water, as well as impacts on hard objects. The system also allowed recognition of the loss of grip on slippery roads. A similar study using deep learning was proposed in [74, 75] for use in mobile wheeled robots. The authors analyzed noise on nine various surface types: asphalt, cut grass, medium to high grass, paving stones, cobblestone, off-road, wood, linoleum, and carpet. The classification was based on a DCNN, which was trained on spectrograms of prerecorded audio signals of driving on different surfaces. It also considered the temporal evolution of audio signals by introducing the LSTM structure. The system exhibited high efficiency up to 98%, including on noisy samples recorded under real conditions.

2.3 Weak Points of Current Setups

Limits of visual processing

Although computer vision algorithms are able to extract much information, such as the 3D positions of single points, 3D poses of objects, movement of the objects, and classes of many different objects highly accurately, visual systems are still not robust enough.

In certain weather conditions, computer vision algorithms are faced with many problems that they cannot handle. If it is very sunny, for example, multiple shadows appear and their projections onto the camera start moving alongside moving objects, or the ego-motion of the car moving relative to a static object casts a shadow onto the road. Furthermore, in fog, snow, or rain, the reliability of computer vision algorithms heavily decreases simply because the contrast of the scene is reduced. Finally, the sensing range is limited to the FoV and the resolution of the projection of an object of constant size decrements is inversely proportional to the distance. Hence, object detection and classification only work in the near field. The same limits apply to the resolution of depth estimates of multiple camera 3D reconstructions because the depth of any projected 3D point is

inversely proportional to the disparity; moreover, the proportional factor is limited by the baseline between the cameras for a given focal length and pixel size [75]. In sum, localization by cameras and proper object detection only work for certain working ranges between 100 m and 150 m, whereas the measurement resolution within the working range reduces with distance. The position of the working range relative to the car depends on the focal length of the camera. The larger the focal length, the smaller the FoV.

Limits of LiDAR measurements

As mentioned above, LiDAR is a laser-based sensor that can monitor 360° around the vehicle and capture high-definition data from the environment as point clouds. This sensor is mainly used for the detection and 3D localization of multiple objects in a larger distance range up to 250 m and a fairly constant measurement resolution within that range compared with cameras. In harsh weather conditions, the performance of LiDAR drops dramatically, especially in foggy environments because the laser beam disperses and more power is required to receive the main reflection. In [8, 31], comparative analyses of the most popular lidars were provided, such as the Velodyne VLP-16, HDL-64S2, and HDL-64S3; Ibeo LUX and LUX HD; as well as Valeo Scala. Studies have demonstrated that all of the current lidars on the market are highly dependent on weather conditions. For example, fog and heavy rain will dramatically decrease their detection range and classification capabilities.

For safety reasons, the power of LiDAR sensors is limited to an average power of 2 mW, which equals safety category 1. However, the classification of point cloud objects remains challenging. The sampling of an object with a fixed size sampled by a laser ray decreases inversely proportionally with the distance until the number of object points is too small for classification. Additionally, same as for digital cameras, the occlusion of visible parts of an object introduces the same problems for classification tasks.

Limits of radar measurements

Radar is a sensor for object detection based on radio waves. It is used for detecting the objects in front of a car. Modern long-range radars provide a distance range of detection up to 250 m. In addition, long-range radars have a small width of detection in a short range, which creates the necessity of using both types of radars simultaneously. Compared with LiDAR, radar

has better performance in foggy environments, but in rainy conditions the number of false detections increases while the distance of operation becomes shorter [71]. Classification by radar as well as detection in general has the same problems as lidar if objects are occluded. Due to range-based differentiation, for a wide field of coverage, both short- and long-range radars should be installed and operated simultaneously.

Common limitations

Since all current autonomous vehicle perception systems are, in one way or another, reproductions of human vision, they share common weaknesses.

The main problem inherent in vision systems is of course occlusion. Like the driver, vision-related systems cannot detect, classify, or localize objects that are out of sight (e.g., those blocked by another vehicle). In some cases, the driver is able to see the brake lights of the occluded vehicle through the windows of the vehicle in front.

Thus, in dense urban traffic or in a traffic jam, an autonomous vehicle's sensor system will only have access to data on vehicles and other objects in the immediate area, which can lead to errors in traffic forecasting and the selection of the wrong strategy for further action.

In the context of EmV detection and classification, lidars and radars are completely unable to provide such functions. Nevertheless, several approaches to camera-based EmV detection have been proposed [37, 62, 64]. Because vision systems can only detect and classify objects from the FoV, these systems are based on the detection of the flashing lights of EmVs. This approach is the only EmV detection facility in the current state of the art. The main disadvantage of visual EmV detection is that the beacon must be in the FoV of the camera and not be occluded by other traffic participants; therefore, in inner-city traffic, the range of detection would not be longer then 10–20 m.

Figure 2.12: Illustration of possible occlusions.

3 Sound Processing

3.1 Sound Source Localization

Sound Source Localization (SSL) is a widely researched topic in the robotics community. Localization algorithms are used in mobile robotics, where robotic audition is one of the components of the human - machine interface.

3.1.1 Direction of Arrival Estimation

Localization of the sound source is usually performed using two main methods. The first is beamforming, while the second is a method based on the Time Difference Of Arrival (TDOA).

Beamforming method: The principle of operation of the beamforming method is to weigh the spatial signal, amplify it in all directions and determine the DOA, for which the direction with the highest power is taken. In 2000, DiBiase proposed the Steered-Response Power with PHAT (SRP-PHAT) as an implementation of the beamforming method. SRP-PHAT has several advantages, such as resistance to noise and echo and reliability.

However, real-time implementation requires a large amount of computing power. Computational limitations make it unsuitable for real-time applications. Therefore, several acceleration and optimization approaches have been proposed to improve SRP-PHAT, such as reducing the two-dimensional search space to a pair of one-dimensional ones [11], or an acceleration method that uses spatial and frequency implementation of coarse-to-fine strategies [88].

Method based on TDOA: TDOA is a GCC based method for the localization of sound sources proposed by K. Knapp in 1976 [40]. It includes the application of several frequency weighting functions, such as Maximum Likelihood (ML), the Roth autocorrelation weighting function (ROTH) and the Phase Transformation weighting function (PHAT).

The principle of the algorithm's operation can be divided into two phases. First, the time difference for each microphone is calculated, and then, based

on the geometric relationship of the microphones and the known time delays, the localization of the sound source is performed [78].

Generalized Cross Correlation with Phase Transform weighting function (GCC-PHAT)

The TDOA method can be explained using the example of a dual-microphone array, since it is a pairwise technique. In application to linear and circular arrays, this method is based on pairwise array processing. Suppose that two signals $s(t)$ and $n(t)$ are not related to each other and represent the signal of the sound source and the signal of noise source, respectively. Without considering reverberation, the signals received by the pair of microphones are as follows:

$$x_1(t) = \alpha_1 s(t - \tau_1) + n_1(t), \tag{3.1}$$

$$x_2(t) = \alpha_2 s(t - \tau_2) + n_2(t), \tag{3.2}$$

where α_1 and α_2 are the attenuation coefficients of the sound propagation to MIC_1 and MIC_2 respectively, τ_1 and τ_2 are the propagation time to MIC_1 and MIC_2.

Considering the reverberation effect, in addition to signals related to the original sound and noise, microphones accept reverberation, which in fact comprises multiple reflections of the original signals from various surfaces and obstacles. In this case, the microphone receives the following signals:

$$x_i(t) = h_i(t) * s(t - \tau_i) + n_i(t) = \alpha_i s(t - \tau_i) + \sum_{p=0}^{\infty} \alpha_{ip} s(t - \tau_{ip}) n_i(t), \tag{3.3}$$

$$x_j(t) = h_j(t) * s(t - \tau_j) + n_j(t) = \alpha_j s(t - \tau_j) + \sum_{p=0}^{\infty} \alpha_{jp} s(t - \tau_{jp}) n_j(t), \tag{3.4}$$

where $h_i(t)$ and $h_j(t)$ are the impulse response function of environment to signal, α_{ip} and α_{jp} are the attenuation coefficients that the sound source signals reach the MIC_i and MIC_j after the p-th reflection, τ_{ip} and τ_{jp} are the time when the original sound travels to MIC_i and MIC_j after the p-th reflection.

Because the signal model also depends on the distance to the sound source, it can be conditionally divided into models of far and near fields according to Equation (3.5).

$$R = \frac{2L^2}{\lambda}, \tag{3.5}$$

where R is the distance from the sound source to the center of the microphone array, L is the diameter of the microphone array, λ is the wavelength of sound.

The near-field model is applicable if $R < \frac{2L^2}{\lambda}$. For the near-field model, the sound wave is considered to be spherical and to propagate in an outward direction from the sound source. In this case, it is necessary to consider the amplitude difference of the signals. Accordingly, in the far-field model, the amplitude of the signals is ignored, and only the time delay of the signals is considered as it is assumed that the wave is flat. This model is a simplification of the real signal model and applicable if $R > \frac{2L^2}{\lambda}$.

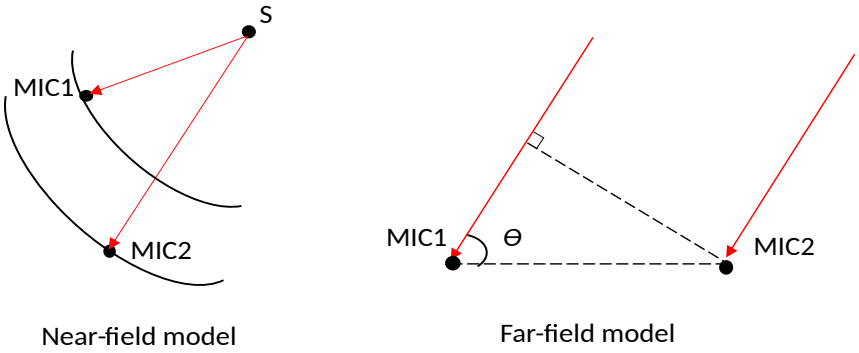

Figure 3.1: Near- and far-field propagation models.

The time delay in accordance with the near- and far-fields models can be calculated using (3.6) and (3.7), respectively:

$$\tau_1 = \frac{|P - M_2| - |P - M_1|}{c}, \tag{3.6}$$

$$\tau_2 = \frac{L\cos(\theta)}{c}, \tag{3.7}$$

where P is the position of sound source, M_1 and M_2 are the positions of microphones, c are the speed of sound.

The time delay estimation is determined by the cross-correlation of signals received by different microphones. The TDOA algorithm provides sufficiently high accuracy when determining time delays as well as a relatively simple implementation and few calculations. These advantages allow this algorithm to be used for real-time applications.

Suppose that the signals received by different microphones in the array were emitted by a single source. Then, calculating the cross-correlation function of these two signals would result in a delay in their arrival time. This time delay is caused by the difference in the distance traveled by the sound from the source to the array. Calculating the difference between these distances results in the estimated direction to the sound source.

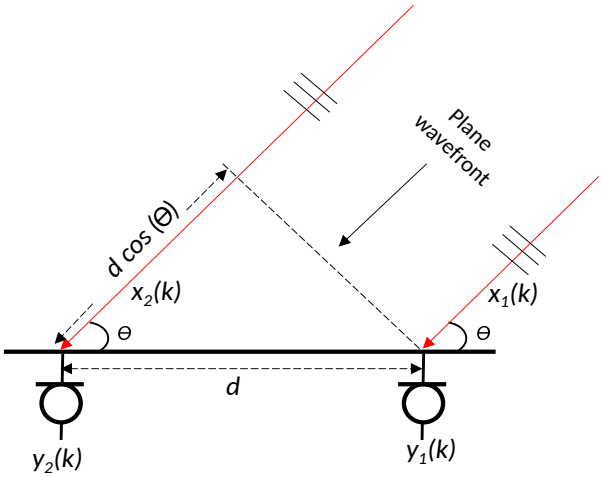

Figure 3.2: Sound propagation model in far field applications.

The cross-correlation function is a signal processing measure represented by the degree of correlation of two time series. The cross-correlation function $G_{x_1x_2}(\tau)$ for two signals $x_1(t)$ and $x_2(t)$ can be expressed as follows:

$$G_{x_1x_2}(\tau) = E(x_1(t)x_2(t-\tau)). \tag{3.8}$$

The GCC function (3.9) is applied to the signal models from Equation (3.3) and (3.4):

$$G_{x_1x_2}(\tau) = \alpha_1\alpha_2 E(s(t-\tau_1)s(t-\tau_2-\tau))+ \\ + \alpha_1 E(s(t-\tau_1)n_2(t-\tau))+ \\ + \alpha_2 E(s(t-\tau_2-\tau)n_1(t)) + E(n_1(t)n_2(t-\tau)). \quad (3.9)$$

The simplification of Equation (3.9) is based on the assumption that $n_1(t)$ and $n_2(t)$ are uncorrelated Gaussian white noise, hence:

$$G_{x_1x_2}(\tau) = \alpha_1\alpha_2 E(s(t-\tau_1)s(t-\tau_2-\tau)) = \alpha_1\alpha_2 G_S(\tau-(\tau_1-\tau_2)). \quad (3.10)$$

The value of the time delay between two microphones is calculated by searching for the maximum value of $G_{x_1x_2}(\tau)$, which is achieved when $\tau = (\tau_1 - \tau_2)$. Under the same environmental conditions, such as temperature and pressure, the speed of sound will be constant, therefore, based on the time delay τ, the difference in the distance to the sound source for each microphone, as depicted in Figure 3.2, can be calculated as follows:

$$d\cos(\theta) = c\tau. \quad (3.11)$$

Due to the high computational complexity of calculating the cross-correlation function in accordance with convolution in the time domain, the calculations are performed in the frequency domain, since convolution in the time domain corresponds to multiplication in the frequency domain. The cross-correlation function and the cross-power spectrum are related as follows:

$$G_{x_1x_2}(\tau) = \int_0^x S_{x_1x_2}(\omega)e^{-j\omega\tau}d\omega = \int_0^x X_1(\omega)X_2^*(\omega)e^{-j\omega\tau}. \quad (3.12)$$

The accuracy of time delay estimation is reduced when noise and reverberation occur. The peak value of $G_{x_1x_2}(\tau)$ ceases to be highly explicit. To suppress interference and reverberation as well as to increase peak allocation, weighted functions can be applied to the cross-power spectrum in the frequency domain. Next, the generalized cross-correlation function $G_{x_1x_2}(\tau)$ is obtained by performing the inverse Fourier transform:

$$G_{x_1x_2}(\tau) = \int_0^x \varphi(\omega)X_1(\omega)X_2^*(\omega)e^{-j\omega\tau}d\omega. \quad (3.13)$$

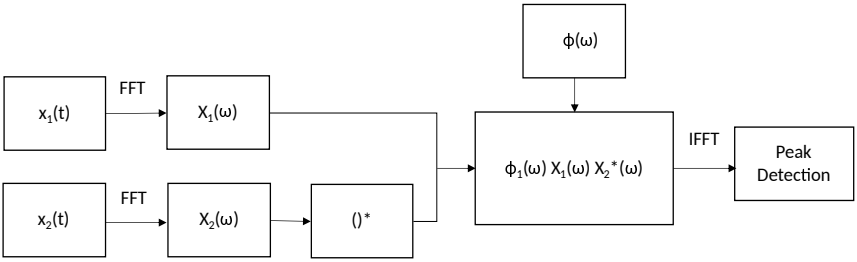

Figure 3.3: Algorithm of time delay estimation by GCC.

The process of estimating the time delay is depicted in Figure 3.3. Since the amplitude information is not considered, the estimate depends on the phase information of the cross-spectrum. The method of using phase transformation is also known as GCC-PHAT, where the weight function of the phase transformation can be expressed as follows:

$$\varphi(\omega) = \frac{1}{|S_{x_1x_2}(\omega)|}. \tag{3.14}$$

By applying function (3.14) to Equation (3.13), the GCC-PHAT equation is obtained:

$$G_{x_1x_2}(\tau) = \int_0^x \frac{X_1(\omega)X_2^*(\omega)}{|X_1(\omega)X_2^*(\omega)|} e^{-j\omega\tau} d\omega. \tag{3.15}$$

In essence, the PHAT weight function is a type of bleaching filter that smooths the cross-power spectrum, thus increasing the sharpness of the generalized cross-correlation function. The expression of the generalized cross-correlation function $G_{x_1x_2}(\tau)$ after weighting PHAT has the following form:

$$G_{x_1x_2}(\tau) = \alpha_1\alpha_2\delta(\tau - \tau_{12}). \tag{3.16}$$

After weighting PHAT, the cross-power spectrum has a similar expression with a single pulse response. The peak of the time delay can be highlighted, which contributes to reverberation and noise suppression as well as to improved estimation accuracy. The GCC algorithm is computationally efficient and can be executed almost instantly. It is suitable for application to AD systems due to its strong tracking ability. However, its operation in a highly reverberated environment is problematic, yet the range of applications is still quite wide.

Once the time delay has been determined, the sound source is localized based on the received data. This is the second stage of the TDOA. The time delay τ is caused by the difference in the distance that the sound travels from the source to each of the microphones. Therefore, if the speed of sound is c, then the difference in distance is:

$$c\tau = |P - M_2| - |P - M_1|, \tag{3.17}$$

where P is the position of sound source, M_1 and M_2 are the positions of microphones, c is the speed of sound.

Equation (3.17) determines the constancy of the difference in distance between the microphones and the sound source. Consequently, the expected sound source position lies in one of the hyperbola branches with a focus on the position of the microphone.

Steered-Response Power with Phase Transformation weighting function (SPR-PHAT)

The SRP-PHAT algorithm is based on a combination of steered-response power and phase transformation. This algorithm combines several advantages of phase transformation and SRP itself, such as robustness and short-term analysis, with low sensitivity to environmental characteristics. The principle is to search for the maximum value of power in all of the expected directions of the sound signal's arrival using delay and sum units.

The SRP-PHAT beamformer output power is expressed as follows:

$$P(q) = \sum_{k=1}^{N} \sum_{l=k+1}^{N} \int_{-\infty}^{\infty} \varphi(\omega) X_k(\omega) X_l^*(\omega) e^{-j\omega\tau}, \tag{3.18}$$

where $X_k(\omega)$ and $X_l(\omega)$ are the Fourier transform of the $k-th$ and $l-th$ microphone signal multiplied by window function, τ is the steered delay between the array and the sound source, q is the spacial vector of the sound source.

The PHAT weighting factor $\varphi(\omega)$ is expressed as follows:

$$\varphi(\omega) = \frac{1}{X_k(\omega) X_l^*(\omega)}. \tag{3.19}$$

From Equation (3.18), it follows that SRP-PHAT represents the sum of the cross-correlations of all possible pairs of microphones in the array. Calculating the cross-correlation between all pairs requires a large amount

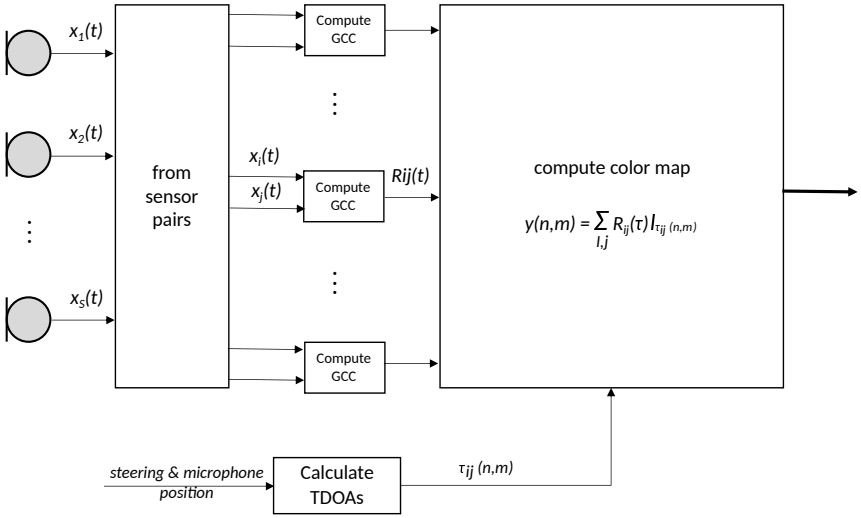

Figure 3.4: SRP-PHAT algorithm.

of computations which increases with the number of microphones in the array. The SRP-PHAT algorithm in the time domain can be expressed as:

$$P(q) = 2\pi \sum_{k=1}^{N} \sum_{l=k+1}^{N} G_{kl}(\tau_l - \tau_k), \tag{3.20}$$

where G_{kl} is the cross-correlation function of the signal from the $k-th$ and $l-th$ microphone weighted by PHAT.

The SRP-PHAT operating principle is the peak-searching of $P(q)$ in all directions of possible sound arrival, and can be expressed as follows:

$$\hat{q}_s = \arg\max_{q} P(q). \tag{3.21}$$

Consequently, the DOA can be calculated through peak value determination. Despite high resistance to noise and reverberation, the effects of interference remain. Extraneous peaks unrelated to the original source can occur, which can make the original peak less obvious.

3.1.2 Distance Estimation

Triangulation

As mentioned before, a microphone array is able to determine the DOA of an audio signal. Specifically, the triangulation method is applied, where two arrays are used and the DOA is determined for both of them, resulting in the position of the sound source being determined.

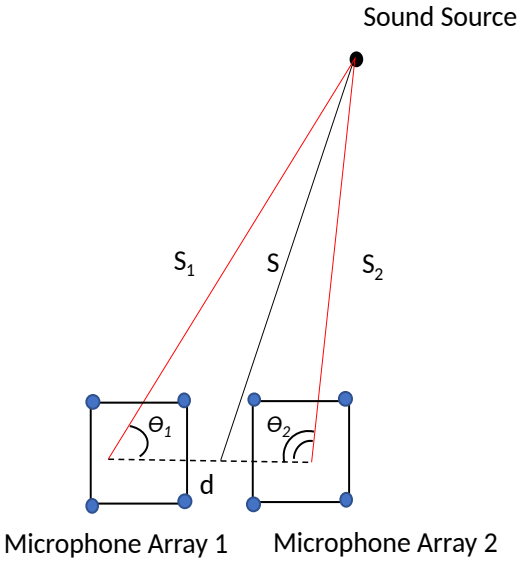

Figure 3.5: Triangulation localization method.

According to the law of sines, by determining the two DOA angles of the sound signal θ_1 and θ_2 for two arrays of microphones and knowing the distance between the centers of these arrays d the distance to both of arrays can be calculated as follows:

$$\frac{d}{\sin(180-\theta_1-\theta_2)} = \frac{R_1}{\sin(\theta_2)}, \tag{3.22}$$

$$\frac{d}{\sin(180-\theta_1-\theta_2)} = \frac{R_2}{\sin(\theta_1)}, \tag{3.23}$$

$$R_1 = \frac{d\sin(\theta_2)}{\sin(180-\theta_1-\theta_2)}, \tag{3.24}$$

$$R_2 = \frac{d\sin(\theta_1)}{\sin(180 - \theta_1 - \theta_2)}. \tag{3.25}$$

The distance between the midpoint of the two microphone arrays and the sound source can be calculated in accordance with the law of cosine as follows:

$$R^2 = R_1^2 + (\frac{d}{2})^2 - 2R_1\frac{d}{2}\cos(\theta_1), \tag{3.26}$$

$$R = \sqrt{R_1^2 + (\frac{d}{2})^2 - 2R_1 d\cos(\theta_1)}. \tag{3.27}$$

Theoretically, the triangulation method can be applied to determine the exact position of the sound source. This method also implies a sufficiently high resolution for DOA estimation. For example, with a real distance of 40 m, an error of only 0.3° in the DOA calculation results in an error of approximately 11 m in the distance estimation, which in this case is more than 25%. Therefore, the triangulation method applied to far-field applications cannot be considered satisfactory.

Method of Sound-Level Reduction

The principle of the sound level reduction method is based on the analysis of sound pressure caught by the microphone array. Sound pressure is variable overpressure that arises in the medium. It is caused by the passage of a sound wave through it, which is measured in Pascals. The transmitting medium alternately contracts and expands with the propagation of the sound wave; therefore, the deviation takes a positive and negative value in turn. Usually, the effective sound pressure is taken as the sound pressure value, that is, the Rated Maximum Sinusoidal (RMS) value of the pressure, which is always positive. The peak pressure P_{peak} and the RMS pressure P_{rms} are related as follows:

$$P_{rms} = \frac{P_{peak}}{\sqrt{2}}. \tag{3.28}$$

The human auditory system is most sensitive in the frequency range of 2000–5000 Hz , where the audibility threshold p_0 is considered to be 2×10^{-5} Pa. The sound pressure p is inversely proportional to the distance r. Sound pressure is related to the distance covered by the sound wave as follows:

$$p \propto \frac{1}{r}. \tag{3.29}$$

Hence, the sound pressure p is inversely proportional to the distance r. The expression of captured sounds through sound pressure may have a value range of more than six orders of magnitude. Therefore, the sound pressure is expressed through the Sound Pressure Level (SPL), which is a logarithmic scale relative to a reference value:

$$SPL = 20 \lg \frac{p}{p_0} dB. \tag{3.30}$$

Consequently, the distance can be determined based on the relationship between the sound pressure and the distance covered by the sound wave. If the sound pressure level L_{p1} at distance r_1 is known, then at distance r_2 the sound pressure level L_{p2} can be calculated as follows:

$$L_{p2} = L_{p1} - 20 \lg \frac{r_1}{r_2} dB. \tag{3.31}$$

Sound intensity is the transmitted sound power per unit time in the direction of propagation across a unit of space perpendicular to the direction of propagation. Essentially, intensity is the objective physical measure of sound strength in W/m^2 corresponding to the loudness, which is a subjective measure perceived by an individual. Sound intensity is related to sound pressure as follows:

$$I = pv, \tag{3.32}$$

where v the particle velocity. Sound intensity is a vector pointing in the direction of the average energy flow, and can also be expressed in terms of the acoustic impedance z of the transmission medium, measured in Rayl, as follows:

$$I = \frac{p^2}{z}. \tag{3.33}$$

The audible threshold of sound intensity can be calculated from the acoustic impedance of the air ($z = 400$ Rayl) and the audible threshold of sound pressure ($p = 2 \times 10 - 5$ Pa) as follows:

$$I_0 = \frac{(2 \times 10^{-5})^2}{400} = 10^{-12} W m^{-2}. \tag{3.34}$$

The intensity of a spherical sound wave in the radial direction is expressed according to the inverse square law as

$$I(r) = \frac{P}{A(r)} = \frac{P}{4\pi r^2}, \tag{3.35}$$

where $A(r)$ is the area of sphere with radius r, and P is the sound power. The sound intensity I is inverse square proportional to the distance r.

$$I \propto \frac{1}{r^2}. \tag{3.36}$$

Similar to the sound pressure level, sound intensity can be expressed through the Sound Intensity Level (SIL) as:

$$SIL = 10 \lg \frac{I}{I_0} dB, \tag{3.37}$$

where $I_0 = 10^{-12} Wm^{-2}$ is set as the reference sound intensity. For example, $SIL = 20dB$ corresponds to $\frac{I}{I_0} = 100$.

According to the sound pressure and sound intensity, the distance to the sound source can be estimated as follows:

$$r_2 = r_1 \frac{p_1}{p_2} = r_1 \frac{I_1^2}{I_2^2}. \tag{3.38}$$

According to the SPL and SIL, the distance to the sound source can be estimated as follows:

$$r_2 = r_1 \cdot 10^{(\frac{|L_{p1} - L_{p2}|}{20})} = r_1 \cdot 10^{(\frac{|L_{I1} - L_{I2}|}{20})}. \tag{3.39}$$

This method of distance estimation is based on a comparison of original values of the sound source with values of the received signal. Thus, this method is suitable only for sound sources with known original characteristics, such as EmV sirens or horns. Therefore, this method is applicable only in several situations.

3.2 Environmental Sound Classification

The earliest definition of environmental sounds appeared in a PhD thesis from Cornell University by N.VanDerveer in 1979. In the thesis, the following four characteristics were proposed for identifying environmental sounds [76]:

- *"It is produced by real events;*
- *It has meaning by virtue of causal events;*
- *It is more complicated than laboratory-generated sounds such as pure tones;*
- *It is not part of a communication system such as speech."*

Most research related to sound recognition is based on speech recognition. Some of those speech recognition systems perform quite well, but only in a specialized domain. With environmental sound, they perform unsatisfactorily. Time Frequency Feature Extraction (TFFE) and ASR are the most discussed topics related to environmental sound recognition.

3.2.1 Mel-frequency Cepstral Coefficients (MFCC)

The use of MFCC methods to solve the environmental sound recognition problem was proposed in [13]. This method later became widely used in speech recognition. Four tests were conducted with this method on a database that contained 23 different sounds, and the accuracy was close to 100%. However, some problems could occur due to sound source separation [28].

Similar to any vibrations, sound can be represented by the sum of harmonic vibrations. Thus, the signal can be represented by a set of coefficients of harmonic signals of different frequencies. Such a representation is called a spectrogram. One approach to such decomposition is Fourier transform – an operation that combines one function of a real variable with another function of a real variable. The new function describes the coefficients (amplitudes) when decomposing the original function into harmonic components of different frequencies. A Discrete Fourier Transform (DFT) is used to decompose a discrete signal:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn}, \qquad k = 0, ..., N-1, \tag{3.40}$$

where $x_0, x_1, ...x_N$ are the signal amplitude values in the interval, $X_0, X_1, ..., X_N$ are the complex values of the amplitudes of sinusoidal signals, k is the frequency index.

The value of the frequency of the k-th signal is equal to $\frac{k}{T}$, where T is the duration of the time interval of transformation.

DFT decomposes a discrete function into a sum of sinusoidal signals with frequencies from one oscillation per period to N oscillations per period. A feature of discrete transformation is that the high-frequency component of decomposition cannot be correctly represented.

DFT allows one to represent a signal as a set of numbers. However, the description of the entire signal with just one set provides low approximation accuracy. To increase the accuracy, the original signal is divided into several equal parts (windows), and a transformation is applied to each of them. Several sets of coefficients are obtained, each of which characterizes only a part of the original signal. These sets are combined into a frequency spectrogram – a matrix of numbers, where rows are different frequencies and columns are different time windows into which the signal has been split. The number at the intersection of row i and column j determines the amplitude of the i-th frequency in the j-th window.

Human hearing is more sensitive to low-frequency sounds and less sensitive to high-frequency sounds. Considering this peculiarity, there is a disadvantage in representing sound in the form of a frequency spectrogram, namely the use of equal intervals between frequencies. With this approach, the frequency density is ineffective when considered in relation to human hearing. This representation is oversampled at high frequencies. A Mel spectrogram can be used as another signal representation that does not have this disadvantage.

Mel is a psychophysical value for measuring pitch. The quantitative assessment of pitch is based on the statistical processing of a large amount of data on the subjective perception of the pitch of sound tones. This dependence is described as follows:

$$f_{mel} = 2595 \log_{10}(1 + \frac{f}{700}), \tag{3.41}$$

where f is the physical frequency in Hz, f_{Mel} is the perceived frequency [14].

The initial data of the transformation is the frequency spectrogram obtained by the DFT - $X_k, k = 1, ..., N$, where N is the number of signals of various frequencies that form the spectrogram. The result of the transformation is the Mel-spectrogram. The conversion of ordinary frequencies to the Mel-scale is performed using a set of M overlapping triangular windows.

The Mel spectrum of the amplitude spectrum $X(k)$ is calculated by multiplying the spectrum by each of the triangular Mel weighting filters:

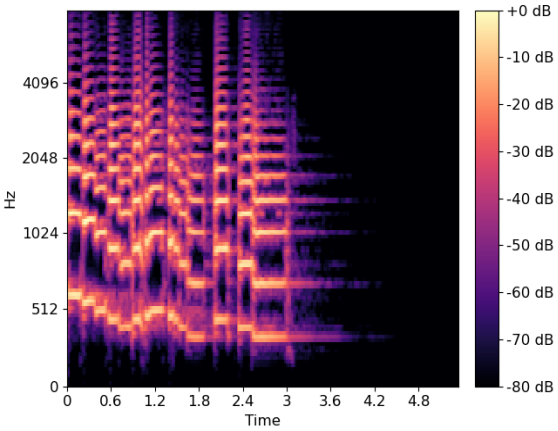

Figure 3.6: Mel specrtrogram.

$$s(m) = \sum_{k=0}^{N-1} [|X(k)|^2 H_m(k)], \qquad 0 \leq m \leq M-1, \tag{3.42}$$

where M is the total number of triangular Mel weighting filters [23, 87], $H_m(k)$ is the weight assigned to the k-th cell of the energy spectrum contributing to the m-th output band, which is expressed as follows:

$$H_m = \begin{cases} 0, & k < f(m-1), \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1) \leq k < f(m), \\ \frac{f(m+1)-k}{f(m-1)-f(m)}, & f(m) \leq k \leq f(m-1), \\ 0, & k > f(m+1), \end{cases}$$

where $0 \leq m \leq M-1$.

The MFCC is calculated by applying Discrete Cosine Transform (DCT) to the transformed Mel-frequency coefficient to obtain a set of cepstral coefficients [58]:

$$c(n) = \sum_{m=0}^{M-1} \log_{10}(s(m)) \cos(\frac{\pi n(m-0,5)}{M}), \qquad n = 0, 1, 2, ..., C-1, \tag{3.43}$$

where $c(n)$ is the cepstral coefficients, C is the number of MFCCs.

MFCC systems mainly use only 8 - 13 cepstral coefficients. The zero coefficient is the average logarithmic energy of the input signal, which carries only a small amount of specific information, therefore, it is excluded [61].

3.2.2 Convolutional Neural Network Architecture

A CNN is a neural network that has a convolutional layer. Usually, convolutional neural networks also have a pooling layer and a fully connected layer.CNNs are used for optical pattern recognition [44], image classification [65], object detection [25], semantic segmentation [48] and other tasks [29].

The foundations of the modern architecture of CNN were presented in one of the first widely known CNN, namely LeNet-5 by Yann LeCun [44], the architecture of which is presented in Figure 3.7.

In CNN, convolution and downsampling layers comprise several "layers" of neurons called feature maps or channels. Each neuron of such a layer is connected to a small section of the previous layer called the receptive field. In the case of an image, a feature map is a 2D array of neurons, or simply a matrix. Other dimensions can be used if another type of data is accepted as the input, such as audio data (one-dimensional array) or volumetric data (3D array) [73].

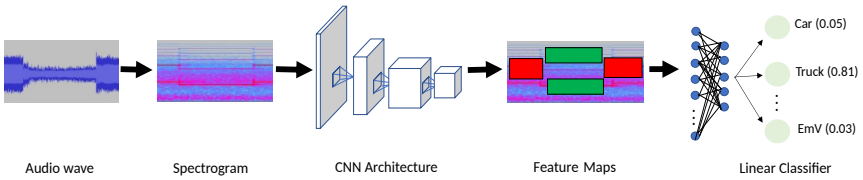

Figure 3.7: Audio classification process.

Convolutional layer

The convolutional layer is based on a discrete convolution. The operation is performed by shifting one function (the convolution kernel) relative to another, thus multiplying the samples and summing the products. The convolution core is a matrix of width kW and height kH. Most often, the matrix is square.

As a result of the convolution operation, an output feature map is formed. The source image is also a matrix (map) that is processed by the

convolution kernel. The original matrix is viewed through the convolution kernel. Furthermore, the elements of the original matrix are multiplied element by element by the corresponding values in the convolution core. The resulting products are summed, and the obtained values are transmitted to the next layer in the form of a feature map.

A 2D convolutional layer aims to extract numerous varied features from the input. Such a layer's input is a tensor with shape (height H × width W × channels C). A kernel is a tensor with shape (height h × width w × channels C), which implies that it has as many channels as the input. The input of the layer is convolved by N kernels with a uniform stride (vertical s_v, horizontal s_h) and padding (vertical p_v, horizontal p_h), and is then added to N corresponding bias. Afterwards, the outcome is applied to an activation function to generate the layer output, which is called feature maps, with the following shape:

$$([\frac{H + 2p_v - h}{s_v} + 1] \times [\frac{W + 2p_h - w}{s_h} + 1] \times N). \tag{3.44}$$

In general, the vertical stride s_v is set identically to the horizontal one s_h . Moreover, only two methods of padding are recommended. The first one, called "valid," sets p_v and p_h to zero, such that the output has a smaller height and width than the input. The other method, called "same", adds in zero elements to the edges to enlarge the input height and width after computing the proper p_v and p_h, thus, that the output has the same height and width as the input.

A 2D convolutional layer with N kernels has a total of $(h \times w \times C + 1) \times N$ trainable parameters, including weights and biases. By contrast, the height h and width w of the kernel and the stride (s_v, s_h) are chosen as hyperparameters before each training.

Mathematically, the discrete convolution operation is presented as follows:

$$(f * g)[m, n] = \sum_{k,l} f[m - k, n - l] * g[k, l], \tag{3.45}$$

where f is the original matrix of the image, g is the convolution kernel.

The convolution core is shifted by a certain step (stride) vertically and horizontally, thus obtaining a matrix as the output. The stride may be the same for the x and y axis offset or it may differ.

The size of the resulting map decreases when the convolution core is shifted relative to the original image. The dimension of the output feature map is determined by the following formula:

$$n_{out} = [\frac{n_{in} + 2p - k}{s}] + 1, \tag{3.46}$$

where n_{in} is the number of input features, n_{out} is the number of output features, k is the convolution kernel size, p - the convolution padding size, and s is the convolution stride size.

Pooling layer

The output of a convolutional layer, namely feature maps, is fed into a pooling layer to drop away potential computation.

According to the receptive field size, pooling layers are divided into the local and global layers. A local pooling layer affects a small area every time, such as 2×2 on each channel and shifts until all the elements have been drawn in. Moreover, padding should choose "valid" and stride should choose s_v as long as s_h to prevent the receptive field overlapping. Hence, the output has a smaller height and width, while the number of channels remains constant. By contrast, a global pooling layer acts on each channel merely once. There is no need to care about stride and padding. In other words, an input tensor with shape $(height \times width \times channelsN)$ ends up with the shape $(1 \times 1 \times N)$.

In terms of mathematics, two types of pooling layers exist, namely Max and Average. The former acquires the maximum value of the receptive field, whereas the latter obtains the average value. Figure 3.8 illustrates how a pooling layer works when faced with an input with a single channel.

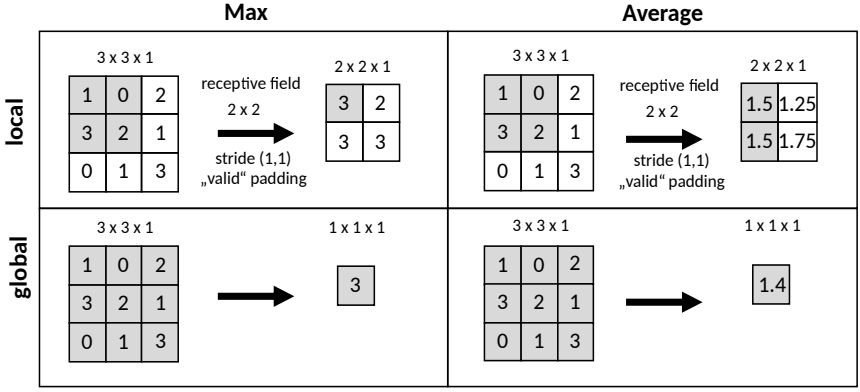

Figure 3.8: Pooling principle.

Fully connected layer

Fully connected layers act as a classifier of the whole network, since they map the distributed feature representation learned in the convolutional layers to the sample label space.

Each fully connected layer consists of dozens of neurons. Each neuron has multiple inputs and a single output. Every neuron's output $y_j^{[l-1]}$ of the former layer $[l-1]$ with p neurons connects to every neuron's input of the current layer $[l]$ with q neurons. The inputs of a neuron are multiplied by corresponding weights $w_{i,j}^{[l]}$ and are summed together with a bias $b_i^{[l]}$. Fed into an activation function $f(\cdot)$, the result forms a neuron's output $y_i^{[l]}$. Noted that for $l = 0$ there is $y^{[0]} = x$, this process can be interpreted as follows:

A single neuron:

$$z_i^{[l]} = \sum_{j=1}^{p} w_{i,j}^{[l]} y_j^{[l-1]} + b_i^{[l]} = (w_i^{[l]})^T y^{[l-1]} + b_i^{[l]}, \tag{3.47}$$

$$y_i^{[l]} = f(z_i^{[l]}). \tag{3.48}$$

Whole layers:

$$z^{[l]} = \begin{bmatrix} z_1^{[l]} \\ \vdots \\ z_q^{[l]} \end{bmatrix} = W^{[l]} y^{[l-1]} + b^{[l]}, \tag{3.49}$$

$$y^{[l]} = \begin{bmatrix} y_1^{[l]} \\ \vdots \\ y_q^{[l]} \end{bmatrix} = \begin{bmatrix} f(z_1^{[l]}) \\ \vdots \\ f(z_q^{[l]}) \end{bmatrix} = f(z^{[l]}). \tag{3.50}$$

Because a fully connected layer requires a vector input, the last convolutional layer's feature maps must pass a global pooling layer or be flattened in order to change into a vector.

As Figure 3.9 indicates, scores of fully connected layers form a neural network, known as the Multi-Layer Perceptron (MLP), which is the most widely used structure in deep leaning. The first and last layers of an MLP are called the input layer and the output layer, respectively, while the internal layers are called hidden layers. In basic CNN architecture, output of fully connected layer is the result of classification.

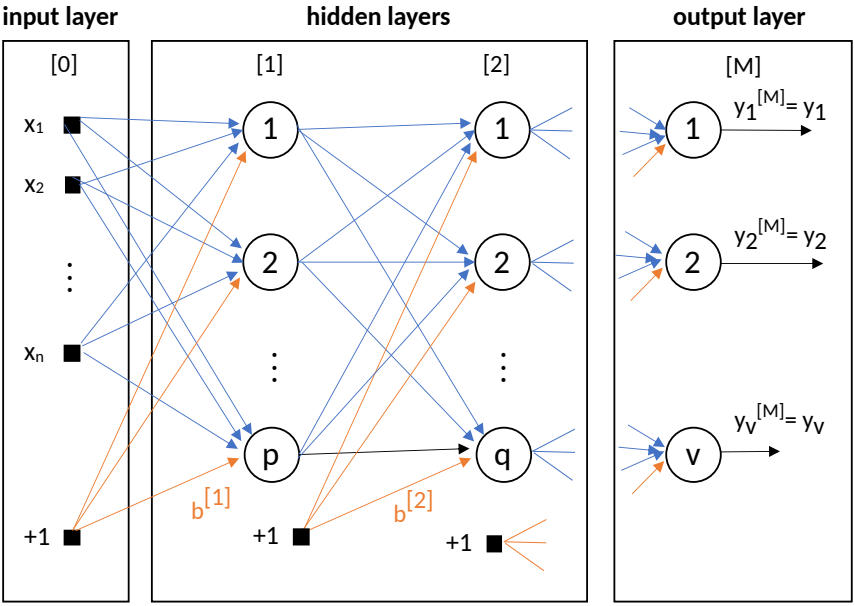

Figure 3.9: Fully connected layer principle.

4 Developed Sound Processing System for Autonomous Driving

In this chapter, we present a developed system for processing audio signals that could be valuable for an autonomous vehicle's perception of the environment. This system is able to analyze the sounds of the vehicle's surroundings to determine the location as well as the type and motion of various traffic objects. The data obtained from ambient sounds could be valuable for the decision making and path planning of an autonomous vehicle, which increases the safety and reliability of the AD system, especially in mixed traffic conditions [19]. Conventionally, the system can be divided into four parts: sensors, the preprocessing module, the classification module, and the localization module, each of which is described in detail in this chapter.

Microphones act as sensors for the sound processing system. A set of several microphones is combined into a microphone array, which can have different shapes and different numbers of microphones.

After preprocessing, the received audio signals are sent to the classification module, which is necessary to determine the significance and influence of this information on the autonomous vehicle. In this case, priority is given to the loudest sounds because, based on the basic rules of acoustics, the louder the sound, the closer it is and thus the greater its direct impact on the vehicle.

Once the audio signal is classified and its importance is understood, it is necessary to determine its location. The localization module in this system is divided into two parts, namely the localization of the sirens of EmVs and the localization of other sounds. Since the accuracy of the equipment does not allow for determining the distance to the sound source by the triangulation method, the DOA of the sound signal is determined.

For the sirens of EmVs, determining the distance is possible using the amplitude-based method, since the siren is a standardized sound signal;

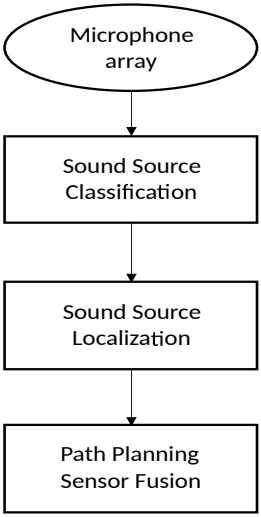

Figure 4.1: System structure.

therefore, all of its initial values are known. If the sound is known to be a damped vibration, having data on the amplitude of the received signal enables one to calculate the distance to the source by comparing it with the initial signal amplitude.

After receiving data on the type of object, its action, and its location, this information is transmitted to the path planning or fusion module, which were not a part of the present work.

4.1 Microphone Array Configuration

A microphone array is a system consisting of several acoustic sensors (usually microphones) in certain arrangement, which are used to select and process an audio signal. By collecting sound signals in different directions and applying various algorithms, the microphone array can perform many functions, such as localization of the sound source, reverberation cancelation, and speech recognition.

Several of the most widely used microphone array arrangements are presented in Figure 4.2. In a linear array, microphones are arranged in a row, which is the simplest form of microphone array. Although it can be used for sound source localization, it has limitations in possible DOA

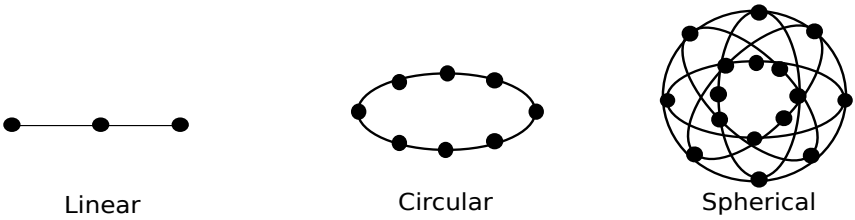

Figure 4.2: Microphone array arrangements.

estimation. Due to its linear form, detectable sound sources can be located only in front or behind the plane (which depends on the type of microphone). Furthermore, the localization accuracy of a linear array depends on the direction corresponding to the axis of the array; specifically, decreasing the angle decreases the accuracy. Linear arrays are applicable to the tasks where a suspected sound source is located in a specific narrow field, or for tasks where the purpose of direction estimation is not to determine the exact location of the sound source but rather to determine a general relative location, such as in front or behind.

A circular plane microphone array eliminates some disadvantages of the linear arrangement, such as the working range of directions. The location of sound sources has no direction requirements due to the omnidirectivity of this arrangement. Nevertheless, the plane shape of the array does not allow elevation estimation.

Finally, the spherical microphone array arrangement provides a solution to all of the abovementioned limitations. The spherical omnidirectional form has wide localization facilities, including direction estimation and elevation. However, such a configuration increases the complexity of algorithms for data processing compared with linear and circular arrangements.

4.2 Sound-Based Object and Action Classification

4.2.1 Taxonomy

The first step in environmental sound classification is the taxonomical categorization of different environmental sounds, which has been extensively researched in the context of perceptual soundscape studies [67]. However, not many investigations have been conducted into the soundscape applicable

to self-driving cars, although a few studies have investigated the sounds of the urban environment.

The UrbanSound study presented a dataset and a broad taxonomy of various urban sounds. The following requirements for the taxonomic categorization of urban environmental sounds were also proposed [66]:

- It should factor in previous research and proposed taxonomies;
- It should aim to be as detailed as possible, going down to low-level sound sources such as a car horn (as opposed to "transportation") and a jackhammer (as opposed to construction);
- It should, in its first iteration, focus on sounds that are of specific relevance to urban sound research, such as sounds that contribute to urban noise pollution.

Based on these requirements and the taxonomy presented in UrbanSound, we propose a taxonomic categorization for traffic-related sounds that can affect the driving of an autonomous vehicle on inner-city public roads. We suggest that all sounds related to the transport landscape are divided into the following three general groups:

- Humans and Animals
- Non-motorized Transportation
- Motorized Transportation.

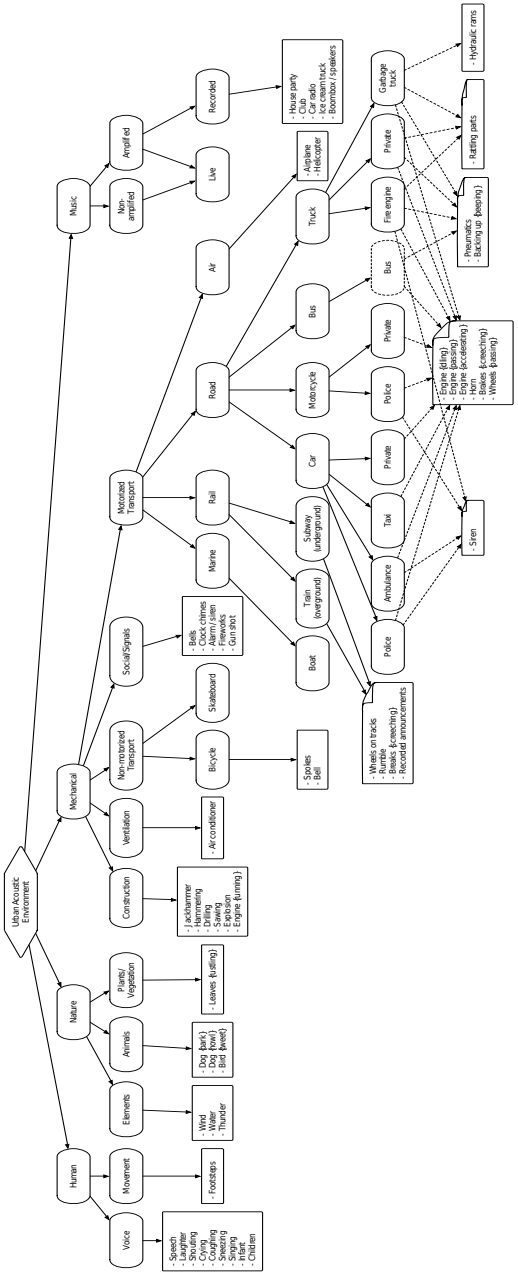

Figure 4.3: UrbanSound Taxonomy [66].

The Humans and Animals group is also divided into two classes, as in the UrbanSound study, namely voice and motion.

The Non-motorized Transportation group, is divided into three classes, namely bicycle, skateboard and wheelchair, which are the three most common non-motorized modes of urban transport.

Finally, the Motorized Transportation group is divided into two subgroups, namely rail and road.

Railway transport, in the context of urban traffic, is usually trams. Based on point 2 of the requirements for categorizing the urban environment, the classes should be as detailed as possible. Hence, a simple definition of the object type is not sufficient. In application to AD systems, it is not so much the value of the object type but rather the action performed by that object. The obtained data can be crucial for decision making and trajectory planning. That is why the type "tram" is further divided into four types of actions performed: bell, passing, acceleration, and braking. These four actions are performed the most frequently and also have the greatest impact on road traffic.

The last type is a road vehicle, which is a predominant participant in road traffic. We identified five main types of vehicles: car, motorcycle, bus, truck, and EmV.

The classes of car, truck, and bus are divided into five actions: horn, idling, passing, acceleration, and braking.

The motorcycle class is divided into four actions, namely idling, passing, acceleration, and braking, since most of the motorcycles are not equipped with a special sound horn and mainly use the ego-noise of the engine as a warning signal.

The EmV class represents the sounds of sirens and special signals installed on this type of vehicle. In the UrbanSound taxonomy, sounds of this type were categorized as social, along with the bell of a clock and the sound of a gunshot. Because such sounds are produced exclusively by different types of vehicles, they should be classified as road- or traffic-related. Traditionally, the sound of a siren is divided into three types: fire, police, and ambulance. Some vehicles are equipped with a siren but do not belong to one of these three types, which are specific to each country; therefore, this class was not included in the overall picture in the present study.

The final categorization is presented in Figure 4.4.

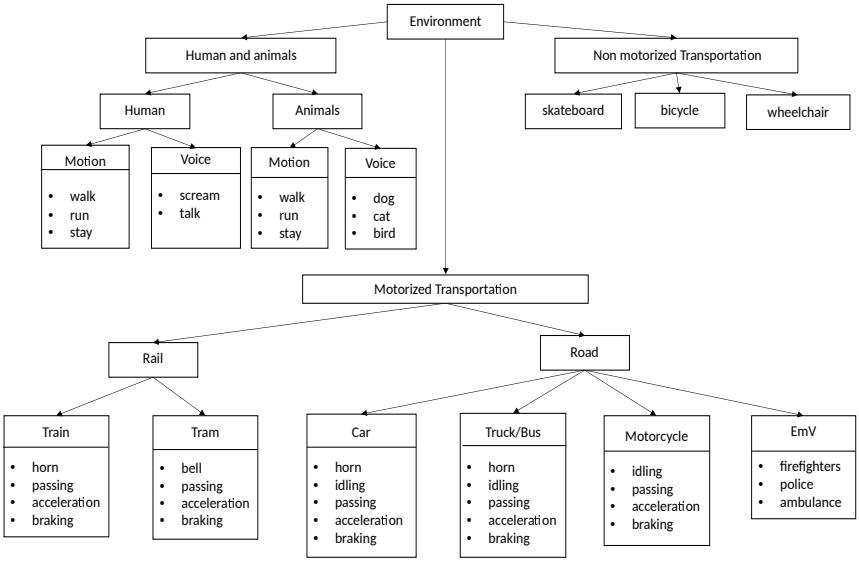

Figure 4.4: Taxonomical categorization of traffic soundscape.

4.2.2 Dataset

In this subsection, we present the dataset that we created to classify objects and their actions.

Acquisition

The transport sound recordings for the dataset were obtained using two methods.

First, various videos that captured events of interest were selected to formation of the dataset. The video files were found mainly on YouTube. Data for classes related to motorbikes were collected, in this way along with parts of the training sets for the emergency braking of cars and trucks, tram bells, and car and truck horns. This method is suitable for sound classes of events that are rare and difficult to reproduce under experimental conditions. Videos of these events were downloaded from YouTube and the audio tracks were extracted in .wav format. The data was then further filtered and manually labeled.

The second method involved recording the sounds of the urban environment. The recording was performed in Darmstadt, Germany. Several of

the most congested intersections in the city with traffic lights were selected as locations for the static recording of braking and acceleration sounds. A recorder was attached to a traffic light pole and all events were manually fixed. Dynamic sound recording was performed using microphones mounted on the roof of the experimental car in parallel with camera recording. When planning the recording route, the task was to plan the route such that it passed through the busiest intersections in the city, as well as through junctions where car traffic intersects with tram lines. Furthermore, the time between 17:00 and 19:00 was chosen for the recording sessions. This time interval is evening rush hour, so the street and road network of the city is maximally loaded and saturated with various types of transport objects.

The map in Figure 4.5 presents the route of the predominant transit traffic in red, the tram tracks in yellow, and the route of the test vehicle at the time of the recording in blue, green circles indicate the intersections of static recording.

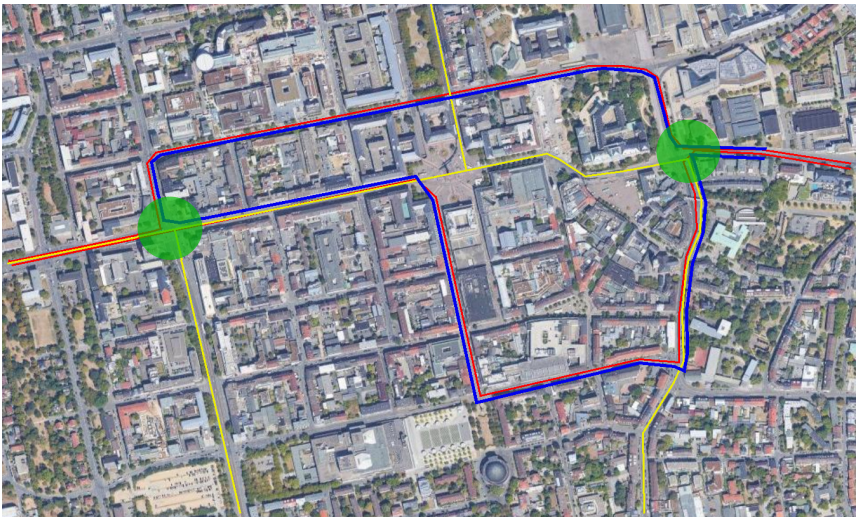

Figure 4.5: Route of experimental vehicle.

Special mention should be made of the recording of sirens. To record sirens, we purchased a siren generator with the capacity to reproduce seven types of sirens. The characteristics of the emitter were close to those of real EmV warning equipment.

Motion Classification Features

In application to AD perception systems, detecting the action of different objects plays a major role. To plan the trajectory and apply the optimal interaction strategy, it is necessary to have as much information about other road users as possible. Regarding the application of sound processing to autonomous vehicles, it is worth noting that sound contains much information that can be useful in the context of road situation assessment. The analysis of the recorded sounds of urban traffic revealed certain features that indicate actions taken by other traffic participants.

Wheeled vehicles with internal combustion engines, cars, motorbikes, and trucks emit a certain type of sound when accelerating, which changes in frequency depending on the rpm of the engine. When accelerating sharply from a standstill, the tires produce a high-frequency sound when the wheels spin, which is also a feature of the vehicle's acceleration.

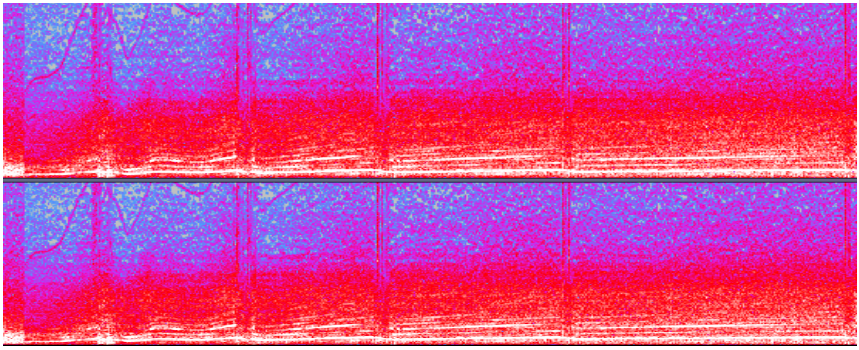

Figure 4.6: Acceleration features.

During emergency braking of cars and trucks, tires sliding on the road surface also produce a high-frequency sound. The sudden braking of a road user provokes sharper braking from every vehicle that follows, so a chain reaction occurs. This sign is therefore critical, especially if emergency braking is performed by the driver in front. Furthermore, as the speed of a car or truck gradually reduces, a particular engine sound is produced. Road traffic regulations recommend braking using the engine in conjunction with the hydraulic braking system. In some cases, a harsh sound is produced when the brake pads interact with the brake plate, which is rare and mostly indicative of a faulty brake system; nevertheless, this sign was also considered in this study. In the case of trucks, specific noises are also

associated with the pneumatic braking system installed on most trucks.

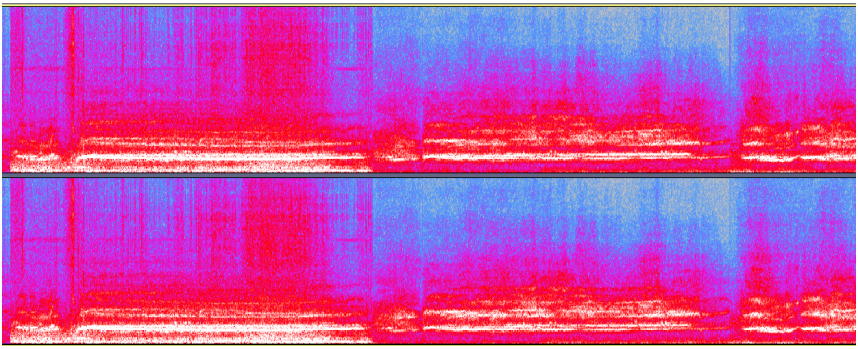

Figure 4.7: Braking features.

The idling class for cars, trucks, and motorbikes is based on the sound made by the engine when running at idling speed. If a car, truck, or motorbike increases engine speed while standing still, this event should be classified as acceleration as it indicates a possible sharp start. When applied to trajectory planning and interaction, this feature should be considered "possible acceleration."

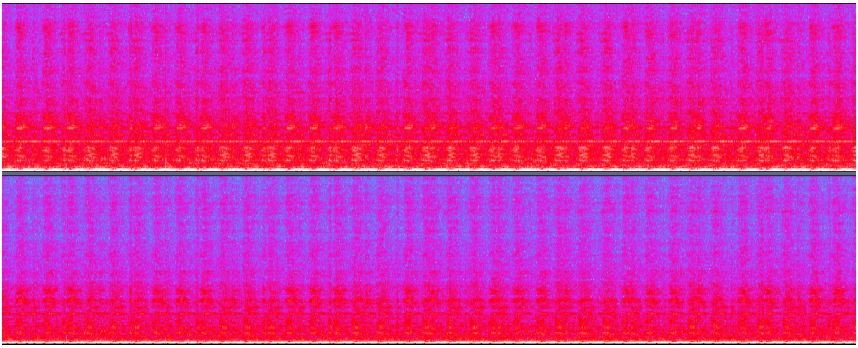

Figure 4.8: Idling features.

Notably, the features applicable to trams are not similar to those of wheeled vehicles. The tram is an electric rail vehicle, the acceleration of which is associated with a particular acceleration sound from the electric motor, which audibly represents an increase in frequency over time. When

a tram brakes, the sound of the electric motor is the opposite, with the frequency decreasing over time. Furthermore, the tram uses a braking system that presses the pad against the rail, which creates a particular sound. Trams are not characterized by an idling sound due to the absence of an internal combustion engine, but they do have a characteristic sound produced in the interaction of the wheelsets with the rails when it passes.

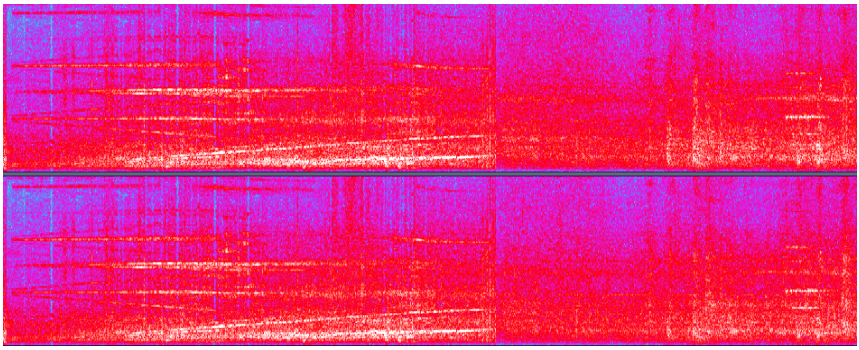

Figure 4.9: Tram features.

EmV sirens are the most obvious and specific sound made by vehicles in urban environments. The characteristics of sirens can vary from country to country. In Germany, a DIN standard has been adopted for vehicle warning systems. The standard requires that the sirens consist of two continuous sounds of differing frequencies. The sounds emitted by vehicles of different services also differ from each other. Fire service vehicles use a pneumatic siren generator, whereas police and ambulance vehicles use an electronic siren generator and emit sounds of different frequencies. In our research, we used a siren generator that emitted different types of siren sounds similar to the real. This type of sound is subject to retraining depending on local regulations and standards applied to sirens of EmVs.

Dataset Structure

In our research, a dataset was recorded that consisted of 17 low-level classes corresponding to the Motorized Transport group. According to the taxonomy presented above, 14 classes represent objects and actions as well as three classes for different types of EmVs.

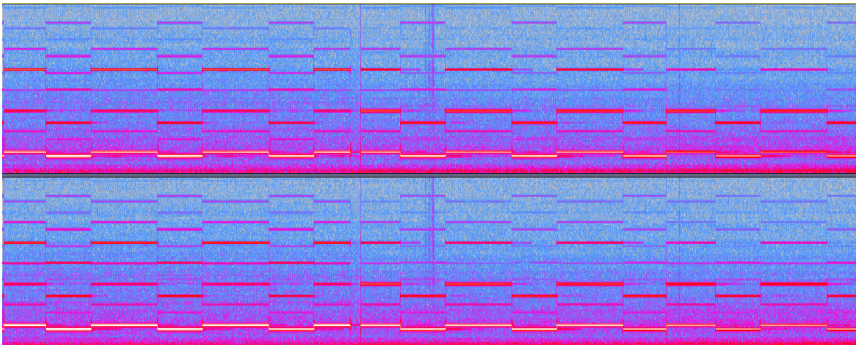

Figure 4.10: Siren features.

Car	Truck	Motorcycle	Tram	EmV
horn	horn	idling	bell	Fireworks
idling	idling	accelerating	passing	Police
braking	braking		braking	Ambulance
accelerating	accelerating		accelerating	

Table 4.1: Dataset structure.

We combined the bus and truck types into one type due to the small difference in the sounds they produce as well as the similarity of their behavior patterns.

An analysis of the sound recordings revealed no features of motorbike braking. Engine braking is rarely used in the case of smooth speed reduction. In the case of hard braking, a sound similar to the hard braking of a car is produced, but this is an extreme case that often results in the motorbike tipping over as well as further uncontrollable sliding. No sound data could be found that offered any indication of motorbike braking. Since it was not possible to reproduce sounds for this class, the definition of motorbike braking was excluded from this dataset.

The event "Passing" for cars, motorbikes, and trucks was also included in the "Acceleration" class as this sound is barely distinguishable when driving. In the case of a stationary autonomous vehicle, considering the difference in speed, the same strategies should be applied to a passing object as to accelerating vehicles.

Convolutional neural network training

For the classification of objects and actions, a convolutional neural network was used. Its principle of operation was to classify MFCC frames. For feature extraction, the Librosa library was used. The audio files were divided into separate fragments and spectral features were extracted from each fragment. The features were combined into a Numpy array, thus forming a matrix. The normalization library is trained by calculating normalization data on the training array using StandardScaler function based on the parameters of normal distribution. Next, a normalization operation is performed on each matrix so that the data has approximately the same amplitude and all the features are analyzed by the neural network according to their importance in training. The matrix is then fed to the input of the neural network.

The neural network consisted of one convolutional layer with 4 filters that performed input matrix convolution, and then a zero padding layer was used to add zeros to the resulting matrix as well as to help preserve the information after the poolling operation.

Next, we used a poolling layer, which reduced the dimensionality and highlighted the main features. Then, we had a flatten layer, which adjusted the dimensionality of the output matrix, followed by a fully connected layer of 200 neurons and a classifier of fully connected neurons. The number of classes was 17, which corresponded to the number of neurons in the output layer.

L1, L2, dropout, and batch normalization were used to reduce the effect of overfitting.

The dataset consisted of 400 fragments of 3 seconds in each of the 17 classes. The division into training and validation samples was performed at a ratio of 75% training and 25% validation (i.e., 300 and 100 fragments, respectively).

Number	Class
0	Car acceleration
1	Car braking
2	Car horn
3	Car idling
4	Motorcycle acceleration
5	Motorcycle idling
6	Siren 1
7	Siren 2
8	Siren 3
9	Tram passing
10	Tram acceleration
11	Tram braking
12	Tram bell
13	Truck acceleration
14	Truck braking
15	Truck horn
16	Truck idling

Table 4.2: Classification classes.

Table 4.3 presents the results of object classification independently from motion class. This includes all the true predictions and false prediction within the object type.

Class	Accuracy
Car	83.75 %
Motorcycle	91 %
Siren	99.3 %
Tram	88.25 %
Truck	89.25 %

Table 4.3: Object classification accuracy.

Results of motion classification independently from object type are presented in Table 4.4. Accuracy was calculated by summing all the true predictions with false predictions within the same motion class.

Class	Accuracy
Acceleration	84 %
Braking	77 %
Horn / Bell	86.66 %
Idling	91.25 %

Table 4.4: Motion classification accuracy.

Figure 4.11 illustrates the results of neural network performance on the validation dataset. Overall, the classification accuracy is 84.7%. Best classification results 99-100% are reached by the EmV siren sound sources, because they are very loud and obvious for classification algorithms. "Truck acceleration" class has the lowest accuracy, and 12 of false predictions as "Truck braking". This is caused by the sound of pneumatic system which is involved in both actions.

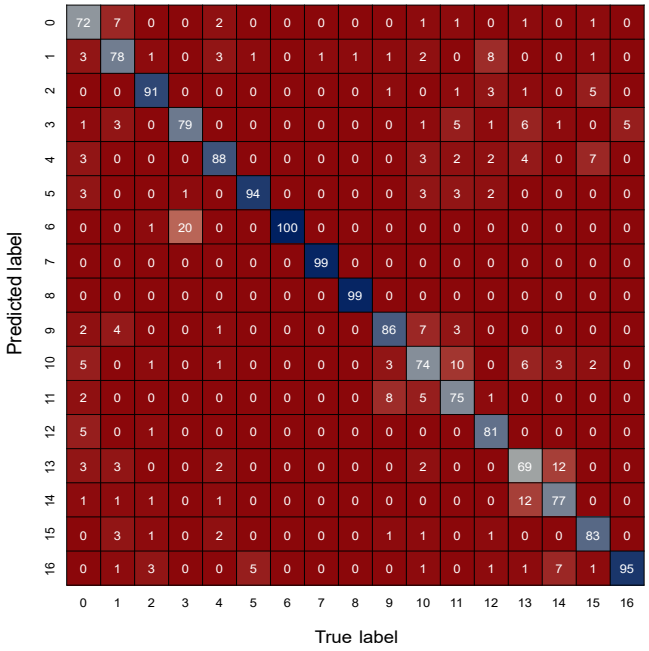

Figure 4.11: Error matrix of classification.

4.3 Sound-based Object Localization

4.3.1 SRP-PHAT Direction of Arrival Estimation Method with Preliminary Sector of Arrival Selection for Circular Microphone Array

Steered response power (SRP) is a method of sound source localization that uses beamforming. Unlike TDOA, it is a one-step process that uses the GCC of multiple microphones instead of pairwise directional estimation. The operating principle is identical to scanning. Beamforming searches for the source in space; if the direction of focus is the same as the actual direction to the sound source, then SRP creates a peak. The idea behind this approach is the assumption that more energy is emitted in the direction of the sound source than in other directions. In application to circular microphone arrays, the search is performed over the entire FoV, which makes the process more complex and time-consuming.

To reduce the time spent on peak search, we propose a two-step localization. Using GCC algorithms on several pairs of microphones, a sector of arrival search is performed. For this purpose, several layouts of circular microphone array pairs are proposed. Once the sound arrival sector is determined, the SRP algorithm applies only to the selected sector, determining the exact DOA of the sound source.

GCC-PHAT for a pair of microphones

Consider a coordinate system with an origin point in the middle between the pair of microphones. The relationship can be presented as follows:

$$c\tau = \sqrt{(x - \frac{d}{2})^2 + y^2} - \sqrt{(x + \frac{d}{2})^2 + y^2}, \tag{4.1}$$

where d is the distance between microphones, (x, y) is the coordinate of sound source.

A standard hyperbolic formula can be obtained through converting Equation (4.1) as:

$$\frac{4x^2}{\tau^2 c^2} - \frac{4y^2}{d^2 - \tau^2 c^2} = 1. \tag{4.2}$$

The real sound source location will be on the right half of the hyperbola.

In application to far-field signal models, the incidence angle θ for each microphone will be the same, hence, the distance difference and the time delay are interrelated as follows:

$$d\cos(\theta) = c\tau. \tag{4.3}$$

Consequently, the sound source angle of incidence is defined as follows:

$$\theta = \arccos(\frac{c\tau}{d}) \tag{4.4}$$

The processing of one pair of microphones is insufficient for determining the DOA because the angle of incidence would have two directions. Thus, to determine the exact DOA, it is necessary to have a microphone array that consists of at least three microphones, which will enable one to calculate two or more incidence angles by combining several pairs of microphones. Based on the data from two or more pairs, the clustering method can then be applied to determine the DOA.

Pair selection

In this study, we used a circular eight-microphone array for sound processing. It is possible to use several layouts of microphone pairs to determine the sector of arrival. The directional accuracy depends on the distance between the microphones and the number of pairs. The greater the number of pairs and the distance between the microphone pairs, the higher the accuracy.

Because the first step of direction estimation is sector searching, the processing of the first stage does not have to be very precise, but it should be fast and therefore low in cost in terms of computational power. Three pairing strategies are proposed for this purpose.

The first strategy is called the "pair with following" strategy. The pairing is formed with each successive microphone in a clockwise direction. Thus, eight pairs of microphones are formed, but the distance between them is quite small.

The second strategy is called the "pair with opposite". Each microphone is paired with the opposite microphone in the circular array. Thus, only four pairs of microphones are formed, but with the largest possible inter-microphone distance, which is equal to the diameter of the microphone array.

The third strategy is called the "two-square" strategy. Microphone pairs are formed in a pattern similar to a square microphone array. The circular

eight-microphone array is conventionally divided into two squares of four microphones each. One square consists of microphones with even numbers, while the other consists of odd-numbered microphones. This results in two square grids offset by 45° with respect to each other. Thus, there are eight pairs of microphones. This method is a compromise. Compared with the "pair with following" strategy, the distance between the microphones increases but the ability to use eight pairs is maintained. Compared with the "pair with following" strategy, the inter-microphone distance is smaller but there are eight pairs of microphones instead of four.

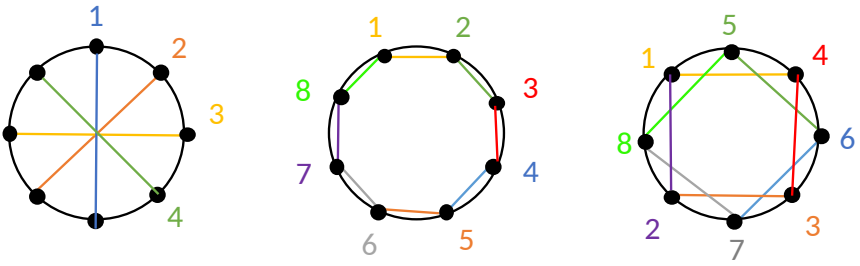

Figure 4.12: Pairs Selection.

Sector of arrival estimation

The arrival sector is determined by applying the GCC-PHAT algorithm to all selected microphone pairs. This generates eight parabolas, which are the decision area for each pair individually. Using a two-square scheme, the pairs are distributed as presented in Table 4.5:

1	2	3	4	5	6	7	8
1-3	3-5	5-7	7-1	2-4	4-6	6-8	8-2

Table 4.5: Microphone pairs distribution.

After linearization and simplification of the curve, a linear curve is obtained, as depicted in blue in Figure 4.13. Since the distance from the sound source to the microphone array is large enough, the far-field assumption is satisfied and thus the linearized curve approximates the nonlinear curve, as depicted in red in Figure 4.13. Applying GCC-PHAT

to a pair of microphones obtains two values, since based on the time delay the two microphones can determine the direction relative to the axis in the positive and negative direction. As the figure demonstrates, the solution has two adjacent angles.

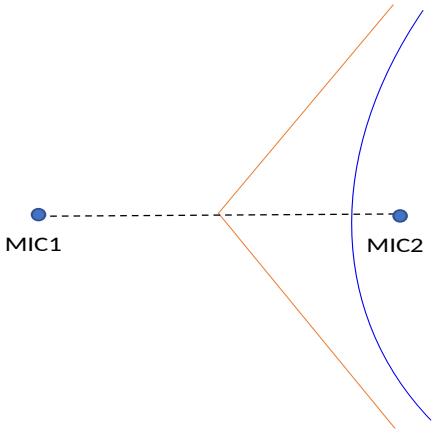

Figure 4.13: Solution curve.

For each pair of microphones, the sound source angle relative to the midpoint of that pair is calculated and converted into a direction relative to the microphone array. The incident angle of the microphone pair is θ, then, the arrival angle will be the sum of θ and the angle of the microphone pair's relationship to the plane of the microphone array.

Applying the GCC to each selected pair produces 16 directional angle values, eight of which are true while the other eight are "reflections" of the angle relative to the plane of the microphone pair. Figure 4.14 illustrates the process of calculating the angle DOA through an angle distribution.

This process results in an estimated sound DOA based on the GCC-PHAT of eight designated microphone pairs. This result cannot be considered accurate, however, as the experiments demonstrated that the standard deviation of the DOA error, at different distances from the sound source, is no more than 2.5°. Taking possible inaccuracy errors caused by reverberations and a noisy environment into account, the width of the investigated sector can be assumed to be 10°. Thus, the sector is defined as a GCC-PHAT value ±5°.

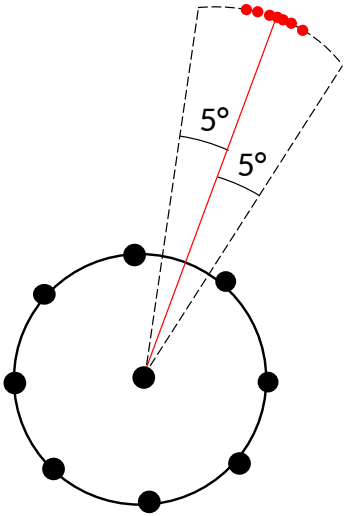

Figure 4.14: Angle distribution.

SRP-PHAT

The SRP-PHAT algorithm searches for the direction that maximizes the function and creates a peak. For circular microphone arrays, the FoV is 360°. Using preliminary search data of the sector of arrival, the determined sector is set as the FoV for SRP-PHAT, which it is offered to designate as 10°, ±5° from an arrival direction defined by GCC-PHAT. Thus, the field of possible solutions is formed and substituted into the formula.

$$P(q) = 2\pi \sum_{k=1}^{N} \sum_{l=k+1}^{N} G_{kl}(\tau_l - \tau_k), \tag{4.5}$$

This approach reduces the processing time while maintaining the advantages of the algorithm, including accuracy.

4.3.2 Amplitude-based Distance Estimation Method for Emergency Vehicle Localization

A sound signal is known to be a damped oscillation. Its amplitude decreases according to the distance traveled by the sound wave.

Thus, the amplitude A of the audio signal is related to the distance r to the audio source and is proportional to the sound pressure p. This approach to distance estimation is only applicable if it is possible to determine the sound source initial characteristics. Furthermore, in the context of to far-field applications, the sound source must have a high volume, otherwise the sound-to-noise ratio (SNR) even at short range will not be sufficient for processing.

For a loud sound source with defined sound parameters, this method is sufficiently reliable for determining the distance r. In application to EmVs, the siren sound is standardized for each country, which means that all siren characteristics are known in advance or can be determined. Moreover, parameters do not vary from vehicle to vehicle. Therefore, knowing the initial sound pressure of the siren enables the accurate determination of the distance to the siren sound source and therefore to the EmV.

The ratio of sound pressures $\frac{p}{p_0}$ at different distances r_0 and r is inversely proportional to the ratio of these distances:

$$\frac{p}{p_0} = \frac{r_0}{r}. \tag{4.6}$$

Since sound pressure is proportional to amplitude, it is possible to determine the distance r given the current amplitude A:

$$r = r_0 \frac{A_0}{A}. \tag{4.7}$$

Consequently, to determine the distance to the sound source, the amplitude of the received signal must be determined, along with the original amplitude A_0 of the signal and distance r_0.

The dataset in section 4.2.2 presented possibilities for classifying different sound sources, including sirens. In our study, three types of sirens that meet the German DIN standards for public alarm systems were used. As a siren is a loud sound source, its detection range can be quite high and its features are clear.

To determine amplitude A from the frame wave form, the sum of the 100 highest peak values and the sum of the 100 lowest peak values are calculated and divided by 200 to obtain the average amplitude of the sound samples with a high signal-to-noise ratio. The frame is stretched in time for 3 seconds to enable full coverage of the most common siren period.

$$\begin{aligned} \overrightarrow{S} &= [S_1, S_2 ... S_T] \\ \hat{S} &= sort(\overrightarrow{S}) \\ A &= \frac{1}{200}(\sum_{i=T-100}^{T} \hat{S}_i - \sum_{i=1}^{100} \hat{S}_i)] \end{aligned} \tag{4.8}$$

As demonstrated in the formula, to calculate the distance for a siren, the initial distance and amplitude r_0 and A_0 must be determined. These parameters can be calculated for each siren individually. For each classified type of siren, the specific constant c_j is set, where $j = [1...n]$ depending on the type of siren. The distance to the sound source is calculated as follows:

$$r = \frac{c_j}{A}. \tag{4.9}$$

Similar approach is applicable to SIL and SPL based distance estimation. Sound level L is determined by the sum of the 100 highest peak values and the sum of the 100 lowest peak values and divided by 200 to obtain the average level value in dB.

$$\begin{aligned} \overrightarrow{S} &= [S_1, S_2 ... S_T] \\ \hat{S} &= sort(\overrightarrow{S}) \\ L &= \frac{1}{200}(\sum_{i=T-100}^{T} \hat{S}_i - \sum_{i=1}^{100} \hat{S}_i)] \end{aligned} \tag{4.10}$$

First, the initial value of sound level L_0 on distance r_0 is determined. Then, distance estimation to a siren sound source is performed as follows:

$$r = r_j \cdot 10^{(\frac{|L_j - L|}{20})}, \tag{4.11}$$

where $j = [1...n]$ depending on the type of siren.

This approach makes it possible to estimate the distance to the sound source with predefined initial characteristics.

4.4 System Structure

The overall structure of the system implies the ability to acquire the full range of audio data of interest for the acoustic analysis of the environment of an autonomous vehicle.

A circular microphone array is used to record the audio environment of the autonomous vehicle. Recording is performed by eight microphones mounted outside on the roof of the vehicle. This configuration is optimal for evaluating the environment of the autonomous vehicle. Furthermore, this design can be used to accommodate other sensors necessary for the functioning of the AD system.

The preprocessing module preprocesses the received data, including filtering and sound separation. Depending on the environment of use, this module is configured according to the tasks to be performed.

After preprocessing, the data is fed to the classification module, which performs the classification of the audio signals. In Section 4.2, a categorization was proposed for the classification of traffic sounds along with a dataset for the definition of some objects and actions. Based on the proposed approach, the signals are classified into 14 types of objects and motions and three types of EmV sirens.

Depending on the classification results, the data is allocated to two localization modules. The first is the sound source arrival direction detection module, which applies the DOA estimation algorithms described above. The second is the localization module, which only receives signals that are classified as one of the siren types of EmVs, as an acoustic distance detection approach has been proposed for this class. Consequently, siren-equipped vehicles can be located based on purely audio data.

Any data obtained that is valuable for environmental perception can then be transferred to a path planning module, decision making module, or perception module, depending on the overall configuration of the AD system. These systems were not a part of this research.

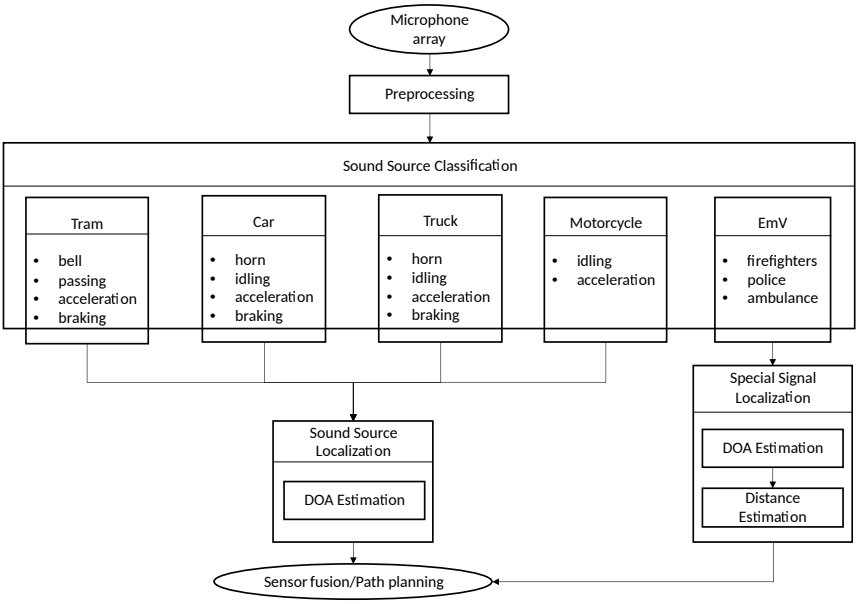

Figure 4.15: Structure of sound processing system for autonomous driving.

4.5 Experiments and Evaluation

4.5.1 Experimental Setup

Microphones

The aim was not to accurately reproduce the audio signal, but rather to reliably recognize patterns of different audio signals or events associated with them. If the accurate reproduction of a silent audio signal was paramount, then the use of a capacitive microphone would obviously be necessary. In this case, however, the properties of a dynamic microphone, such as its reliability and insensitivity to moisture, were more important. Therefore, we decided to use a dynamic microphone.

Dynamic microphones can be divided into two types, namely conventional moving coil microphones and the rarer ribbon microphones. Moving coil microphones are almost always what is meant when dynamic microphones are mentioned in general. In the context of the present work, however, the term dynamic microphone also refers to moving coil microphones.

Dynamic microphones convert sound pressure $\Delta p(t)$ into a voltage signal $u_s(t)$ according to the electrodynamic principle. As the name suggests, the change occurs with a moving coil. The basic design is identical to a loudspeaker. In principle, dynamic microphones are inverted speakers. A moving coil attached to a diaphragm, also known as a voice coil, is mostly immersed in a magnetic gap. Sound waves hitting the membrane cause it and hence the plunger coil to vibrate.

This alternating motion changes the magnetic flux Φ flowing through the plunger coil. The voltage induced in the coil is calculated using the following equation:

$$u_{ind} = -N\frac{d\Phi}{dt}, \tag{4.12}$$

where N is the number of turns in the coil. A dynamic microphone is a passive sensor that does not require an external power source. The greatest advantage of dynamic microphones is that they only comprise a few discrete parts, making them highly durable. In addition, compared with condenser microphones, they are completely insensitive to moisture, making them suitable for outdoor use [26].

Determination of the DOA is only possible with a more complicated analysis of transit time differences and the use of several microphones. This is easier for microphones with directional characteristics, but it requires a larger number of microphones to cover the entire FoV. However, since up to eight microphones were available, this condition was met. The cardioid directional characteristic of the microphone offered the clear advantage of signals almost only from the front half-space being detected. Thus, it was possible to determine the DOA.

Shure SM58 microphones were selected for the experiments. They have an output and transmission impedance of 300 ohms and an open-circuit sensitivity S 0 of 1.85 mV/Pa in the range of 50 Hz to 15,000 Hz. The advantage of the Shure SM58 is its integrated balloon filter, which additionally dampens wind and clatter. Due to the airflow, interfering wind noise is a fundamental problem in a planned acoustic environmental analysis. In addition, the Shure SM58 features a damping system to reduce noise from vibrations, which is unavoidable when a microphone is mounted on a moving vehicle.

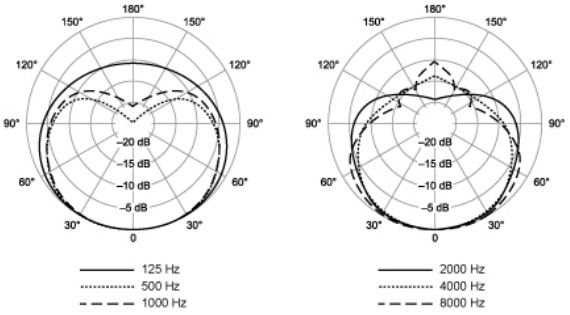

Figure 4.16: Directional characteristics of Shure SM58 microphone.

Experimental Vehicle

For the experimental evaluation of the presented algorithms, a circular eight-microphone array with diameter of 1 m was built. To attach the array on a car roof, a wooden frame with microphone mounts was constructed. The experimental setup consisted of the following:

- 8 × SM58 Shure microphones with foam and fur windscreen
- 8 × 5-meter XLR cables
- U-Phoria UMC1820 USB audio interface with sampling frequency up to 96 kHz
- Roof rack frame
- 12V DC / 220V AC converter

To generate siren sounds, a special loudspeaker was used that could produce seven different types of siren sounds, all of which had similar characteristics to real EmV sirens.

All of the equipment was installed on the experimental vehicle, which was a 2016 model Volkswagen Polo.

4.5.2 Sound Source Classification Evaluation

For the classification experiment, a test drive around the city was performed. The route was planned such that it passed the busiest parts of

Figure 4.17: Experimental vehicle.

the city. Furthermore, if a rare object was detected, such as a motorbike, it was "followed" through several traffic light areas. Moreover, static recording of three road intersections was conducted. The recorder was placed immediately on the border of the road crossing and directed toward the center of the intersection.

Siren sound signals were recorded in a parking lot. The siren generator was static while the experimental vehicle with recording equipment was driven around in circles and figures of eight.

All of the recorded data were reviewed and the additional real-test dataset was formed.

Figure 4.18 presents the performance of the neural network on the real-rest dataset. Obviously, the optimal classification accuracy was achieved for EmV sirens as such sounds are loud, easy to recognize, and not similar to any other sound signals in the traffic acoustic soundscape.

Six classes exhibited accuracy lower then 70%, namely the Car, Tram, and Truck acceleration and braking classes. Most false classifications occurred between motions of the same object.

Furthermore, the Car acceleration class had 12 of 100 false classifications with the Tram Bell class. This was caused by the possible occurrence of

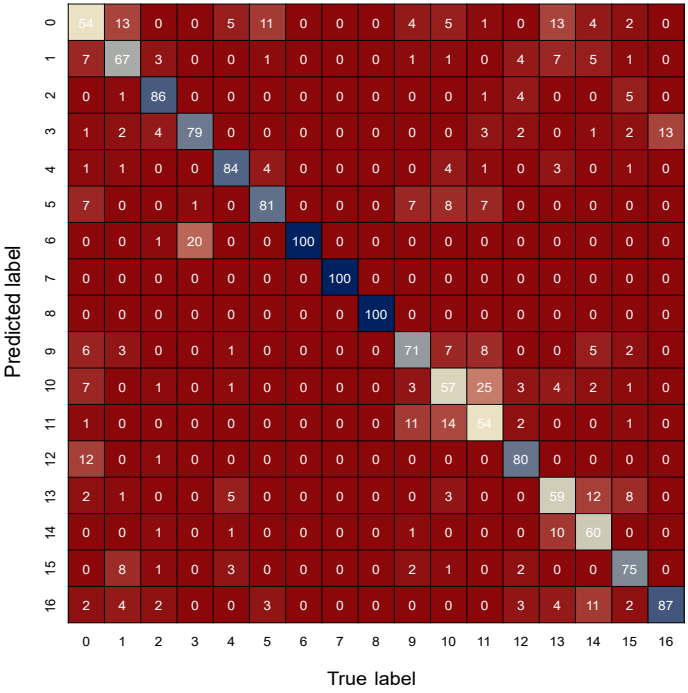

Figure 4.18: Classification error matrix.

other vehicles around the tram on the road crossing. Furthermore, the Car idling class had 20 false classifications with one of the siren sounds. This could be related to the experimental vehicle's ego-noise during siren recordings at far distances.

4.5.3 Sound Source Localization Evaluation

Direction of arrival estimation

The proof-of-concept experiment was conducted in a parking lot in the town of Rossdorf, in a suburban area of Darmstadt. A siren and car horn were recorded at distances of 10, 20, 40, 60, and 80 m and with an arrival direction of 0°, 10°, and 25°.

The following table presents the localization results using pure GCC-PHAT for the car horn and siren.

Range \ Angle	0°	5°	25°
10	0.66	2.33	2.72
20	2.63	3.42	1.2
40	1.57	2	0.44
60	1.7	1.93	0.68
80	2.34	6.6	1.52

Table 4.6: Results of GCC-PHAT angle estimation for car horn.

Range \ Angle	0°	5°	25°
10	0.32	0.82	-0.5
20	1	1.68	-0.68
40	0.44	1.53	-1.22
60	-0.2	0.71	-0.95
80	1.89	2.96	0.37

Table 4.7: Results of GCC-PHAT angle estimation for siren sound.

After performing an arrival direction analysis with the GCC-PHAT algorithm, a SRP-PHAT algorithm with arrival sector pre-selection was applied to the same dataset.

The results of the siren and car horn angle estimation are presented in the following tables. The results in bold are those that were worse after the algorithm was applied.

Range \ Angle	0°	5°	25°
10	**-1.3**	-1.6	**1.8**
20	0	2.5	0.1
40	-0.1	0.7	0.3
60	0.3	0.7	0.6
80	-0.1	0.8	0.4

Table 4.8: Results of SRP-PHAT angle estimation for car horn.

Range \ Angle	0°	5°	25°
10	**-0.9**	0.3	-0.3
20	-0.3	0.7	-0.4
40	-0.2	0.2	-1
60	**-1.3**	-0.3	**-1.2**
80	**2.9**	-0.1	**0.7**

Table 4.9: Results of SRP-PHAT angle estimation for siren sound.

The second experiment was performed in the same parking lot in Rossdorf, Germany. The siren generator was employed and three different types of siren signal were recorded. The types were selected in accordance with the DIN standard.

Signals were recorded in the ranges of 5–70 m with steps of 5 m and angles of 0°, 5°, and 15°. The following tables present the error of direction estimation for each type of siren sound in different directions.

Range \ Angle	0°	5°	15°
5	-2.6	0.1	-0.1
10	-0.4	-0.1	-0.3
15	3.1	-0.4	-0.8
20	-14.5	0.5	0.3
25	0.6	-1.5	-1.9
30	-2.6	0.6	1
35	-0.3	-0.3	-0.3
40	-0.7	-1.2	-0.3
45	-2.2	-1.2	-0.6
50	-1.1	-1.1	-0.1
55	-1.2	-0.5	-0.3
60	2.7	0.9	0.7
65	5.7	0.9	0.6
70	-4	-0.2	-0.4

Table 4.10: Results of DOA estimation for siren type 1.

Range \ Angle	0°	5°	15°
5	0.6	0.2	0.2
10	0.3	0.1	0.2
15	4.6	0	0.6
20	0.5	-0.1	0.8
25	0	-0.4	-0.5
30	0.3	0.5	0.6
35	0	0.2	-0.1
40	-1.4	0.1	-0.3
45	0.4	0.4	0.5
50	0.8	0.2	0.2
55	-0.3	0.2	0.3
60	0.1	0.6	-0.4
65	-1.1	0.8	0.6
70	-1.1	-0.2	0.5

Table 4.11: Results of DOA estimation for siren type 2.

Range \ Angle	0°	5°	15°
5	1.6	0	0.5
10	-0.9	0	-0.5
15	-1.5	-10.1	-10
20	4.7	0.2	1.3
25	0.6	-0.3	0
30	-0.5	-0.1	0.3
35	-0.3	-0.4	-0.2
40	-0.9	1.1	-0.6
45	-11	-0.3	-0.4
50	-0.5	0.2	-0.6
55	-1.4	-0.1	-0.4
60	9.5	-0.2	-0.1
65	-2.6	-0.3	-0.5
70	-0.8	-0.9	0

Table 4.12: Results of DOA estimation for siren type 3.

Distance estimation

The experiment for determining the distance to the siren sound source was conducted in a parking lot in Weiterstadt. The signals were recorded at distances between 5 and 300 m. Recordings from 5 to 30 m were conducted every 5 m, and from 50 to 300 m every 25 m. Due to the parameters of the siren sound source, which was not as powerful as real EmV equipment, the SNR was not high enough in the range of 150–175 m depending on the type of signal.

Figure 4.19: Distance estimation experiment.

Distance real	Distance estimated	Error in [m]
10	10.76	0.76
15	14.55	-0.45
20	20.82	0.82
25	26.12	1.12
30	30.83	0.83
50	48.08	-1.92
75	72.77	-2.23
100	102.79	2.79
125	129.41	4.41
150	155.23	5.23
175	180.91	5.91

Table 4.13: Results of distance estimation for siren type 1.

Distance real	Distance estimated	Error in [m]
10	10.26	0.26
15	16.56	1.56
20	19.45	-0.55
25	25.06	0.06
30	28.94	-1.06
50	49.71	-0.29
75	77.53	2.53
100	105.67	5.67
125	129.56	4.56
150	153.98	3.98
175	166.31	-8.69

Table 4.14: Results of distance estimation for siren type 2.

Distance real	Distance estimated	Error in [m]
10	10.39	0.39
15	15.36	0.36
20	20.92	0.92
25	26.36	1.36
30	32.32	2.32
50	47.97	-2.03
75	72.02	-2.98
100	103.03	-1.97
125	127.19	2.19
150	136.14	-13.86
175	133.14	-41.86

Table 4.15: Results of distance estimation for siren type 3.

4.6 Conclusion

This chapter has presented a sound processing system for AD systems. Sound can be a critical source of information for determining the optimal driving trajectory as well as for decision making.

Section 4.1 presented the taxonomic categorization of traffic sounds for classification and tracking. A dataset and transport classification system based on the analysis of the acoustic landscape of the surrounding cars were also presented. The presented system has the ability to classify the actions performed by a particular object, such as acceleration and braking. This information can be of great interest and play crucial roles in planning the optimal trajectory as well as in the decision making of an autonomous vehicle.

In Section 4.2, an approach was proposed to improve the DOA algorithm for unknown audio signals. The SRP algorithm was reasonably accurate, but it was long and computationally demanding. The predetermination of the arrival sector reduced the time required to process the data array by finding a valid value – one that was not in the entire FoV of the microphone array but only in the predetermined sector.

Furthermore, as part of the sound processing system for AD, an approach for determining the distance to the sound source of EmVs based on the amplitude of the sound signal was proposed. Since EmVs are quite loud in the context of the soundscape and interactions with them are different from those with ordinary vehicles, the recognition of this type of vehicle could be critical. Since our system has the ability to classify the sirens of EmVs, all of the initial characteristics of the sound source of the siren can be predetermined. Therefore, analyzing the amplitude characteristics of the received signal and comparing them with the initial characteristics of the emitter allows one to determine the distance traveled by the sound wave. In other words, determining the distance to an EmV traveling far behind or ahead is possible.

5 Calibration

In this chapter, we introduce cameras to the microphone array calibration approach with the aim of visualizing sound sources in camera images. This would have huge potential for the integration of audio signal processing in AD sensor setups as well as provide a mechanism for audio–visual data fusion.

5.1 Introduction

The previous chapters have covered methods and approaches for audio signal processing for AD systems. Sound is a crucial source of information, not only in general but also in the context of road traffic. From personal experience with driving, all drivers knows that when a sharp and suspicious sound or horn becomes audible, they must look for the source of the sound or attempt to determine where it came from. Of course, however, sound alone is not sufficient to ensure safe driving.

Consequently, in human perception, sound is a marker for attracting attention. This is why all cars are equipped with a horn and all EmVs have a powerful sound source in the form of a siren, which are used to attract the attention of other traffic participants.

Drawing an analogy with a driver, an autonomous vehicle must have the ability to visually identify the sound source itself in addition to the abilities of sound-based recognition and localization. Thus, it is necessary to integrate and combine acoustic data with data from other perception systems, including machine vision.

Audio information can be used to confirm data provided by machine vision systems, but simultaneously it can provide entirely new information about events that are only acoustically representable. At large distances, audio source detection can be even more reliable than image recognition, since distant objects are only a few pixels in a camera image. At a poor resolution, the accuracy of image recognition drops dramatically [42].

The aim of this chapter is to find a way to integrate a machine listening system into the overall system of a self-driving car. At the moment, an

autonomous vehicle is represented by a system of several cameras that share a common reference system. The array of microphones can be thought of as a single sensor that gives the coordinates of the position of the sound source. In order to transfer these coordinates into the common reference frame of the camera system, the microphone array needs to be externally calibrated. This means that its relative position in the camera system has to be determined. Obtaining the external parameters of the array practically completes its integration. The joint system visualizes the localization results from the array by marking them on the camera images.

The aim of this chapter is to determine an approach for integrating a machine listening system into the overall system of AD. At the present, an autonomous vehicle is represented by a system of several cameras that share a common reference system. An array of microphones can be thought of as a single sensor that provides the coordinates of the position of the sound source. In order to transfer these coordinates into the common reference frame of the camera system, the microphone array must be externally calibrated. This means that its relative position in the camera system must be determined. Obtaining the external parameters of the array practically completes its integration. The joint system visualizes the localization results from the array by marking them on the camera images.

5.2 Camera to Microphone Array Calibration

5.2.1 Camera Calibration

Cameras are one of the most crucial sources of information for autonomous vehicle perception systems and are used for visual odometry, navigation, and obstacle recognition. Cameras are quite small and inexpensive compared with other sensors used for environmental sensing, and they have a sufficiently wide operating range in a variety of conditions. Most often, a set of cameras is combined into a multi-camera system capable of covering a 360° field around the vehicle, and their calibration must meet sufficiently high accuracy requirements. This chapter describes how to use cameras to determine the external parameters of the microphone array and to visualize the location of sound sources in the camera images.

The calibration of computer vision systems is not the focus of this research; therefore, the camera system discussed in this chapter is quite simple, but it is nevertheless capable of providing a proof of concept for

combining audio and video data for AD systems. It consists of several monocular cameras with overlapped FoV. This overlap is necessary because the world origin (checkerboard) must be in the FoV of each camera to obtain the relative position and orientation. The following two subsections describe the intrinsic and extrinsic calibration of the cameras.

Intrinsic camera calibration

The process can be explained on ideal pinhole camera is one without a lens but with a tiny aperture. The lack of a lens allows the distortion of the lens to be neglected. Thus, the system as a whole considers three coordinate systems: the world coordinate system in which P has X, Y, and Z coordinates, the camera coordinate system, in which P has x, y, and z coordinates, and finally the image plane coordinate system in which P has u, and v 2D coordinates.

The world coordinate system is fixed in relation to the camera. The relative position between the camera and the world can be described by the rotation R and translation t. The values of R and t are determined through external calibration as:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = R \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + t, \tag{5.1}$$

P in pixel coordinates u and v are calculated as:

$$x' = x/z, \tag{5.2}$$

$$y' = y/z, \tag{5.3}$$

$$u = f_x x' + c_x, \tag{5.4}$$

$$v = f_y y' + c_y, \tag{5.5}$$

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \underbrace{\begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}}_{K} \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \tag{5.6}$$

where x' and y' describe the projection of P onto the image plane of the pinhole camera, u and v are the position of the actual pixel, K is the camera's own matrix, describing the entire transformation from P as a three-dimensional point in the camera coordinate frame to the corresponding two-dimensional pixel, (f_x, f_y) are the focal lengths, and (c_x, c_y) are the focal points.

Determining the focal lengths and the main point is the task of the internal calibration procedure. K is independent of the scene being viewed as long as the focal length is not changed with the zoom lens.

A pinhole camera is only an approximate model of a real camera. Unlike a pinhole camera, a real camera has a lens that introduces distortion into the image (Figure 5.1). The distortion parameters are part of the camera properties and must be considered. With distortion considered, the projection of image P on pixel coordinates u and v can be described as follows:

$$x' = x/z, \tag{5.7}$$

$$y' = y/z, \tag{5.8}$$

$$x" = x' \frac{1 + k_1 r^2 + k_2 r^4 + k_3 r^6}{1 + k_4 r^2 + k_5 r^4 + k_6 r^6} + 2p_1 x' y' + p_2(r^2 + 2x'^2), \tag{5.9}$$

$$y" = y' \frac{1 + k_1 r^2 + k_2 r^4 + k_3 r^6}{1 + k_4 r^2 + k_5 r^4 + k_6 r^6} + p_1(r^2 + 2y'^2) + 2p_2 x' y', \tag{5.10}$$

$$r^2 = x'^2 + y'^2, \tag{5.11}$$

$$u = f_x x'' + c_x, \tag{5.12}$$

$$v = f_y y'' + c_y, \tag{5.13}$$

where k_1, k_2, k_3, k_4, k_5 and k_6 are the radial distortion coefficients, p_1 and p_2 are the tangential distortion coefficients.

Figure 5.1 shows the effect of barrel distortion ($k_1 > 0$) and pincushion distortion ($k_1 < 0$). Non-parallelism of lens and the image plane causes the tangential distortion occurrence.

Figure 5.1: Types of distortion.

Extrinsic camera calibration

External camera calibration is performed to determine the exact position of the camera within the world frame. If the camera is mounted on a vehicle, the world frame is at a fixed location inside the vehicle and is the common reference point for all sensors.

By taking pictures of the checkerboard with previously known parameters, it is possible to estimate the relative position between the camera and the checkerboard. Checkerboard tracking is only possible if the internal characteristics of the camera are known. Figure 5.2 depicts the geometric contour for cam1. The position of the checkerboard frame inside the world frame must be measured beforehand.

5.2.2 Microphone Array Transfer Function

The transfer function is a model of sound propagation characteristics in the environment relative to the spatial sensitivity of the microphone array. It is assumed that the array is part of a linear time-invariant system. This means that the environment and equipment do not change over time.

$$y_m(t) = h_m(t) * s(t), \tag{5.14}$$

where $s(t)$ is an audio signal, $y_m(t)$ is the microphone m response, $h_m(t)$ is the transfer function and $*$ denotes the convolution.

If the signal $s(t)$ is a pulse $\delta(t)$, then the pulse response itself is the transfer function:

$$h_m(t) = h_m(t) * \delta(t). \tag{5.15}$$

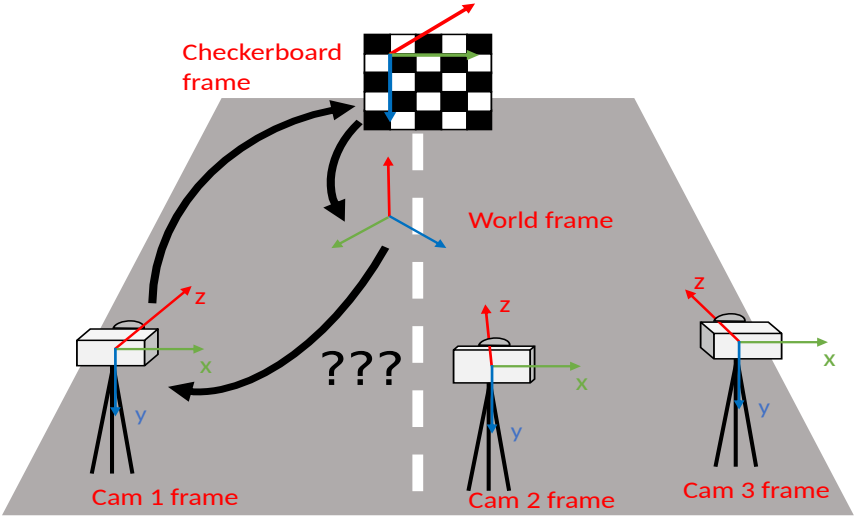

Figure 5.2: Geometric loop for obtaining camera extrinsics.

Thus, the transfer function can be determined when reproducing the pulse with subsequent recording of the response. However, an impulse is an almost instantaneous sharp and very loud signal, which can be roughly compared to a shot and cannot be reproduced using standard audio equipment.

Therefore, Time-Stretched Pulse (TSP) are used to determine the transfer function as presented in Figure 5.3. A TSP is a linear signal stretched in the frequency domain and can be reproduced. The convolution of TSPs with Inverse TSP (ITSP), obtains the initial pulse as:

$$\delta(t) = tsp(t) * itsp(t). \tag{5.16}$$

The principle presented in (5.17) can be used to obtain the impulse response, and hence the transfer function, from the TSP responses:

$$h_m(t) = h_m(t) * tsp(t) * itsp(t), \tag{5.17}$$

where $h_m(t) * tsp(t)$ is the response of the microphone m to the TSP.

To determine the transfer function, the TSP is reproduced sequentially approximately 10–20 times. The response is then averaged over the number of recorded TSPs to minimize the impact of noise during recording.

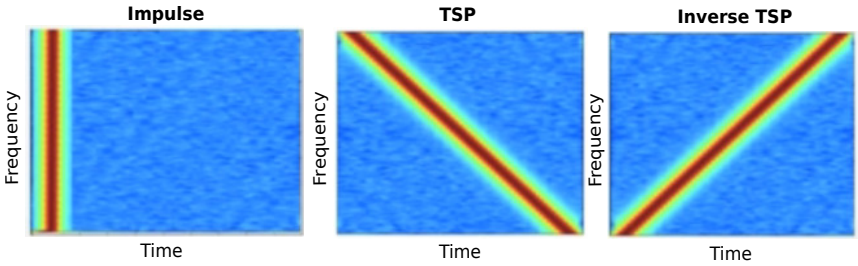

Figure 5.3: Spectrogram of the impulse, TSP, and inverse TSP.

Measurements of the transfer function are conducted at predetermined points at a certain distance around the microphone array. The layout of the calibration points is presented in Figure 5.4. The result is a matrix with four dimensions:

$$\underbrace{\underline{\underline{H}}(\Theta_a, r_d, \omega_k)}_{M \times N_\Theta \times N_r \times N_\omega} = [h_1(\Theta_a, r_d, \omega_k), \cdots, \underbrace{h_m(\Theta_a, r_d, \omega_k)}_{N_\Theta \times N_r \times N_\omega}, \cdots, h_M(\Theta_a, r_d, \omega_k)], \tag{5.18}$$

where $\underline{\underline{H}}$ is the transfer function, where elements are complex numbers related to the transfer characteristics for each combination of discrete microphone m, azimuth Θ_a, radius r_d and frequency bin ω_k. The possible calibration point layout is presented in Figure 5.4.

5.2.3 Extrinsic Array Calibration

This section presents an approach for determining the external parameters of the microphone array, which describe the position and orientation of objects within the world frame. In the previous section, the external and internal calibration of cameras with an intersecting FoV was discussed. This section describes how the calibrated cameras are used to calibrate the microphone array.

For combined calibration, we propose using an installation that consists of both sound and visual elements rigidly attached to each other. The cameras can track the location of the speaker in the world frame because the relative pose between the checkerboard and the speaker is fixed. The array can track the speaker by sound, but there is still no connection between the world frame and the array frame. The principle is to repeatedly localize

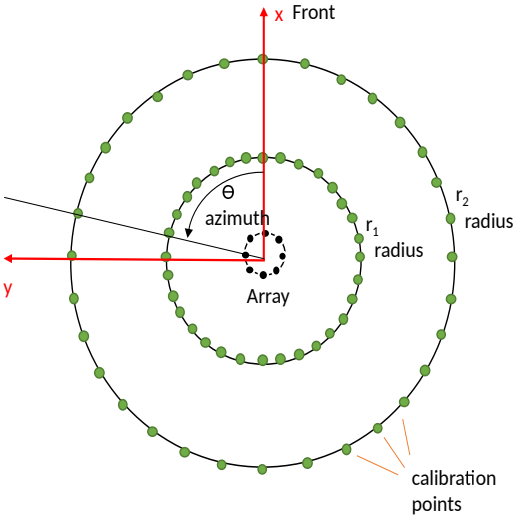

Figure 5.4: TSP calibration points layout.

the speaker using both an array of microphones and cameras, thus enabling the position of the array within the world frame to be estimated.

The microphone array operates only in a plane: the transfer function has been measured in the horizontal plane spanned by the x- and y-axes of the array. Therefore, the array will automatically place the speaker in this plane regardless of its true height along the z-axis of the array.

Figure 5.5 presents a simplified geometric loop excluding 3D objects. The mathematical description can be presented as follows:

$$W_{tWS_n} = W_{tWA} + W_{AR}A_{AtS_n}, \tag{5.19}$$

$$W_{tWA} + W_{AR}A_{tAS_n} - W_{tWS_n} = 0, \tag{5.20}$$

$$A_{tAS_n} = a_n A_{dAS_n}, \tag{5.21}$$

$$W_{AR} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}, \tag{5.22}$$

where W is a vector in the world frame of reference, WS indicates that the vector points from the world source W to the speaker S, n is the

number of the current reference point as the calibration unit moves across the room, and W_{AR} describes the rotation of the array frame with respect to the world frame.

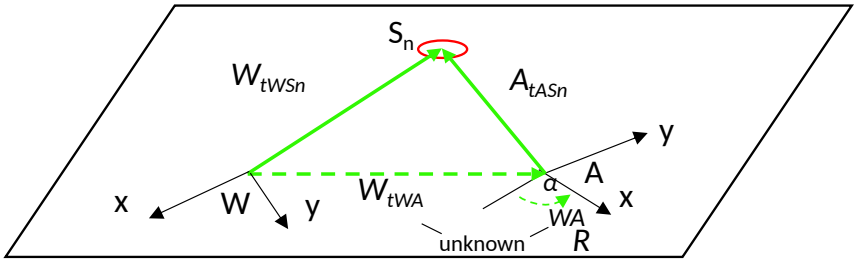

Figure 5.5: Simplified geometric loop.

Because distance estimation to the speaker is imprecise, only DOA is considered. Therefore, A_{tAS_n} is assumed to be a direction vector A_{dAS_n} of unknown length and scaling factor a_n. The components x and y can be decomposed as follows:

$$W_{tWA_x} + a_n(A_{dAS_{nx}} \cos\alpha - A_{dAS_{ny}} \sin\alpha) - W_{tWS_{nx}} = 0, \tag{5.23}$$

$$W_{tWA_y} + a_n(A_{dAS_{nx}} \sin\alpha + A_{dAS_{ny}} \cos\alpha) - W_{tWS_{ny}} = 0, \tag{5.24}$$

Hence a scalar a_n is represented as:

$$a_n = \frac{W_{tWS_{nx}} - W_{tWA_x}}{A_{dAS_{nx}} \cos\alpha - A_{dAS_{ny}} \sin\alpha}. \tag{5.25}$$

Substituting the a_n value into equation obtains the following:

$$f_a = W_{tWA_y} + (W_{tWS_{nx}} - W_{tWA_x}) \frac{A_{dAS_{nx}} \sin\alpha + A_{dAS_{ny}} \cos\alpha}{A_{dAS_{nx}} \cos\alpha + A_{dAS_{ny}} \sin\alpha} - W_{tWS_{ny}} = 0, \tag{5.26}$$

$$x = \begin{bmatrix} W_{tWA_x} \\ W_{tWA_y} \\ \alpha \end{bmatrix}, \tag{5.27}$$

where W_{tWA_x}, W_{tWA_y} and α are the unknown external parameters of the microphone array.

The remaining parameters are determined by localizing the calibration rig. To obtain a system of three non-linear equations such as (5.26) and solve them for all three external parameters, the calibration setup should theoretically only be adjusted at 3 different points (n = 1, 2, 3). In reality, due to the relatively low directional resolution of the array (5°), the data points are affected by large uncertainties. Higher accuracy of the measured array position can be achieved by making additional determinations.

Each data point results in one additional non-linear equation f_n. A non-linear least squares fitting method is used to approximate the solution for this overdetermined system to minimize the sum of the residuals squares for each individual non-linear equation. The solution of the nonlinear least squares can be implemented using linear search methods or trust region methods.

To minimize the squares of the residuals of n functions, f_n, in the parameters p, x_n, the cost function Φ is defined:

$$\Phi(x) = \frac{1}{2}||f(x)||^2 = \frac{1}{2}\sum_{i=1}^{n} f_i(x_1, \cdots, x_p)^2. \tag{5.28}$$

The number of data points n must be greater or equal to the number of parameters p. In case of external array calibration, it is three parameters W_{tWA_x}, W_{tWA_y} and α.

The model function $m_k(\delta)$ is generated to approximate $\Phi(x)$ around unknown point x_k in order to determine the minimization of $\Phi(x)$, where x_k is the vector of parameters x at iteration k. Basically, $m_k(\delta)$ is a second-order Taylor series:

$$\Phi(x+\delta) \approx m_k(\delta) = \Phi(x) + g_k^T\delta + \frac{1}{2}\delta^T B_k \delta, \tag{5.29}$$

where $g_k = \nabla\Phi(x_k) = J_k^T f(x_k)$ is the gradient vector at point x_k, $B_k = \nabla^2\Phi(x_k)$ is the Hessian matrix at point x_k, B_k is approximated as $\approx J_k^T J_k$, and J is the n-by-p Jacobian matrix:

$$J_{ij} = \frac{\partial f_i}{\partial x_j}. \tag{5.30}$$

The solution to the minimization Φ is to find the correct step size δ for all parameters x_i. This is achieved by solving a trust region subproblem (TRS), where the model function $m_k(\delta)$ is minimized:

$$\min_{\delta \in R^p} m_k(\delta) = \Phi(x_k) + g_k^T \delta + \frac{1}{2}\delta^T B_k \delta, \qquad s.t||D_k\delta|| \leq \Delta_k. \tag{5.31}$$

Here, δ must stay within the region where $m_k(\delta)$ is a good approximation of Φ. $\Delta_k > 0$ is the radius of the trust region, and D_k is the diagonal scaling matrix. GSL proposes a More scaling strategy that makes D_k scale-invariant and the scaled step $D_k\delta$ has entries of the same magnitude. Several algorithms are available for solving the trust region subproblem. In this chapter, the Levenberg–Marquardt algorithm is used, which provides an exact TRS solution.

For each δ, a trial step is first performed to check whether the objective function $\Phi(x)$ decreases. A useful indicator is the following ratio:

$$\rho = \frac{\Phi(x_k) - \Phi(x_k + \delta_k)}{m_k(0) - m_k(\delta_k)}. \tag{5.32}$$

The numerator calculates the reduction of the function caused by step δ_k. The denominator estimates the expected decrease of Φ based on the m_k model. If ρ_k is the negative, the target function is increased by δ_k and therefore, it must be rejected. A positive ρ_k value indicates that δ_k successfully decreased the target function, and hence, it can be accepted. Furthermore, ρ_k indicates how well m_k approximates Φ. If the value is near 1, this represents a good match, and the trust area can be expanded, to allow larger steps to be taken.

After the trial step has been accepted, $\Phi(x + \delta)$ is accepted as the new $\Phi(x)$, and the search for the next step is repeated. The minimization is stopped if the minimum Φ is found within a given accuracy, a given maximum number of iterations is reached, or if errors occur.

Based on Equation (5.26), the squared residuals obtained from the nonlinear function f_n must be minimized. However, to initialize the solver, the Jacobian matrix and some initial values for the vector x are required. The Jacobian matrix can be built based on the following equation:

$$\frac{\partial f_n}{\partial x_1} = \frac{\partial f_n}{\partial W_{tWA_x}} = \frac{A_{dAS_{nx}} \sin\alpha + A_{dAS_{ny}} \cos\alpha}{A_{dAS_{nx}} \cos\alpha + A_{dAS_{ny}} \sin\alpha}, \tag{5.33}$$

$$\frac{\partial f_n}{\partial x_2} = \frac{\partial f_n}{\partial W_{tWA_y}} = 1, \tag{5.34}$$

$$\frac{\partial f_n}{\partial x_3} = \frac{\partial f_n}{\partial \alpha} = \frac{A_{d^2AS_{nx}} + A_{d^2AS_{ny}}}{(A_{dAS_{ny}} \sin\alpha - A_{dAS_{nx}} \cos\alpha)^2}. \tag{5.35}$$

Almost any starting point works to ensure the algorithm converges. The condition is that the value of α_0 must be within $\pm\frac{\pi}{2}$ of the true value. Otherwise, the algorithm will stop at the pose where the array is rotated 180°.

5.2.4 Combined Intrinsic and Extrinsic Array Calibration

The array was extrinsically calibrated and its transfer function measured. When comparing the two procedures, several similarities were noted: in both cases, the speaker was carried across the array and placed in separate locations for sound reproduction. By connecting the checkerboard and the loudspeaker, while the TSP was being recorded, the extrinsic array calibration could be performed simultaneously. The exact position of the speaker in the array coordinate system (A_{tAS_n}) was known for each TSP. The sensor calibration points were considered to have 100% accuracy, and therefore, the error propagation would not occur due to limited resolution or noise.

Figure 5.4 presents a schematic representation of the TSP recording session. The calibration points are located in concentric circles around the array, with one point for each given direction (azimuth) and distance (radius). However, this operation was performed simultaneously with the loudspeaker and checkerboard. Thus, the cameras could record the position of the sound source. All of the coordinates of the TSP were stored in a file, as seen in the array. The array itself was not involved in the extrinsic calibration process and only recorded the TSP. However, the geometric loop in Figure 5.5 is valid:

$$W_{tWS_n} = W_{tWA} + W_{AR}A_{AtS_n}, \tag{5.36}$$

$$W_{tWA} + W_{AR}A_{tAS_n} - W_{tWS_n} = 0, \tag{5.37}$$

where A_{tAS_n} is the location of the TSP. Decomposing Equation (5.37) into its x and y components results in two different non linear equations:

$$f_{na} = W_{tWA_x} + A_{tAS_{nx}} \cos\alpha - A_{tAS_{ny}} \sin\alpha - W_{tWS_{nx}} = 0, \tag{5.38}$$

$$f_{nb} = W_{tWA_y} + A_{tAS_{nx}} \sin\alpha - A_{tAS_{ny}} \cos\alpha - W_{tWS_{ny}} = 0, \tag{5.39}$$

$$x = \begin{bmatrix} W_{tWA_x} \\ W_{tWA_y} \\ \alpha \end{bmatrix}. \tag{5.40}$$

During every loudspeaker detection by the camera within the TSP playback process, two new equations f_{na} and f_{nb} are added to solve the parameter vector x, which contains the same three external parameters as in the previous section.

$$\frac{\partial f_n a}{\partial x_1} = \frac{\partial f_n a}{\partial W_{tWA_x}} = 1, \tag{5.41}$$

$$\frac{\partial f_n a}{\partial x_2} = \frac{\partial f_n a}{\partial W_{tWA_y}} = 0, \tag{5.42}$$

$$\frac{\partial f_n}{\partial x_3} = \frac{\partial f_n a}{\partial \alpha} = -A_{tAS_{nx}} \sin\alpha - A_{tAS_{ny}} \cos\alpha, \tag{5.43}$$

$$\frac{\partial f_n b}{\partial x_1} = \frac{\partial f_n b}{\partial W_{tWA_x}} = 1, \tag{5.44}$$

$$\frac{\partial f_n b}{\partial x_2} = \frac{\partial f_n b}{\partial W_{tWA_y}} = 0, \tag{5.45}$$

$$\frac{\partial f_n}{\partial x_3} = \frac{\partial f_n a}{\partial \alpha} = A_{tAS_{nx}} \cos\alpha - A_{tAS_{ny}} \sin\alpha. \tag{5.46}$$

Thus, a total of four equations with three unknowns requires at least two data points. Furthermore, a non linear least-squares algorithm is applied. The Jacobian matrix must be provided by the user and consists of partial derivatives of both equations f_{na} and f_{nb}.

5.3 Experiments

5.3.1 Experimental Setups

Calibration rig

The calibration rig consisted of an active loudspeaker and a checkerboard rigidly connected to each other, as presented in Figure 5.6. This design allowed the sound source to be located by the cameras as the position of the checkerboard in relation to the loudspeaker was known and did not

Figure 5.6: Experimental calibration rig.

vary. The rig was mounted on a platform on wheels, allowing it to be moved freely to the points required for recording.

Mc Crypt PA 10/2 Aktiv loudspeaker:

- Max output power: 160 Watts
- R.M.S power: 80 Watts
- Frequency range: 50 - 20000 Hz
- Impedance: 4 ohms
- SPL: 115 dB

Checkerboard:

- Board: 1200 × 900 mm
- Vertices: 8 × 6
- Squares: 9 × 7
- Squares dimensions: 114.6 × 114.6 mm

Microphone setup

The same setup as in the previous chapter was used as an experimental microphone array: a circular eight-microphone array with a radius of 67 cm and foam and fur windscreens, mounted on a wooden frame with an eight analog channel Audio–USB interface. Every input channel was equipped with a preamplifier and a 24 bit A/D converter.

8 × SM58 Shure microphones:

- Microphone Type: Dynamic
- Polar Pattern: Cardioid
- Frequency Response: 50Hz-15kHz
- Max SPL: 94dB SPL
- Output Impedance: 150 ohms
- Sensitivity: -54.5 dBV/Pa (1.88 mV)
- Connector: XLR

U-Phoria UMC1820 USB audio interface:

- Digit capacity of DAC / ADC: 24 bit / 24 bit
- Maximum DAC Frequency (Stereo): 96 kHz
- Maximum ADC frequency: 96 kHz
- Output analog connectors: 10
- Independent headphone outputs: 2
- Analog Input Channels: 8
- Input connectors jack 6.3 mm: 8
- XLR input connectors: 8
- Microphone inputs: 8

Figure 5.7: Camera setup.

Cameras

For the calibration experiments, two webcams were mounted on the microphone array, 10 cm apart from the frame of the microphone array, as depicted in Figure 5.7. Therefore, the distance between the camera lenses was 54 cm.

Camera Logitech C270HD:

- Max Resolution: 720p/30fps
- Camera mega pixel: 0.9
- Focus type: fixed focus
- Lens type: plastic
- Diagonal field of view (dFoV): 55°

5.3.2 Procedure

Camera calibration

By capturing several consecutive images of a checkerboard with known parameters and applying a calibration algorithm, several 3D world points and corresponding 2D image points were obtained. All of the necessary functions are already presented in the OpenCV and ROS libraries. The

launch file started the camera driver node and the calibrator node. The calibrator automatically started capturing the checkerboard as soon as it recognized the pattern. After calibration, the .yaml file with intrinsic parameters was generated.

The ROS library provides a package called tf2, which tracks every coordinate frame in the system and can output transformations between two arbitrary frames if they can be geometrically related. In this process, the user is almost never in contact with the internal transformation procedures. Figure 5.8 presents the tf2 tree during an external camera calibration. In position A, cam1 does not detect the checkerboard, so the connection between the camera and the world frame is not established. In frame B, cam1 detects the checkerboard, so a geometric relationship between the camera frame and the world frame can be established and the camera extrinsics are found.

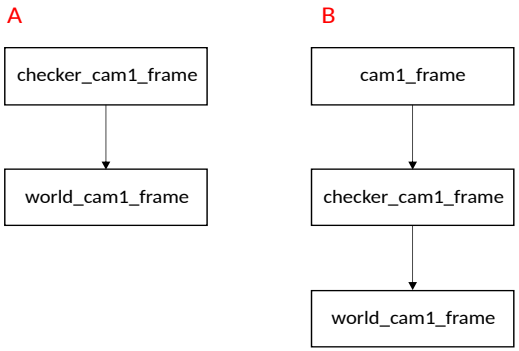

Figure 5.8: tf2-tree during extrinsic camera calibration (A: No checkerboard detected, B: Checkerboard detected).

Array transfer function

Hark software was used to determine the transfer function of the array, which provides a graphical user interface and a tool for generating the transfer function Harktool. It is installed as part of the Hark package. This package allows the user to set the array geometry and calibration points. The paths to the TSP files used must also be set. If the generation of the transfer function is successful, the output will be three files: the file that stores the description of the microphone settings for the future session, the

file containing the TSP location description, and the transfer function itself. All files must be saved in the directory defined by the ROS launch files.

Extrinsic array calibration

All nodes required for off-system array calibration, including Hark, are started using a launch file. This requires a system with at least one internally and externally calibrated camera. Furthermore, the transfer function of the microphone array must be predefined. If the default parameters are not appropriate, it is possible to edit the launch file manually. Once all nodes have been started, all further actions are conducted according to an instruction that is generated automatically. On request, the system performs a presence check of each camera in the FoV of the calibration unit. Simultaneously, the speakers produce a short sound to which the microphone array responds. If both cameras and the microphone array can locate the speaker, a new data point is added. A nonlinear least squares fitting algorithm can be initialized after at least three data points have been determined. This results in an approximation of the array pose. The program generates a request to continue collecting data points or to save the result to a file.

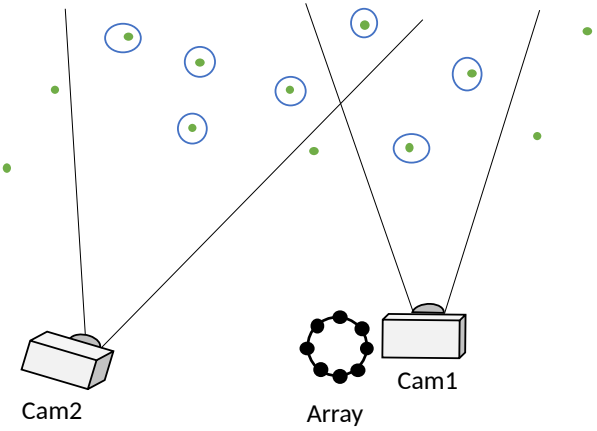

Figure 5.9: Data points addition.

Combined intrinsic and extrinsic array calibration

All necessary nodes for combined internal and external array calibration are also started through the launch file. These include the camera drivers, the checkerboard detector (one for each camera), and several tf2 translator nodes, which establish the geometric loop between the world frame and the loudspeaker, but only if the calibration setup is detected by the camera. The launch file also starts the main process node, which manages the data acquisition and accommodates a nonlinear least-squares algorithm. The node is controlled via the terminal interface. On startup, the node reads the TSP location file. It is then possible to select which TSP response is to be written next. On request, a simultaneous recording and playback of the TSP signal is performed.

Furthermore, the cameras search for a checkerboard, and if found, the position of the speaker in the world frame is determined and a data point is added.

An .xml file is also generated in which the speaker position and the corresponding TSP sequence number are stored. Each time a new data point is added, this file is overwritten.

Moreover, on request, the calculation of external parameters based on existing data points can be started and written to the .yaml file. Once all responses have been written, the TSP can start generating the transfer function.

5.3.3 Evaluation

The indoor experiment was conducted in the conference room. The TSP points were marked every 5° with a radius of 5 and 10 m. Due to space constraints, the calibration was performed only for the 40° sector in the cameras' FoV and not for the whole circle.

The extrinsic array calibration experiment achieved satisfactory results. The translation vector between the world frame and the array was computed from a total of 15 data points and then measured manually for comparison (all values in [m]):

$$W_{tWA_{computed}} = \begin{bmatrix} -2.643 \\ -0.159 \\ 1.5 \end{bmatrix}, \tag{5.47}$$

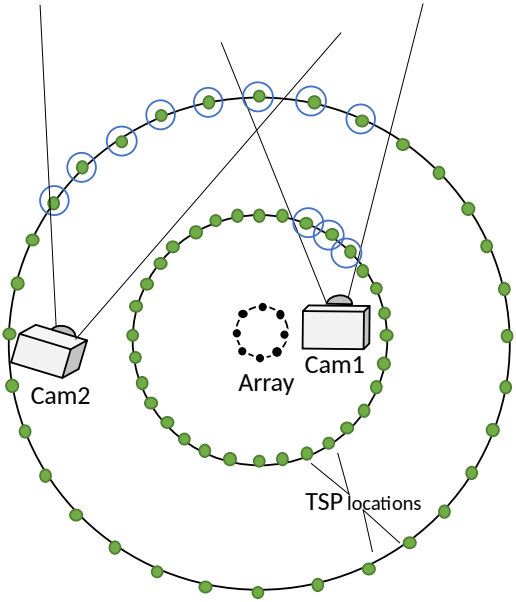

Figure 5.10: Camera calibration in combination with TSP recordings.

$$W_{tWA_{real}} = \begin{bmatrix} -2.670 \\ -0.12 \\ 1.5 \end{bmatrix}. \tag{5.48}$$

The difference between the calculated and real values for the x and y directions was less than 4 cm, which is a satisfactory result considering the low resolution of the array. As the array works only in the 2D plane, only W_{tWA_x} and W_{tWA_y} were calculated, and the z-height was set to approximately 1.5 m. The rotation angle was defined as 3.2°, which is very close to the actual value; however, it was not possible to measure the angle correctly for comparison.

Testing of the combined internal and external array calibration method was performed in the same room. Twelve data points were used to estimate translation and rotation between the world frame and the array, and manual measurements were taken for comparison (all values in [m]):

$$W_{tWA_{computed}} = \begin{bmatrix} -2.625 \\ -0.169 \\ 1.5 \end{bmatrix}, \tag{5.49}$$

$$W_{tWA_{real}} = \begin{bmatrix} -2.670 \\ -0.12 \\ 1.5 \end{bmatrix}. \tag{5.50}$$

The deviation from the actual position was 4.4 cm in x and 4.8 cm in y. This result can be considered satisfactory.

Both methods provided satisfactory results that were much closer to each other than the manually measured array position. This indicated a possible shift error common to both methods. Incorrect intrinsic or extrinsic camera calibration resulted in inaccurate speaker location readings W_{tWS_n}.

5.4 Visualization

To display the results of sound processing on camera images, a sound visualization approach – also known as an acoustic camera – was developed. Since a stream of coordinates is difficult to place in context, their visualization makes real-time interpretation of the data possible. Furthermore, the audio source visualization tool can have applications for data fusion; for example, to confirm visual classification in case the area of an object in the camera image and the audio signal coincide. Moreover, visual tracking of the sound source allows one to quickly assess the sensor calibration quality and the performance of current settings.

Figure 5.11 demonstrates how the sound source imaging algorithm works. The corner points of each sector (labeled A, B, C, and D) are the base area for the visualization frame, which refers to the calibration position S. When one moves the sound source from one area to another, the new position coincides with the next calibration position, and hence, it is displayed in a different box.

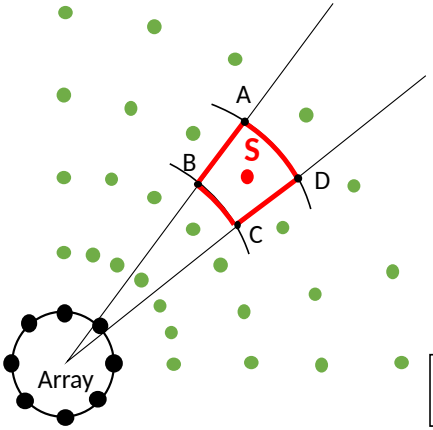

Figure 5.11: SSL visualization.

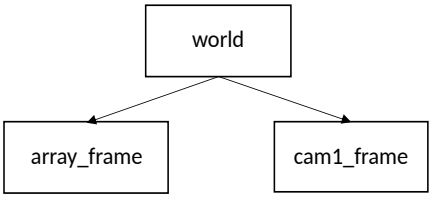

Figure 5.12: The tf2 tree visualization process.

The visualization tool requires precalibration of the cameras as well as intrinsic and extrinsic calibration of the microphone array. In this case, extrinsic calibration allows the source coordinates, or more precisely the source location field, to be converted from the array frame to the camera frame. The geometric relationship between both sensors is established using tf2. The external parameters are loaded from the .yaml files obtained during calibration. Figure 5.12 presents the tf2 process tree as an example of a single camera, the possible number of cameras is not limited.

By using more cameras, it is possible to visualize the sound source from several angles as well as to monitor a wider area. It is currently possible to visualize three sound sources simultaneously, with boxes of different colors assigned to different sources for improved differentiation. After a box is projected into the camera frame, additional transformation into pixel coordinates is performed through the OpenCV library and the intrinsic camera matrix. After applying the distortion, the pixel coordinates of the unrectified image are obtained. The box can then be projected onto the camera image. Figure 5.13 illustrates the workflow.

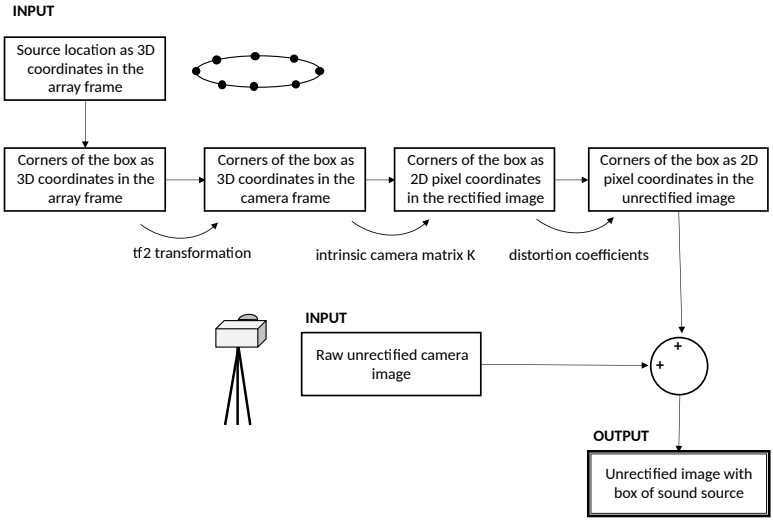

Figure 5.13: Visualization process.

Figure 5.14 presents an example of the visualization of a sound source using the developed software. A loudspeaker makes a siren sound while it is moved around the room. The checkerboard is not involved in this example as it is not localized by cameras and the visualization tool only uses pure sound. That is, the microphone array is the only sensor that locates the speaker.

A box is displayed around the speaker, which indicates the 3D space in which the sound source is assumed to be located. The dimensions of the box are determined by the template from which the transfer function was calibrated. If a greater resolution is required, it is possible to arrange the calibration points densely on the calibration template to achieve a higher matrix resolution and hence a smaller box size.

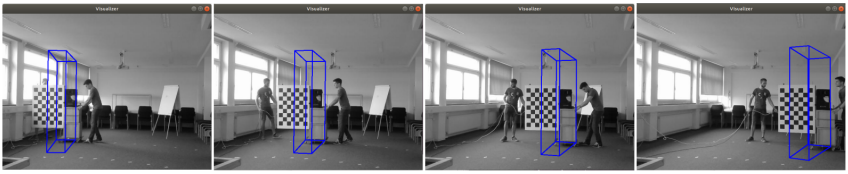

Figure 5.14: Speaker tracking visualization [5].

The visualization tool was also tested on a vehicle roof-mounted installation. For this purpose, an array of microphones on a wooden frame with cameras mounted on it was used, as indicated in the description. The complete unit was then mounted on the roof of the experimental VW Polo and the moving sound of the siren was recorded. Several shots are provided in Figure 5.16, where a person is moving around on the floor in front of the car with a Bluetooth speaker in hand, which is playing the sound of EmV siren.

Figure 5.15: Camera to microphone array calibration experiment.

Figure 5.16: Siren sound source tracking visualization.

5.5 Conclusion

This chapter has presented methods for the extrinsic and combined calibration of a microphone array to a camera for use in AD systems. Camera and microphone calibration has great potential when implementing sound processing systems in autonomous vehicles and also when integrating acoustic data with other sensors in an autonomous vehicle. As data fusion approaches are dominant in safety improvement, any kind of new data for environment representation should be available for fusion with the rest of the autonomous vehicle sensor setup data.

In addition to calibration, an acoustic data visualization tool was introduced, which also allows approaches for combining data from different sensors. Simultaneously, visualization of the sound source can represent an additional support system for testing and tuning the sound processing system. Data visualization enables real-time monitoring of the correct operation of sound processing algorithms as well as monitoring of the accuracy of the system's operation.

6 Conclusions and outlook

6.1 Summary

This dissertation has presented methods and approaches for audio signal processing as part of perception systems for autonomous vehicles.

The first chapter provided an overview of the current challenges faced by the autonomous transport industry in introducing modern technology into the established transport system. It also introduced some crucial aspects of safety and public acceptance. These included statistics on accidents involving EmVs, which may indirectly indicate a lack of localization capability for EmVs, even for human-driven vehicles.

Section 2.1 reviewed current autonomous vehicle perception systems and sensors such as LiDAR, radar, cameras, and ultrasonic sensors, as well as methods for data fusion from different sensors and systems. The design and operation of these systems as well as the limitations of their application in different environments were analyzed.

In Section 2.2, some methods and approaches for sound processing in vehicular and traffic applications were discussed. The literature review that we conducted was presented including sound processing methods in applications such as EmV identification and the recognition and classification of vehicles and other traffic-related objects, as well as approaches to acoustic road marking crossing detection and surface classification, both by cars and wheeled robots.

Section 2.3 provided a comparative analysis of existing solutions as well as several disadvantages of current perception systems, which are related to their range of operation and various harsh road and weather conditions.

Chapter 3 discussed the methods and approaches for acoustic signal processing used in various areas of robotics. Methods for sound source localization, including the determination of the sound signal DOA as well as distance estimation. Some methods and algorithms were considered and a comparative analysis was performed to determine the possibility of their application to AD systems. In the same chapter, a review of existing solutions and principles of sound signal classification systems using machine

learning algorithms and neural networks was presented.

Acoustic-based object and motion classification

Section 4.2 proposed the use of sound classification algorithms to classify traffic objects and determine motion and further actions. A taxonomic categorization of traffic acoustic landscape sounds was presented, based on which we proposed determining the current behavior of objects, such as acceleration or deceleration, in addition to object type determination.

In accordance with this taxonomy, a dataset belonging to the Motorized Transport group was assembled. The dataset was manually tagged and consisted of audio tracks downloaded from the Internet, particularly from YouTube; audio files that were recorded during test drives around the city; and recordings of several busy intersections in the city center. In addition to standard vehicles, three types of EmV sirens were recorded using a siren generator. The sounds chosen for training were as close as possible to the DIN standard requirements.

Next, a convolutional neural network with a basic architecture was trained, and the results obtained for most classes could be considered satisfactory; however, a need existed to improve the classification accuracy between acceleration and deceleration for all presented objects.

Sound direction of arrival and distance estimation

In Section 4.3.1, an accelerated method for applying the SRP-PHAT algorithm was proposed. Since the SRP method is quite complex and computationally intensive, we proposed applying this algorithm only to a small sector of arrival. Using the GCC-PHAT algorithm and a few pairs of microphones, a preliminary determination of the DOA of the sound signal was performed; then, based on the data obtained, the SRP-PHAT algorithm was applied to a sector corresponding to 10°. Thus, the peak search was not conducted over the entire 360° field, but only in a sector of 10°, which was previously defined. The method was thus significantly accelerated, allowing it to be used in real-time systems.

In Section 4.3.2, an approach was proposed for determining the distance to EmVs by analyzing the amplitude of the received signal. As the section above presented possibilities for classifying traffic sounds, including EmV sirens, it was therefore possible to identify the specific type of siren generator. Since most sirens are standard and their characteristics (e.g., amplitude) can be determined in advance, it is possible to estimate the

distance to the EmV by determining the amplitude of the received signal and comparing it to the amplitude of the transmitter. This approach will enable the detection and localization of EmVs in advance at a sufficiently long distance. Thus, autonomous vehicles will be able to apply separate interaction algorithms and have more time to find an opportunity to give way. This approach will also avoid collisions with EmVs, especially at intersections in situations where EmVs appear from outside the FoV.

The general structure of the proposed system was provided in Section 4.4. The system covers most of the acoustic analyses of the transport landscape that can be performed based on acoustic signals.

Section 4.5 described the experiments and tests of the proposed approaches in a real traffic environment on public roads. A prototype of an experimental setup was also presented, which can be used to analyze the acoustic landscape as part of an AD system. The tests achieved satisfactory results, which indicated the possibility of applying the presented algorithms in practice as well as the importance of continuing research in this area.

Calibration and Visualization

Chapter 5 presented an approach to camera and microphone array calibration, both individually and in combination. The basic camera calibration capabilities and their application were discussed along with algorithms for microphone array calibration. For the mutual calibration of the cameras and the microphone array, a calibration setup consisting of a speaker and a checkerboard was constructed. Thus, the sound source together with the checkerboard in the FoV of the cameras and the position of the sound source in relation to the checkerboard is known and does not change.

After calibration, sound sources that have been captured by the microphone array can be visualized in camera images based on microphone data only, without requiring the recognition capabilities of computer vision systems. This algorithm allows preliminary testing of the microphone array's accuracy, which can be critical when setting up the system and checking its reliability. Moreover, the ability to combine data from the camera and microphones has great potential for data fusion. Combining data from different sensors is currently done in all AD projects as well as in several ADAS systems. Thus, sensor data fusion for environmental perception is a key strategy for improving the reliability of AD systems. An audio analysis system is no exception. Since navigation based on purely acoustic data is not possible, a sound analysis system, such as human hearing, is an auxiliary system for other perception systems. Consequently,

it is necessary to have additional options for integrating this data into the overall picture. The proposed visualization fits these requirements precisely as it can complement or confirm the data from the cameras. In the future, the same method can be applied to LiDAR and radar data as the concept remains the same.

6.2 Future Research

In future research, the dataset should be expanded and supplemented with new recorded in different road and weather conditions. Furthermore, sound fragments recorded in different cities and countries should be used to achieve the greatest versatility of the system. It is also possible to increase the number of classes; for example, the braking class could be divided into emergency braking and deceleration classes, as these types of traffic have different characteristics. That is, one class would have the primary sound produced by car tires during emergency braking, while the other would have the engine as the primary source.

In addition to the current dataset, the identification of nonmotorized vehicles, such as bicycles, scooters, and skateboards, as well as electric vehicles and personal urban mobility vehicles, such as electric scooters, gyroscooters, Segways, and monocars, should be considered. Acoustic signal processing could have great potential for identifying various electric-powered vehicles because electric motors emit a high-frequency sound that is not picked up by the human ear – it is barely audible. The microphone array used in this thesis has the ability to pick up sound signals up to 20,000 Hz, which theoretically makes it possible to distinguish electric motors based on acoustic information. This point requires further investigation.

In sum, acoustic data has great potential for application to autonomous vehicles. As discussed above, sound can be used to identify the type and condition of the road surface; for example, wet or icy roads cause a change in the sound generated when a wheel is in contact with the road. This approach has the disadvantage that only the surface on which the vehicle is already driving can be identified by the sound (i.e., there is no predictive capability). Furthermore, the information could be highly useful when other methods of terrain classification are not applicable. It could also be useful for confirming other systems' data about road conditions. This approach will undoubtedly have a positive impact on the safety and reliability of the system in harsh road and weather conditions.

In addition, acoustic signal analysis can be used for vehicles' ego-noise-

based diagnostics. Most drivers use their hearing for initial vehicle diagnosis. A suspicious noise may indicate the need for a full diagnosis at a service station. Modern vehicles have many diagnostic systems, but it is almost impossible to install sensors on all units; therefore, listening for noise is still the main method for determining certain faults in, for example, the chassis. In addition to the diagnostic systems already installed, noise can be used for so-called predictive diagnostics. While most sensors will signal a malfunction that has already occurred, the analysis of acoustic signals from engine and suspension components can enable faults to be identified in advance as well as signal the extreme wear of a particular component. In the context of autonomous vehicles, this capability could be critical. As the reliability and safety of autonomous vehicles are essential, as is the reliability of the system, the reliability of the vehicle platform itself is an equally critical parameter. Under AD conditions, with no driver present, there is no possibility for human diagnosis of ego-noise. Consequently, this human function must be replaced. Both the roof-mounted microphone array presented in this thesis and additional microphones mounted inside can be used to diagnose ego-noise. Consequently, the application of sound processing algorithms in this area would be both crucial and necessary.

Another important area for future research is the development of methods for calibrating the microphone array to other sensors of AD systems. In this thesis, an algorithm for calibrating cameras to microphones was presented. Similar methods could be developed for LiDAR and Radar in the future, allowing the integration of acoustic data with all major sensors of perception systems. This task will be of great importance when implementing sound processing capabilities in AD systems.

Bibliography

[1] Bosch R (2008). Assistance systems. The Bosch Yellow Jackets, Stuttgart. Technical report, The Bosch Yellow Jackets, Stuttgart, 2008.

[2] Irman Abdic, Lex Fridman, Daniel Brown, William Angell, Bryan Reimer, Erik Marchi, and Björn Schuller. Detecting road surface wetness from audio: A deep learning approach. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3458--3463, 2016.

[3] Julián Alonso, Juan M Lopez Navarro, I. Pavón, M. Recuero, César Asensio, G. Arcas, and Agustin Bravo. On-board wet road surface identification using tyre/road noise and support vector machines. *Applied Acoustics*, 76:407--415, 2014.

[4] Farzin Amzajerdian, Vincent E. Roback, Alexander Bulyshev, Paul F. Brewster, and Glen D. Hines. Imaging flash LiDAR for autonomous safe landing and spacecraft proximity operation. In *2016 AIAA SPACE*, 2016.

[5] Paul Artner. Camera to microphone calibration for sound processing. Master's thesis, TU Darmstadt, 2019.

[6] Koyel Banerjee, Dominik Notz, Johannes Windelen, Sumanth Gavarraju, and Mingkang He. Online camera LiDAR fusion and object detection on hybrid data for autonomous driving. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1632--1638, 2018.

[7] F. Beritelli, S. Casale, A. Russo, and S. Serrano. An automatic emergency signal recognition system for the hearing impaired. In *2006 IEEE 12th Digital Signal Processing Workshop and 4th IEEE Signal Processing Education Workshop*, pages 179--182, 2006.

[8] Mario Bijelic, Tobias Gruber, and Werner Ritter. A benchmark for LiDAR sensors in fog: Is detection breaking down? In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 760--767, 2018.

[9] Horst Bischof, Martin Godec, Leistner Christian, Bernhard Rinner, and Andreas Starzacher. Autonomous audio-supported learning of visual classifiers for traffic monitoring. *IEEE Intelligent Systems*, 25(3):15--23, 2010.

[10] Martin Buczko and Volker Willert. Monocular outlier detection for visual odometry. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 739--745, 2017.

[11] Weiping Cai, Shikui Wang, and Zhenyang Wu. Accelerated steered response power method for sound source localization using orthogonal linear array. *Applied Acoustics*, 71:134--139, 2010.

[12] William M. Campbell, Douglas E. Sturim, Douglas A. Reynolds, and Alex Solomonoff. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, pages I--I, 2006.

[13] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357--366, 1980.

[14] John R. Deller, John G. Proakis, and John H. L. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan Publishing Company, 1993.

[15] Daniel Ellis. Detecting alarm sounds. In *Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis*, pages 59--62, 2001.

[16] Bruno Fazenda, H. Atmoko, Fengshou Gu, Luyang Guan, and Andrew Ball. Acoustic based safety emergency vehicle detection for intelligent transport systems. In *2009 ICROS-SICE International Joint Conference*, pages 4250 -- 4255, 2009.

[17] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of ACM*, 24(6):381--395, 1981.

[18] Elisabete Freitas, Paulo Pereira, Luis Picado Santos, and Adriana Santos. Traffic noise changes due to water on porous and dense asphalt surfaces. *Road Materials and Pavement Design*, 10:587--607, 2009.

[19] Yury Furletov, Volker Willert, and Jürgen Adamy. Auditory scene understanding for autonomous driving. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 697--702, 2021.

[20] Annette Gail and Wolfram Bartolomaeus. Noise emission of structured road markings. *Procedia - Social and Behavioral Sciences*, 48:544–552, 2012.

[21] Annette Gail, Wolfram Bartolomaeus, and Marek Zöller. Influence of surface textures of road markings on tyre/road marking noise. In *2014 43rd International Congress on Noise Control Engineering (INTER-NOISE)*, volume 249, pages 1--10, 2014.

[22] Jonah Gamba. *Radar Signal Processing for Autonomous Driving*. Springer Singapore, Singapore, 2020.

[23] Todor Ganchev, Nikos Fakotakis, and George Kokkinakis. Comparative evaluation of various mfcc implementations on the speaker verification task. In *2005 10th International Conference on Speech and Computer*, pages 191--194, 2005.

[24] Wladyslaw Gardziejczyk. Comparison of vehicle noise on dry and wet road surfaces. *Foundations of Civil and Environmental Engineering*, 9:5--15, 2007.

[25] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580--587, 2014.

[26] Georg Neuman GmbH. Mikrofon-Basics (2).

[27] Martin Godec, Christian Leistner, Horst Bischof, Andreas Starzacher, and Bernhard Rinner. Audio-visual co-training for vehicle classification. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 586--592, 2010.

[28] Richard S. Goldhor. Recognition of environmental sounds. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 149--152, 1993.

[29] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, and Gang Wang.

Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354--377, 2018.

[30] Jürgen Hasch, Eray Topak, Raik Schnabel, Thomas Zwick, Robert Weigel, and Christian Waldschmidt. Millimeter-wave technology for automotive radar sensors in the 77 GHz frequency band. *IEEE Transactions on Microwave Theory and Techniques*, 60(3):845--860, 2012.

[31] Robin Heinzler, Philipp Schindler, Jurgen Seekircher, Werner Ritter, and Wilhelm Stork. Weather influence and classification with automotive LiDAR sensors. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1527--1534, 06 2019.

[32] Sepp Hochreiter and Jürgen Schmidhuber. Long- short-term memory. *Neural computation*, 9:1735--80, 12 1997.

[33] Radu Horaud, Miles Hansard, Georgios Evangelidis, and Clement Menier. An overview of depth cameras and range scanners based on time-of-flight technologies. *Machine Vision and Applications*, 27:1005--1020, 2016.

[34] Shigemi Ishida, Jumpei Kajimura, Masato Uchino, Shigeaki Tagashira, and Akira Fukuda. SAVeD: Acoustic vehicle detector with speed estimation capable of sequential vehicle detection. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 906--912, 2018.

[35] Shigemi Ishida, Song Liu, Kohei Mimura, Shigeaki Tagashira, and Akira Fukuda. Design of acoustic vehicle count system using DTW. In *ITS World Congress*, pages 1--10, 2016.

[36] ISO 11819-1 Acoustics — Measurement of the influence of road surfaces on traffic noise — Part 1: Statistical Pass-By method. Standard, International Organization for Standardization, Geneva, CH, 1997.

[37] Srinivasan Kaushik, Abhishek Raman, and K.V.S Rajeswara Rao. Leveraging computer vision for emergency vehicle detection-implementation and analysis. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1--6, 2020.

[38] Michael Kemmler, Erik Rodner, Esther-Sabrina Wacker (Name at birth: Platzer), and Joachim Denzler. One-class classification with gaussian processes. *Pattern Recognition*, 46:3507--3518, 12 2013.

[39] Osama Furqan Khan. *Multilayer Antenna Design for Automotive Radar at 77 GHz*. PhD thesis, Universität Ulm, 2020.

[40] Charles Knapp and G. Clifford Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320--327, 1976.

[41] Wuttiwat Kongrattanaprasert, Hideyuki Nomura, Tomoo Kamakura, and Koji Ueda. Detection of road surface states from tire noise using neural network analysis. *IEEJ Transactions on Industry Applications*, 130:920--925, 2010.

[42] Micha-l Koziarski and Bogus-law Cyganek. Impact of low resolution on image recognition with deep neural networks: An experimental study. *International Journal of Applied Mathematics and Computer Science*, 28(4):735--744, 2018.

[43] John D. Kraus and Ronald J. Marhefka. *Antennas, 3rd edn.* McGraw-Hill Education, New York, 2001.

[44] Yann Lecun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278--2324, 1998.

[45] You Li and Javier Ibanez-Guzman. LiDAR for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. *IEEE Signal Processing Magazine*, 37(4):50--61, 2020.

[46] Yue Li, Devesh K. Jha, Asok Ray, and Thomas A. Wettergren. Feature level sensor fusion for target detection in dynamic environments. In *2015 American Control Conference (ACC)*, pages 2433--2438, 2015.

[47] Jacqueline Libby and Anthony J. Stentz. Using sound to classify vehicle-terrain interactions in outdoor environments. In *2012 IEEE International Conference on Robotics and Automation*, pages 3559--3566, 2012.

[48] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431--3440, 2015.

[49] Richard Lyon, Andreas Katsiamis, and Emmanuel Drakakis. History and future of auditory filter models. In *2010 IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems*, pages 3809--3812, 2010.

[50] Letizia Marchegiani and Xenofon Fafoutis. How well can driverless vehicles hear? A gentle introduction to auditory perception for autonomous and smart vehicles. *IEEE Intelligent Transportation Systems Magazine*, 14(3):92--105, 2022.

[51] Letizia Marchegiani and Paul Newman. Listening for sirens: Locating and classifying acoustic alarms in city scenes. *IEEE Transactions on Intelligent Transportation Systems*, pages 1--10, 2022.

[52] Letizia Marchegiani and Ingmar Posner. Leveraging the urban soundscape: Auditory perception for smart vehicles. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6547--6554, 2017.

[53] Paul F. McManamon, Terry A. Dorschner, David L. Corkum, Larry J. Friedman, Douglas S. Hobbs, Michael Holz, Sergey Liberman, Huy Q. Nguyen, Daniel P. Resler, Richard C. Sharp, and Edward A. Watson. Optical phased array technology. *Proceedings of the IEEE*, 84(2):268--298, 1996.

[54] Filippo Meucci, Laura Pierucci, Enrico Del Re, L. Lastrucci, and P. Desii. A real-time siren detector to improve safety of guide in traffic environment. In *2008 16th European Signal Processing Conference*, pages 1--5, 2008.

[55] Mahesh Nandwana and Taufiq Hasan. Towards smart-cars that can listen: Abnormal acoustic event detection on the road. In *2016 INTERSPEECH*, pages 2968--2971, 2016.

[56] Sid Odedra, Stephen Prior, Mehmet Karamanoglu, Mehmet Erbil, and Siu-Tsen Shen. Using acoustic sensor technologies to create a more terrain capable unmanned ground vehicle. In *Engineering Psychology and Cognitive Ergonomics*, volume 5639, pages 574--579, 2009.

[57] James O'Neill, William T. Moore, Kevin Williams, and Robert Bruce. Scanning system for LiDAR, 2011.

[58] Joseph W. Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):1215--1247, 1993.

[59] Christopher V. Poulton, Matthew J. Byrd, Erman Timurdogan, Peter Russo, Diedrik Vermeulen, and Michael R. Watts. Optical phased arrays for integrated beam steering. In *2018 IEEE 15th International Conference on Group IV Photonics (GFP)*, pages 1--2, 2018.

[60] Victor Rabinovich and Nicolai Alexandrov. *Antenna Arrays and Automotive Applications*. Springer-Verlag, New York, 2012.

[61] Krothapalli S. Rao and Manjunath K.E. *Speech Recognition Using Articulatory and Excitation Source Features*, pages 85--88. Springer, 2017.

[62] Husniza Razalli, Rusyaizila Ramli, and Mohammed Hazim Alkawaz. Emergency vehicle recognition and classification method using HSV color segmentation. In *2020 16th IEEE International Colloquium on Signal Processing Its Applications (CSPA)*, pages 284--289, 2020.

[63] Mark A. Richards. *Fundamentals of Radar Signal Processing*. McGraw-Hill, New York, 2005.

[64] Shuvendu Roy and Md. Sakif Rahman. Emergency vehicle detection on heavy traffic road from CCTV footage using deep convolutional neural network. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1--6, 2019.

[65] Olga Russakovsky, Jia Deng, and Hao et al. Su. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211--252, 2015.

[66] Justin Salamon, Christopher Jacoby, and Juan Bello. A dataset and taxonomy for urban sound research. In *2014 22nd ACM International Conference on Multimedia*, pages 1041--1044, 2014.

[67] R. Murray Schafer. *The Soundscape: Our Sonic Environment and the Tuning of the World.* Destiny Books, 1993.

[68] Matthias Schreier, Volker Willert, and Jürgen Adamy. Grid mapping in dynamic road environments: Classification of dynamic cell hypothesis via tracking. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3995--4002, 2014.

[69] Jens Schröder, Stefan Goetze, Volker Grützmacher, and Jörn Anemüller. Automatic acoustic siren detection in traffic noise by part-based models. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 493--497, 2013.

[70] Michael Seiter, Hans-Jörg Mathony, and Peter Knoll. *Handbook of Intelligent Vehicles*, chapter Parking Assist. Springer, London, 2012.

[71] Sergey Shadrin. Affordable and efficient autonomous driving in all weather conditions. *World Automotive Congress (FISITA): Disruptive Technologies for Affordable and Sustainable Mobility*, 1:673--682, 2018.

[72] Babak Shahian Jahromi, Theja Tulabandhula, and Sabri Cetin. Real-time hybrid multi-sensor fusion framework for perception in autonomous vehicles. *Sensors*, 19(20), 2019.

[73] Oleg S. Sikorsky. Review of convolutional neural networks for the problem of image classification. *New information technologies in automated systems*, 20:37--42, 2017.

[74] Abhinav Valada and Wolfram Burgard. Deep spatiotemporal models for robust proprioceptive terrain classification. *The International Journal of Robotics Research*, 36(13-14):1521–1539, 2017.

[75] Abhinav Valada, Luciano Spinello, and Wolfram Burgard. *Robotics Research: Volume 2*, chapter Deep Feature Learning for Acoustic-based Terrain Classification, pages 21 -- 37. Springer International Publishing, 2018.

[76] Nancy Van Derveer. Ecological acoustics: human perception of environmental sounds, 1979. Thesis (Ph.D.) -- Cornell University, 1979.

[77] Harry L. Van Trees. *Optimum Array Processing Part. IV*. Wiley Interscience, 2002.

[78] José F. Velasco, Mohammad J. Taghizadeh, Afsaneh Asaei, Hervé Bourlard, Carlos J. Martín-Arguedas, Javier Macias-Guarasa, and Daniel Pizarro-Perez. Novel GCC-PHAT model in diffuse sound

field for microphone array pairwise distance based calibration. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2669--2673, 2015.

[79] Ulla Wandinger. *Lidar*, chapter Introduction to LiDAR. Springer-Verlag, New York, 2005.

[80] DeLiang Wang. *Speech Separation by Humans and Machines*, chapter On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis. Springer, Boston, 2005.

[81] Zhangjing Wang, Yu Wu, and Qingqing Niu. Multi-sensor fusion in automated driving: A survey. *IEEE Access*, 8:2847--2868, 2020.

[82] Volker Willert, Julian Eggert, Jürgen Adamy, Raphael Stahl, and Edgar Korner. A probabilistic model for binaural sound localization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(5):982--994, 2006.

[83] Volker Winkler. Range doppler detection for automotive FMCW radars. In *2007 European Radar Conference*, pages 166--169, 2007.

[84] De Jong Yeong, Gustavo Velasco-Hernandez, John Barry, and Joseph Walsh. Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors*, 21:2140, 2021.

[85] Han Woong Yoo and Georg Schitter. MEMS-based LiDAR for autonomous driving. *Elektrotechnik und Informationstechnik*, 135(6):408--415, 2018.

[86] Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi. 3D-CVF: Generating joint camera and LiDAR features using cross-view spatial feature fusion for 3D object detection. In *Computer Vision – ECCV 2020: 16th European Conference*, page 720–736, 2020.

[87] Fang Zheng, Guoliang Zhang, and Zhanjiang Song. Comparison of different implementations of MFCC. *Journal of Computer Science and Technology*, 16:582–589, 2001.

[88] Dmitry N. Zotkin and Ramani Duraiswami. Accelerated speech source localization via a hierarchical search of steered response power. *IEEE Transactions on Speech and Audio Processing*, 12(5):499--508, 2004.

Curriculum Vitae

Personal Information	
Name	Yury Furletov
Date of Birth	29.06.1993
Place of Birth	Moscow, Russian Federation

Academic Career	
Since October 2017	Research associate and PhD-student at the Control Methods and Robotics Lab (RMR), Department of Electrical Engineering, TU Darmstadt
2015 -- 2017	Master of Science in Automotive Engineering, Moscow Automobile and Road Construction State Technical University (MADI), Specialization: Autonomous Vehicles Research and Testing
2011 -- 2015	Bachelor of Science in Vehicle Operation Engineering, Moscow Automobile and Road Construction State Technical University (MADI), Specialization: Transport Telematics
2011	General university entrance qualification (Russian Unified State Exam), State Comprehensive School №1909, Moscow, Russian Federation